Sensitivity analysis in general metric spaces

Fabrice Gamboa, Thierry Klein, Agnès Lagnoux, Leonardo Moreno

Sensitivity analysis in general metric spaces

Fabrice Gamboa[1], Thierry Klein[2], Agnès Lagnoux[3], and Leonardo Moreno[4]

[1]Institut de Mathématiques de Toulouse and ANITI; UMR5219. Université de Toulouse; CNRS. UT3, F-31062 Toulouse, France.
[2]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; ENAC - Ecole Nationale de l'Aviation Civile, Université de Toulouse, France
[3]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT2J, F-31058 Toulouse, France.
[4]Departamento de Métodos Cuantitativos, FCEA, Universidad de la República, Uruguay

January 19, 2021

Abstract

Sensitivity indices are commonly used to quantity the relative influence of any specific group of input variables on the output of a computer code. In this paper, we introduce new sensitivity indices adapted to outputs valued in general metric spaces. This new class of indices encompasses the classical ones; in particular, the so-called Sobol indices and the Cramér-von-Mises indices. Furthermore, we provide asymptotically Gaussian estimators of these indices based on U-statistics. Surprisingly, we prove the asymptotic normality straightforwardly. Finally, we illustrate this new procedure on a toy model and on two real-data examples.

Keywords: Sensitivity analysis, Cramér-von-Mises distance, Pick-Freeze method, U-statistics, general metric spaces.

1 Introduction

In the last decades, the use of computer code experiments to model physical phenomena has become a recurrent task for many applied researchers and engineers. In such simulations, it is crucial to understand the global influence of one or several input variables on the output of the system. When considering these inputs as random elements, this problem is generally called (global) sensitivity analysis. We refer, for example to [7] or [29] for an overview on practical aspects of sensitivity analysis.

One of the most popular indicator to quantify the influence of some inputs is the so-called Sobol index. This index was first introduced in [26] and then considered by [30]. It is well tailored when the output space is $\mathbb{R}$. It compares using the so-called Hoeffding decomposition (see [14]) the conditional variance of the output (knowing some of the input variables) with the total variance of the output. Many different estimation procedures of the Sobol indices have been proposed and studied in the literature. Some are based on Monte-Carlo or quasi Monte-Carlo design of experiments (see [18, 25] and references therein for more details). More recently a method based

on nested Monte-Carlo [13] has been developed. 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 [11, 15] and references therein. Some other estimation procedures are based on different designs of experiments using for example polynomial chaos expansions (see [35] and the reference therein for more details).

The case of vectorial outputs was first studied in [19] and tackled using principal component analysis. In [10], the authors recover the indices proposed in [19] and showed that in some sense they are the only reasonable generalization of the classical Sobol indices in dimension greater than 2. Moreover, they provide the theoretical study of the Pick-Freeze estimators and extend their definition to the case of outputs valued in a separable Hilbert space.

Since Sobol indices are based on the variance through the Hoeffding decomposition, they only quantify the input influence on the mean value of the computer code. Many authors proposed other ways to compare the conditional distribution of the output knowing some of the inputs to the distribution of the output. In [23, 24], the authors considered higher moments to define new indices, whereas in [1, 2, 4, 6], divergences or distances between measures are used. In [8, 17], the authors used contrast functions to build goal-oriented indices. Although these works defined nice theoretical indices, the existence of an efficient statistical estimation procedure is still in most cases an open question. The case of vectorial-valued computer codes is considered in [12] where a sensitivity index based on the whole distribution of the output utilizing the Cramér-von-Mises distance is defined. The authors showed that the Pick-Freeze estimation procedure can be used providing an asymptotically Gaussian estimator of the index. This scheme requires $3N$ evaluations of the output code and leads to a convergence rate of order $\sqrt{N}$. This approach has been generalized in [9], where the authors considered computer codes valued in a compact Riemannian manifold. Once again, they used the Pick-Freeze scheme to provide a consistent estimator of their index, requiring $4N$ evaluations of the output. Unfortunately, no central limit theorem was proved.

In this work, we build general indices for a code valued in a metric space and we provide an asymptotically Gaussian estimator based on U-statistics requiring only $2N$ evaluations of the output code while keeping a convergence rate of $\sqrt{N}$. In addition, we explain that all the indices studied in [9, 10, 11, 12, 15] can be seen as particular cases of our framework. Hence, we improve the estimation scheme of [12] and [9] by reducing to $2N$ the number of evaluations of the code. Last but not least, using the results of Hoeffding [14] on U-statistics, the asymptotic normality is proved straightforwardly.

The paper is organized as follows. Section 2 is dedicated to the definition of the new indices and the presentation of their estimation via U-statistics. In Section 3, we recover the classical indices classically used in global sensitivity analysis. Furthermore, we extend the work of [9] and establish the central limit theorem that was not yet proved. We illustrate the procedure in Section 4 on a toy example and on two real-data models. The first application is about the Gaussian plume model and consists in quantifying the sensitivity of the contaminant concentration with respect to some input parameters. Second, an elliptical differential partial equation of diffusive transport type is considered. In this setting, we proceed to the singular value decomposition of the solution and we perform a sensitivity analysis of the orthogonal matrix produced by the decomposition with respect to the equation parameters. Finally, some conclusions are given in Section 5.

2 General setting

2.1 The Cramér-von-Mises indices

The main idea of Cramér-von-Mises indices is to compare the conditional cumulative distribution function (c.d.f.) to the unconditional one via some distance. The supremum norm is used in [3] while in [12] the L^2-norm is chosen. The previous approaches are global. A local approach is considred in [20].

Here, we consider a measurable function f (black-box code) defined on $E = E_1 \times E_2 \times \cdots \times E_p$ and valued in a separable metric space $(\mathcal{X}, d)$. Here, $(E_1, \mathcal{A}_1), \cdots, (E_p, \mathcal{A}_p)$ are measurable spaces. The output denoted by Z is then given by

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

where X_i is a random element of E_i and $X_1, \ldots, X_p$ are assumed to be mutually independent. Naturally, we assume that all the random variables are defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and $\omega \mapsto (X_1(\omega), \ldots, X_p(\omega))$ is a measurable application from Ω to E.

In [12], the authors studied, for $\mathcal{X} = \mathbb{R}^k$, global sensitivity indices of Z with respect to the inputs $X_1, \ldots, X_p$ based on its whole distribution (instead of considering only its second moment as done usually via the so-called Sobol indices). Those indices are based on the Cramér-von-Mises distance. To do so, they introduced a family of test functions parameterized by a single index $t = (t_1, \ldots, t_k) \in \mathbb{R}^k$ and defined by

$$T_t(Z) = \mathbb{1}_{\{Z \leqslant t\}} = \mathbb{1}_{\{Z_1 \leqslant t_1, \ldots, Z_k \leqslant t_k\}}.$$

More precisely, let $\mathbf{u}$ be a subset of $I_p = \{1, \ldots, p\}$ and let $\sim \mathbf{u}$ be its complementary in I_p ($\sim \mathbf{u} = I_p \setminus \mathbf{u}$). We define $X_{\mathbf{u}} = (X_i)_{i \in \mathbf{u}}$. For $t = (t_1, \ldots, t_k) \in \mathbb{R}^k$, let also F be the distribution function of Z:

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

and $F^{\mathbf{u}}$ be the conditional distribution function of Z conditionally on $X_{\mathbf{u}}$:

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

Obviously, for any $t \in \mathbb{R}^k$, $F^{\mathbf{u}}(t)$ is a random variable depending only on t and $X^{\mathbf{u}}$, the expectation of which is $\mathbb{E}[F^{\mathbf{u}}(t)] = F(t)$. Since for any fixed $t \in \mathbb{R}^k$, $T_t(Z)$ is a real-valued random variable, we can perform its Hoeffding decomposition with respect to $\mathbf{u}$ and $\sim \mathbf{u}$:

$$T_t(Z) = F(t) + ((F^u(t) - F(t)) + (F^{\sim u}(t) - F(t)) + R_t(X^u, X^{\sim u}),$$

where

$$R_t(X^u, X^{\sim u}) = T_t(Z) - \mathbb{E}[Y(t)T_t(Z)] - (\mathbb{E}[T_t(Z)|X_{\mathbf{u}}] - \mathbb{E}[T_t(Z)]) - (\mathbb{E}[T_t(Z)|X_{\sim \mathbf{u}}] - \mathbb{E}[T_t(Z)])$$

leading to

$$\begin{aligned}
\mathrm{Var}(T_t(Z)) &= F(t)(1 - F(t)) \\
&= \mathbb{E}\left[(F^{\mathbf{u}}(t) - F(t))^2\right] + \mathbb{E}\left[(F^{\sim \mathbf{u}}(t) - F(t))^2\right] + \mathrm{Var}(R_t(X^u, X^{\sim u})). \quad (2)
\end{aligned}$$

Then, the Cramér-von-Mises index is obtained by integrating over t with respect to the distribution of the output code Z and normalizing by the integrated total variance :

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

In this example, the collection of the test functions $T_t(Z) = \mathbb{1}_{\{Z \leqslant t\}}$ ($t \in \mathbb{R}^k$) is parameterized by a single vectorial parameter t. Since the knowledge of the c.d.f. of Z: $F(t) = \mathbb{E}[\mathbb{1}_{\{Z \leqslant t\}}] = \mathbb{P}(Z \leqslant t)$ characterizes its distribution, the index $S_{2,CVM}^{\mathbf{u}}$ depends as expected on the whole distribution of the output computer code. Using the Pick-Freeze methodology, the authors of [12] proposed an estimator which requires $3N$ evaluations of the output code leading to a convergence rate of $\sqrt{N}$.

This approach has been generalized in [9] to compact Riemannian manifolds replacing the indicator function of half-spaces $\mathbb{1}_{\{Z \leqslant t\}}$ parameterized by t by the indicator function of balls $\mathbb{1}_{\{Z \in \widetilde{B}(a_1, a_2)\}}$ indexed by two parameters a_1 and a_2. In this last work, $\widetilde{B}(a_1, a_2)$ stands for the ball whose center is the middle point between a_1 and a_2 with radius $\overline{a_1 a_2}/2$. Therein a consistent estimation scheme based on $4N$ evaluations of the function is proposed. Nevertheless, the convergence rate of the estimator is not studied.

Now we aim at generalizing this methodology to any separable metric spaces and to any class of test functions parameterized by a fixed number of elements of the metric space.

2.2 The general metric space sensitivity indices

Recall that Z lives in the space $\mathcal{X}$. Generalizing the previous approach, we consider a family of test functions parameterized by $m \geqslant 1$ elements of $\mathcal{X}$. For any $a = (a_i)_{i=1,\ldots,m} \in \mathcal{X}^m$, we consider the test functions

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

We assume that $T_a(\cdot) \in L^2(\mathbb{P}^{\otimes m} \otimes \mathbb{P})$ where $\mathbb{P}$ denotes the distribution of Z. Performing the Hoeffding decomposition on each test function $T_a(\cdot)$ and then integrating with respect to a using $\mathbb{P}^{\otimes m}$ leads to the definition of our new index.

Definition 2.1. The *general metric space sensitivity index* with respect to $\mathbf{u}$ is defined by

$$
S_{2,GMS}^{\mathbf{u}} = \frac{\int_{\mathcal{X}^m} \mathbb{E}\left[\left(\mathbb{E}[T_a(Z)] - \mathbb{E}[T_a(Z)|X_{\mathbf{u}}]\right)^2\right] d\mathbb{P}^{\otimes m}(a)}{\int_{\mathcal{X}^m} \mathrm{Var}(T_a(Z)) d\mathbb{P}^{\otimes m}(a)}, \tag{4}
$$

with $X_{\mathbf{u}} = (X_i)_{i \in \mathbf{u}}$.

Observe that the different contributions $S_{2,GMS}^{\mathbf{u}}$, $S_{2,GMS}^{\sim \mathbf{u}}$ and the integrated remaining term (see (2)) sum to 1.

Particular examples By convention, when the test functions T_a do not depend on a, we set $m = 0$.

1. For $\mathcal{X} = \mathbb{R}$, $m = 0$, and T_a given by $T_a(x) = x$, one recovers the classical Sobol indices (see [31, 30]). In this case, it appears that the parameterized test functions do not depend on the parameter a. For $\mathcal{X} = \mathbb{R}^k$ and $m = 0$, one can recover the index defined for vectorial outputs in [10, 19] by extending (4).

2. For $\mathcal{X} = \mathbb{R}^k$, $m = 1$, and T_a given by $T_a(x) = \mathbb{1}_{\{x \leqslant a\}}$, one recovers the index based on the Cramér-von-Mises distance defined in [12] and recalled in (3).

3. Consider that $\mathcal{X} = \mathcal{M}$ is a manifold, $m = 2$ and T_a is given by $T_a(x) = \mathbb{1}_{\{x \in \widetilde{B}(a_1, a_2)\}}$, where $\widetilde{B}(a_1, a_2)$ stands for the ball whose center is the middle point between a_1 and a_2 with radius $\overline{a_1 a_2}/2$. Here, one recovers the index defined in [9].

Remark 2.2. The previous two first examples can be seen as particular cases of what is called Common Rationale in [4]. More precisely, the first-order Sobol index with respect to X_i corresponds to the index η_i in [4, Equation (4)] while the Cramér-von-Mises index with respect to X_i is based on the distance between the c.d.f. F of Z and its conditional version F^i with respect to X_i. Actually, in our construction, as soon as the class of test functions T_a characterizes the distribution, the index becomes a particular case of the Common Rationale.

Analogously, the authors of [4] also consider as particular cases the expectation of the L^1-distance between the p.d.f. of Z and its conditional version with respect to X_i (index δ_i in [4, Equation (12)] and the expectation of the L^∞-distance between F of Z and its conditional version F^i (index β_i in [4, Equation (13)]). Notice that the integration in δ_i is done with respect to the Lebesgue measure whereas the integration in our general metric space sensitivity index $S_{2,GMS}^{\mathbf{u}}$ in (4) is done with respect to the distribution of the output Z. The benefit is twofold. First, the integral always exists. Second, such an integration weights the support of the output distribution.

2.3   Estimation procedure via U-statistics

Following the so-called Pick-Freeze scheme, let $X^{\mathbf{u}}$ be the random vector such that $X_i^{\mathbf{u}} = X_i$ if $i \in \mathbf{u}$ and $X_i^{\mathbf{u}} = X_i'$ if $i \notin \mathbf{u}$ where X_i' is an independent copy of X_i. Then, setting

$$Z^{\mathbf{u}} = f(X^{\mathbf{u}}), \tag{5}$$

a direct computation leads to the following relationship (see, e.g., [15]):

$$\mathrm{Var}(\mathbb{E}[T_a(Z)|X_{\mathbf{u}}]) = \mathrm{Cov}\left(T_a(Z), T_a(Z^{\mathbf{u}})\right).$$

Now let us define $\mathbf{Z} = (Z, Z^{\mathbf{u}})^\top$ and consider $(m+2)$ i.i.d. copies of $\mathbf{Z}$ denoted by $(\mathbf{Z}_i, i = 1, \ldots, m+2)$. In the sequel, $\mathbb{P}_2^{\mathbf{u}}$ stands for the law of $\mathbf{Z} = (Z, Z^{\mathbf{u}})^\top$. Setting $A = (Z_1, \ldots, Z_m)$. Then the integrand in the numerator of (4) rewrites as

$$\mathbb{E}\left[(\mathbb{E}[T_A(Z)] - \mathbb{E}[T_A(Z)|X_{\mathbf{u}}])^2\right] = \mathbb{E}_A\left[\mathrm{Var}_{X_{\mathbf{u}}}(\mathbb{E}_{Z_{m+1}}[T_A(Z_{m+1})|X_{\mathbf{u}}])\right]$$
$$= \mathbb{E}_A\left[\mathrm{Cov}_{\mathbf{Z}_{m+1}}(T_A(Z_{m+1}), T_A(Z_{m+1}^{\mathbf{u}}))\right].$$

Here the notation $\mathbb{E}_Z$ (resp. Var_Z and Cov_Z) stands for the expectation (resp. the variance and the covariance) with respect to the law of the random variable Z.

Now, for any $1 \leqslant i \leqslant m+2$, we let $\mathbf{z}_i = (z_i, z_i^{\mathbf{u}})^\top$ and we define

$$\Phi_1(\mathbf{z}_1, \ldots, \mathbf{z}_{m+1}) = T_{z_1, \ldots, z_m}(z_{m+1}) T_{z_1, \ldots, z_m}(z_{m+1}^{\mathbf{u}})$$
$$\Phi_2(\mathbf{z}_1, \ldots, \mathbf{z}_{m+2}) = T_{z_1, \ldots, z_m}(z_{m+1}) T_{z_1, \ldots, z_m}(z_{m+2}^{\mathbf{u}})$$
$$\Phi_3(\mathbf{z}_1, \ldots, \mathbf{z}_{m+1}) = T_{z_1, \ldots, z_m}(z_{m+1})^2$$
$$\Phi_4(\mathbf{z}_1, \ldots, \mathbf{z}_{m+2}) = T_{z_1, \ldots, z_m}(z_{m+1}) T_{z_1, \ldots, z_m}(z_{m+2}).$$

Further, we set

$$M(1) = M(3) = m+1 \quad \text{and} \quad M(2) = M(4) = m+2 \tag{6}$$

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

$$I(\Phi_j) = \int_{\mathcal{X}^{M(j)}} \Phi_j(\mathbf{z}_1, \ldots, \mathbf{z}_{M(j)}) d\mathbb{P}_2^{u, \otimes M(j)}(\mathbf{z}_1 \ldots, \mathbf{z}_{M(j)}). \tag{7}$$

Finally, we introduce the application Ψ from $\mathbb{R}^4$ to $\mathbb{R}$ defined by

$$\Psi : \quad \begin{array}{ccc} \mathbb{R}^4 & \to & \mathbb{R} \\ (x, y, z, t) & \mapsto & \frac{x-y}{z-t}. \end{array} \tag{8}$$

Then, $S_{2,GMS}^{\mathbf{u}}$ can be rewritten as

$$S_{2,GMS}^{\mathbf{u}} = \Psi \left(I(\Phi_1), I(\Phi_2), I(\Phi_3), I(\Phi_4) \right). \tag{9}$$

The previous expression of $S_{2,GMS}^{\mathbf{u}}$ will allow to perform easly its estimation. Following Hoeffding [14], we replace the functions Φ_1, Φ_2, Φ_3 and Φ_4 by their symmetrized version $\Phi_1^s, \Phi_2^s, \Phi_3^s$ and Φ_4^s:

$$\Phi_j^s(\mathbf{z}_1, \ldots, \mathbf{z}_{M(j)}) = \frac{1}{(M(j))!} \sum_{\tau \in \mathcal{S}_{M(j)}} \Phi_j(\mathbf{z}_{\tau(1)}, \ldots, \mathbf{z}_{\tau(M(j))})$$

for $j = 1, \ldots, 4$ where $\mathcal{S}_k$ is the symmetric group of order k (that is the set of all permutations on I_k). For $j = 1, \ldots 4$, the integrals $I(\Phi_j^s)$ are naturally estimated by U-statistics of order $M(j)$. More precisely, we consider an i.i.d. sample $(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)$ ($N \geqslant 1$) with distribution $\mathbb{P}_2^{\mathbf{u}}$ and, for $j = 1, \ldots, 4$, we define the U-statistics

$$U_{j,N} = \binom{N}{M(j)}^{-1} \sum_{1 \leqslant i_1 < \cdots < i_{M(j)} \leqslant N} \Phi_j^s \left(\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{M(j)}} \right). \tag{10}$$

Theorem 7.1 in [14] ensures that $U_{j,N}$ converges in probability to $I(\Phi_j)$ for any $j = 1, \ldots, 4$. Moreover, one may also prove that the convergence holds almost surely proceeding as in the proof of Lemma 6.1 in [12]. Then we estimate $S_{2,GMS}^{\mathbf{u}}$ by

$$\widehat{S}_{2,GMS}^{\mathbf{u}} = \frac{U_{1,N} - U_{2,N}}{U_{3,N} - U_{4,N}} = \Psi(U_{1,N}, U_{2,N}, U_{3,N}, U_{4,N}). \tag{11}$$

Remark 2.3. Naturally, a covariance quantity $\mathrm{Cov}(A, B)$ can be estimated using either the expression $\mathrm{Cov}(A, B) = \mathbb{E}[AB] - \mathbb{E}[A]\mathbb{E}[B]$ or the expression $\mathrm{Cov}(A, B) = \mathbb{E}[(A - \mathbb{E}[A])(B - \mathbb{E}[B])]$ leading to the following estimators:

$$\frac{1}{N} \sum_{i=1}^{N} A_i B_i - \left(\frac{1}{N} \sum_{i=1}^{N} A_i \right) \left(\frac{1}{N} \sum_{i=1}^{N} B_i \right) \quad \text{or} \quad \frac{1}{N} \sum_{i=1}^{N} \left(A_i - \frac{1}{N} \sum_{i=1}^{N} A_i \right) \left(B_i - \frac{1}{N} \sum_{i=1}^{N} B_i \right)$$

which are equal. The use of the right-hand side formula enables greater numerical stability (i.e., less error due to round-offs). The Kahan compensated summation algorithm [16] may also be used on these sums. However, the left-hand side formula is generally preferred in sensitivity analysis for the mathematical analysis. This analysis is of course independent of the way the estimators are numerically computed in practice. The same holds for a variance term.

Hence, the estimator $\widehat{S}_{2,GMS}^{\mathbf{u}}$ defined in (11) can be rewritten in the fashion of the right-hand side of the previous display.

Our main result follows.

Theorem 2.4. If for $j = 1, \ldots, 4$, $\mathbb{E}\left[\Phi_j^s \left(\mathbf{Z}_1, \ldots, \mathbf{Z}_{M(j)} \right)^2 \right] < \infty$ then

$$\sqrt{N} \left(\widehat{S}_{2,GMS}^{\boldsymbol{u}} - S_{2,GMS}^{\boldsymbol{u}} \right) \xrightarrow[N \to \infty]{\mathcal{L}} \mathcal{N}_1(0, \sigma^2) \tag{12}$$

where the asymptotic variance σ^2 is given by (13) in the proof below.

Proof of Theorem 2.4. The first step of the proof is to apply Theorem 7.1 of [14] to the random vector $(U_{1,N}, U_{2,N}, U_{3,N}, U_{4,N})^\top$. By Theorem 7.1 and Equations (6.1)-(6.3) in [14], it follows that

$$\sqrt{N} \left(\begin{pmatrix} U_{1,N} \\ U_{2,N} \\ U_{3,N} \\ U_{4,N} \end{pmatrix} - \begin{pmatrix} I(\Phi_1^s) \\ I(\Phi_2^s) \\ I(\Phi_3^s) \\ I(\Phi_4^s) \end{pmatrix} \right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_4(0, \Gamma)$$

where Γ is the square matrix of size 4 given by

$$\Gamma(i,j) = M(i)M(j)\mathrm{Cov}(\mathbb{E}[\Phi_i^s(\mathbf{Z}_1, \ldots, \mathbf{Z}_{M(i)})|\mathbf{Z}_1], \mathbb{E}[\Phi_j^s(\mathbf{Z}_1, \ldots, \mathbf{Z}_{M(j)})|\mathbf{Z}_1]).$$

Now, it remains to apply the so-called delta method (see [36]) with the function Ψ defined by (8). Thus, one gets the asymptotic behavior in Theorem 2.4 where σ^2 is given by

$$\sigma^2 = g^\top \Gamma g \tag{13}$$

with $g = \nabla\Psi(I(\Phi_1^s), I(\Phi_2^s), I(\Phi_3^s), I(\Phi_4^s))$ and $\nabla\Psi = (z-t)^{-2} (z-t, -z+t, -x+y, x-y)^\top$. $\quad\square$

Notice that we consider $(m+2)$ copies of $\mathbf{Z}$ in the definition of $S_{2,GMS}^{\mathbf{u}}$ (see (9)). Nevertheless, the estimation procedure only requires a N sample of $\mathbf{Z}$ (see (11)) that means only $2N$ evaluations of the black-box code which constitutes an appealing advantage of the method presented in this paper. Moreover, the required number of calls to the black-box code is independent of the size m of the class of tests functions unlike in [12] or in [9] where $(m+2) \times N$ calls to the computer code were necessary. In addition, the proof of the asymptotic normality in Theorem 2.4 is elementary and does not rely anymore on the use of the sophisticated functional delta method as in [12].

2.4 Comments

For any output code f, one may consider different choices of the family $(T_a)_{a\in\mathcal{X}^m}$ of functions indexed by $a \in \mathcal{X}^m$ leading to very different indices. The choice of the family must be induced by the aim of the practitioner. To quantify the output sensitivity around the mean, one should consider the classical Sobol indices based on the variance and corresponding to the first particular case presented in Section 2.2. Otherwise, interested in the sensitivity of the whole distribution, one should prefer a family of functions that characterizes the distribution. For instance, in the second particular case presented in Section 2.2, the functions T_a are the indicator functions of half-lines and yield the Cramér-von-Mises indices.

Moreover, since in the estimation procedure the number of output calls is independent of the choice of the family $(T_a)_{a\in\mathcal{X}^m}$, one can consider and estimate simultaneously several indices with no-extra cost. In fact, the only computational challenge relies in our capability to evaluate the functions Φ on the sample.

3 Applications in classical frameworks and beyond

3.1 Particular cases

Sobol indices In the case where $\mathcal{X} = \mathbb{R}$, $m = 0$ and the test functions T_a given by $T_a(x) = x$ (do not depend on the parameter a), we recover the classical Sobol indices. As mentioned in the Introduction, many classical methods of estimation are available. Among them, one can cite estimation procedure based on polynomial chaos expansion [35], quasi Monte-Carlo scheme [18, 25], the classical Pick-Freeze method [11, 15], and more recently a method based on nested

Monte-Carlo [13]. This last method seems to be numerically efficient. Nevertheless, it requires that all the random elements have a density with respect to the Lebesgue measure to be able to simulate under the conditional distribution. In addition, no theoretical asymptotic convergences are given.

As explained in Section 2.3, our method provides a new estimator based on U-statistics for the classical Sobol index. In that case, the estimator is given by (11) and, for $j = 1, \ldots, 4$, the $U_{j,N}$'s are given by

$$U_{1,N} = \frac{1}{N} \sum_{i=1}^{N} Z_i Z_i^{\mathbf{u}}$$

$$U_{2,N} = \frac{1}{N(N-1)} \left(\sum_{i=1}^{N} Z_i \sum_{i=1}^{N} Z_i^{\mathbf{u}} - \sum_{i=1}^{N} Z_i Z_i^{\mathbf{u}} \right) =: \frac{1}{N(N-1)} (\tilde{U}_{2,N} - \tilde{V}_{2,N})$$

$$U_{3,N} = \frac{1}{N} \sum_{i=1}^{N} Z_i^2$$

$$U_{4,N} = \frac{1}{N(N-1)} \left(\left(\sum_{i=1}^{N} Z_i \right)^2 - \sum_{i=1}^{N} Z_i^2 \right) =: \frac{1}{N(N-1)} (\tilde{U}_{4,N} - \tilde{V}_{4,N})$$

leading to
$$\widehat{S}_{2,GMS}^{\mathbf{u}} = \frac{U_{1,N} - U_{2,N}}{U_{3,N} - U_{4,N}} = \Psi(U_{1,N}, U_{2,N}, U_{3,N}, U_{4,N})$$

while in [11], the classical Pick-Freeze estimator $\widehat{S}^{\mathbf{u}}$ of $S_{2,GMS}^{\mathbf{u}}$ is given by

$$\widehat{S}^{\mathbf{u}} = \frac{U_{1,N} - (1 - 1/N^2)\tilde{U}_{2,N}}{U_{3,N} - (1 - 1/N^2)\tilde{U}_{4,N}} = \Psi(U_{1,N}, (1 - 1/N^2)\tilde{U}_{2,N}, U_{3,N}, (1 - 1/N^2)\tilde{U}_{4,N}) \qquad (14)$$

and takes into account the diagonal terms. Both procedures require $2N$ evaluations of the black-box code and have the same rate of convergence. The estimators are slightly different which induces different asymptotic variances. Finally, one may improve the estimation $\widehat{S}^{\mathbf{u}}$ using the information of the whole sample leading to $\widehat{T}^{\mathbf{u}}$ given in [11, Equation (6)]:

$$\widehat{T}^{\mathbf{u}} = \frac{\frac{1}{N} \sum_{i=1}^{N} Y_i Y_i^{\mathbf{u}} - \left(\frac{1}{N} \sum_{j=1}^{N} \frac{Y_i + Y_i^{\mathbf{u}}}{2} \right)^2}{\frac{1}{N} \sum_{i=1}^{N} \frac{(Y_i)^2 + (Y_i^{\mathbf{u}})^2}{2} - \left(\frac{1}{N} \sum_{i=1}^{N} \frac{Y_i + Y_i^{\mathbf{u}}}{2} \right)^2}. \qquad (15)$$

The sequence of estimators $\widehat{T}^{\mathbf{u}}$ is asymptotically efficient in the Cramér-Rao sense (see [11, Proposition 2.5]). In this paper, we also could have constructed a new estimator $\widehat{T}_{2,GMS}^{\mathbf{u}}$ analog version of $\widehat{S}_{2,GMS}^{\mathbf{u}}$ taking into account the whole information contained in the sample. However, based on the same initial design as $\widehat{S}^{\mathbf{u}}$ and $\widehat{T}^{\mathbf{u}}$, neither $\widehat{S}_{2,GMS}^{\mathbf{u}}$ nor $\widehat{T}_{2,GMS}^{\mathbf{u}}$ will be asymptotically efficient. Nevertheless, the estimation procedure proposed in this paper outperforms the procedure presented in [12, 9] as soon as $m \geqslant 1$.

Sobol indices for multivariate outputs   For $\mathcal{X} = \mathbb{R}^k$ and $m = 0$, one may realize the same analogy between the estimation procedure proposed in this paper and that in [10].

Cramér-von-Mises indices For $\mathfrak{X} = \mathbb{R}^k$, $m = 1$ and the test functions T_a given by $T_a(x) = \mathbb{1}_{\{x \leqslant a\}}$, we outperform the central limit theorem proved in [12]. Indeed, the estimator proposed in [12] requires $3N$ evaluations of the computer code versus only $2N$ in our new procedure. In addition, the proof therein is based on the powerful but complex functional delta method while the proof of Theorem 2.4 is an elementary application of Theorem 7.1 in [14] combined with the classical delta method.

3.2 Compact manifolds

A particular framework is the case when the output space is a compact Riemannian manifold $\mathcal{M}$. In [9], a similar index to $S_{2,GMS}^{\mathbf{u}}$ is studied in this special context, taking $T_a(x) = \mathbb{1}_{\{x \in \widetilde{B}(a_1, a_2)\}}$ as test functions, where $\widetilde{B}(a_1, a_2)$ still stands for the ball whose center is the middle point between a_1 and a_2 with radius $\overline{a_1 a_2}/2$. The authors showed that, under some restrictions on the underlying probability measure and the Riemannian manifold, the family of balls $\left(\widetilde{B}(a_1, a_2)\right)_{(a_1, a_2) \in \mathcal{M}}$ is a determining class, that is, if two probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathcal{M}$ coincide on all the events of this family, then $\mathbb{P}_1 = \mathbb{P}_2$. By this property, they proved that if their index, denoted $B_2^{\mathbf{u}}$, vanishes then the distributions of $T_a(Z)$ and $(T_a(Z)|X_{\mathbf{u}})$ coincide. Further, the performance of $B_2^{\mathbf{u}}$ in Riemannian manifolds immersed in $\mathbb{R}^d$ with $d = 2, 3$ and on the cone of positive definite matrices (manifold) is analyzed. Last, an exponential inequality for the estimator $\hat{B}_2^{\mathbf{u}}$ of $B_2^{\mathbf{u}}$ is provided together with the almost sure convergence that is deduced from. Unfortunately, no central limit theorem is given.

As a particular case of $S_{2,GMS}^{\mathbf{u}}$, the asymptotic distribution of $\hat{B}_2^{\mathbf{u}}$ can be found from Theorem 2.4. Given x, since $(a_1, a_2) \mapsto T_{(a_1, a_2)}(x)$ is a symmetric function and $m = 2$, it is verified that,

$$\Phi_1(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = \mathbb{1}_{\left\{\mathbf{z}_3, \mathbf{z}_3^{\mathbf{u}} \in \widetilde{B}(\mathbf{z}_1, \mathbf{z}_2)\right\}},$$
$$\Phi_2(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4) = \mathbb{1}_{\left\{\mathbf{z}_3, \mathbf{z}_4^{\mathbf{u}} \in \widetilde{B}(\mathbf{z}_1, \mathbf{z}_2)\right\}},$$
$$\Phi_3(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = \mathbb{1}_{\left\{\mathbf{z}_3 \in \widetilde{B}(\mathbf{z}_1, \mathbf{z}_2)\right\}},$$
$$\Phi_4(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4) = \mathbb{1}_{\left\{\mathbf{z}_3, \mathbf{z}_4 \in \widetilde{B}(\mathbf{z}_1, \mathbf{z}_2)\right\}}.$$

In this setting, the limiting covariance matrix Γ is given by $\Gamma(i, j) = M(i)M(j)\text{Cov}(L_i, L_j)$, for $i, j = 1, \ldots, 4$ where

$$L_1 = \frac{1}{6} \sum_{\tau \in \mathcal{S}_3} \mathbb{P}\left(Z_{\tau_3}, Z_{\tau_3}^{\mathbf{u}} \in \widetilde{B}(Z_{\tau_1}, Z_{\tau_2})|Z_1\right),$$
$$L_2 = \frac{1}{24} \sum_{\tau \in \mathcal{S}_4} \mathbb{P}\left(Z_{\tau_3}, Z_{\tau_4}^{\mathbf{u}} \in \widetilde{B}(Z_{\tau_1}, Z_{\tau_2})|Z_1\right),$$
$$L_3 = \frac{1}{6} \sum_{\tau \in \mathcal{S}_3} \mathbb{P}\left(Z_{\tau_3} \in \widetilde{B}(Z_{\tau_1}, Z_{\tau_2})|Z_1)\right),$$
$$L_4 = \frac{1}{24} \sum_{\tau \in \mathcal{S}_4} \mathbb{P}\left(Z_{\tau_3}, Z_{\tau_4} \in \widetilde{B}(Z_{\tau_1}, Z_{\tau_2})|Z_1\right).$$

4 Numerical applications

4.1 A non linear model

In this section, we illustrate and we compare the different estimation procedures based on the Pick-Freeze scheme and the U-statistics for the classical Sobol indices on the following toy model:

$$Z = \exp\{X_1 + 2X_2\}, \tag{16}$$

where X_1 and X_2 are independent standard Gaussian random variables. The distribution of Z is log-normal and we can derive both its probability density function and its c.d.f.

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

where Φ stands for the c.d.f. of the standard Gaussian random variable. We have $p = 2$ input variables and tedious exact computations (see [12]) lead to closed forms of the Sobol indices:

$$S^1 = \frac{1 - e^{-1}}{e^4 - 1} \approx 0.0118 \quad \text{and} \quad S^2 = \frac{e^3 - e^{-3}}{e^4 - 1} \approx 0.3738.$$

Further, the Cramér-von-Mises indices $S^1_{2,CVM}$ and $S^2_{2,CVM}$ are also explicitly computable:

$$S^1_{2,CVM} = \frac{6}{\pi} \arctan 2 - 2 \approx 0.1145 \quad \text{and} \quad S^2_{2,CVM} = \frac{6}{\pi} \arctan \sqrt{19} - 2 \approx 0.5693.$$

The reader is refered to [12] for the details of these computations.

In Figure 1, we compare the estimations of the two first-order Sobol indices obtained by both estimation procedures (U-statistics and Pick-Freeze). The total number of calls of the computer code ranges from $n = 100$ to 500000. When estimating the Sobol indices with both methodologies, we have considered samples of size $N = n/(p+1)$ so that each estimation requires a total number n of evaluations of the code. Analogously, when estimating the Cramér-von-Mises indices using U-statistics, we have also considered samples of size $N = n/(p+1)$. In contrast, when estimating the Cramér-von-Mises indices using the Pick-Freeze scheme, we have considered samples of size $N = n/(p + 2)$. We observe that both methods converge and give precise results for large sample sizes. The same kind of convergence can be observed for the estimations of the Cramér-von-Mises indices with both methodologies. Actually, the convergence is a bit slower which is not surprising due to the greater complexity of the Cramér-von-Mises indices. In addition, the estimation procedure with U-statistics outperforms the Pick-Freeze one as soon as $m \geqslant 1$. As already mentioned in Sections 2.3 and 2.4, such a better performance increases with the number m of parameters of the tests functions family. Indeed, for a fixed budget n (in other words, a fixed number of evaluations of the computer code), the needed sample size to estimate the p first-order indices with the standard Pick-Freeze scheme is $N_1 = n/(m + p + 1)$ to be compared to $N_2 = n/(p + 2)$ required in the U-statistics estimation procedure.

4.2 The Gaussian plume model

In this section, the model under study concerns a point source that emits contaminant into a uni-directional wind in an infinite domain. Such a model is also applied, for instance, to volcanic eruptions, pollen and insect dispersals, and is called the Gaussian plume model (GPM) (see, e.g., [5, 34]). The GPM assumes that atmospheric turbulence is stationary and homogeneous.
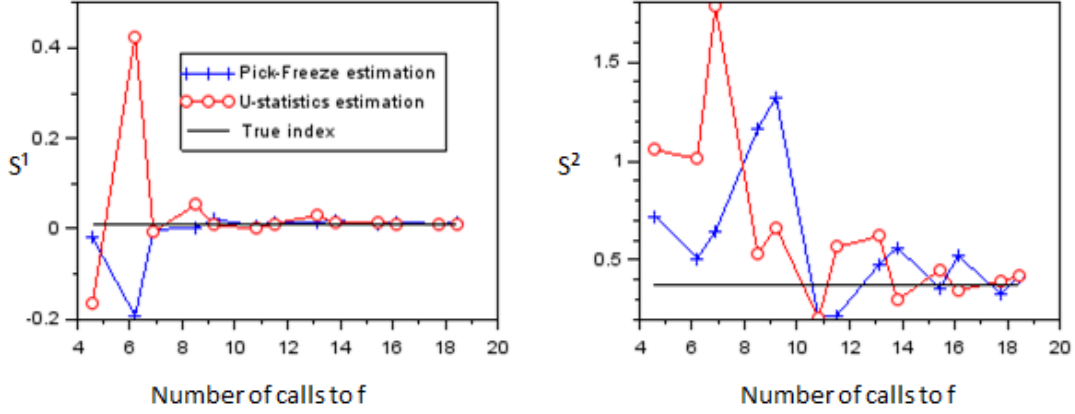
Figure 1: Non-linear model (16). Convergence of both methods when the total number of calls of the computer code increases. The two first-order Sobol indices have been represented from left to right. Several total number of calls of the computer code have been considered ranging from $n = 100$ to $n = 10^8$. The x-axis is in logarithmic scale.

Naturally, in Earth Sciences, it is crucial to analyze the sensitivity of the output of the GPM model regarding the input parameters (see [21, 27]).

The model parameters are represented in Figure 2. The contaminant is emitted at a constant rate Q and the wind direction is denoted by $\mathbf{u} = (u, 0, 0)$ (with $u \geqslant 0$) while the effective height is $H = h(1 + \delta)$ where h is the stack height and δh is the plume rise.

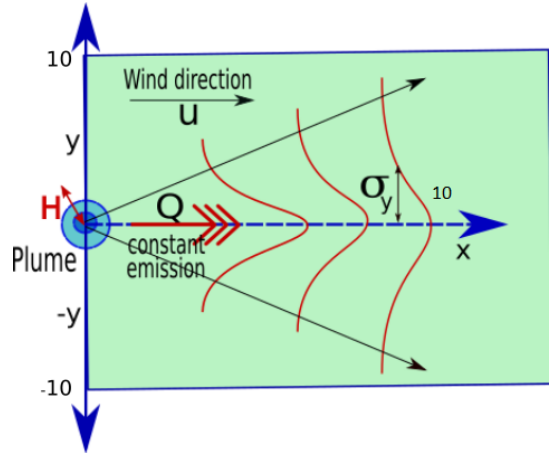Figure 2: Plume model. Cross section at $z = 0$ of a contaminant plume emitted from a continuous point source, with wind direction aligned with the x–axis.

Then the contaminant concentration at location (x, y, z) is given by

$$C(x, y, z) = \frac{Q}{4\pi u r(x)} e^{-\frac{y^2}{4r(x)}} \left(e^{-\frac{(z-H)^2}{4r(x)}} + e^{-\frac{(z+H)^2}{4r(x)}} \right),$$

where r is a parametric function given by $r(x) = \frac{1}{u} \int_0^x K(v) dv$, the function K being the eddy

11

diffusion. In this section, we investigate the particular two-dimensional case: the height is considered as zero (at ground level). In addition, we suppose that $r(x) = Kx/u$ where K is a constant. Hence, the contaminant concentration at location $(x, y, 0)$ rewrites as:

$$C(x, y, 0) = \frac{Q}{2\pi K x} e^{\frac{-u(y^2 + H^2)}{4Kx}}. \tag{17}$$

A first step consists in performing a GSA for spatial data, namely an ubiquous sensitivity analysis. In other words, the sensitivity indices are computed location after location leading to a sensitivity map. See, for instance, [22] for more details on this methodology. The results are presented in Figure 3.
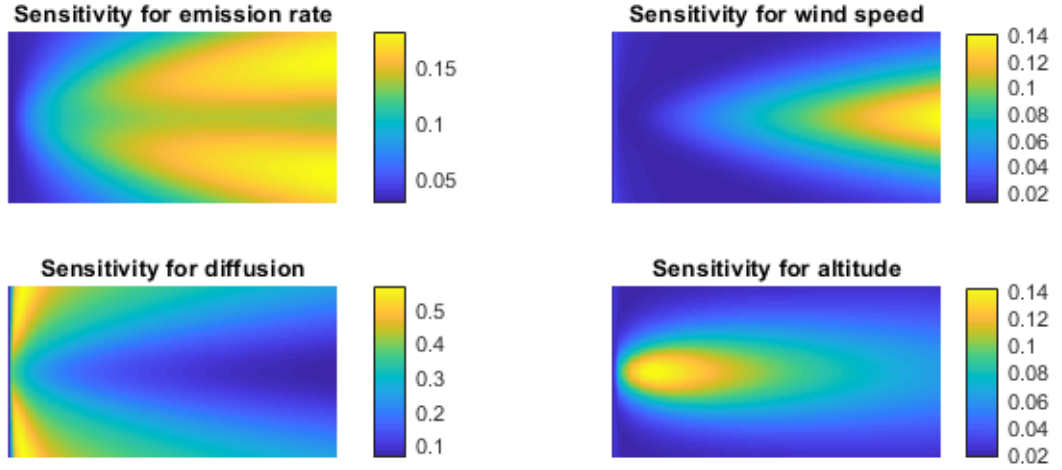

Figure 3: Plume model (17). Ubiquous sensitivity analysis with respect to the emission rate Q (top left), the wind speed u (top right), the diffusion K (bottom left) and the altitude H (bottom right).

Secondly, we wish to perform a sensitivity analysis globally on the contaminant concentration with respect to the uncertain inputs Q, K, and u, while the altitude plume parameter H is fixed in advance. In this setting, the function f that defines the output of interest in (1) is then given by:

$$f: \quad \begin{array}{ccc} \mathbb{R}^3 & \to & L^2(\mathbb{R}^2) \\ (Q, K, u) & \mapsto & f(Q, K, u) = (C(x, y, 0))_{(x,y) \in \mathbb{R}^2}. \end{array} \tag{18}$$

In other words, to any triplet (Q, K, u), the computer code associates one square-integrable field from $\mathbb{R}^2$ to $\mathbb{R}$. Based on reality constraints and guided by the expert knowledge, the stochastic parameters Q, K, and u of the model are assumed to be all independent with uniform distribution on $[0, 10]$. Let C_1 and C_2 be two pollution concentrations with domain in the ground level. The range of values of x and y where C_1 and C_2 are compared is $A = \{(x, y) \in \mathbb{R}^2 : x \in [0, 10], y \in [-10, 10]\}$. The distance used is the classical L^2 distance

$$d(C_1, C_2) = \sqrt{\iint (C_1(x, y, 0) - C_2(x, y, 0))^2 dx dy}.$$

To quantify the sensitivity on the contaminant concentration with respect to Q, K, and u, we consider the family of test functions T_a given by $T_{(a_1, a_2)}(b) = \mathbb{1}_{b \in B(a_1, a_2)}$, where a_1, a_2, and

b are square-integrable applications from $\mathbb{R}^2$ to $\mathbb{R}$ and $B(a_1, a_2)$ stands for the ball centered at a_1 with radius $\overline{a_1 a_2}$ (whence $m = 2$). The values of the indices are presented in Table 1. In this study, we have considered several values of the altitude plume parameter H from 1 to 20 and a sample size N equal to 1000, 2000, and 5000. We observe that, as H increases, the values of the sensitivity indices decrease. When $N = 5000$, we may also observe that the rank of the indices largely varies with respect to the value of H: for large values of H, the parameter K appears to be the most influent on the concentration. In contrast, when $H = 1$, all three parameters seem to have the same influence.

	N=1000			N=2000			N=5000		
	K	Q	u	K	Q	u	K	Q	u
H=1	0.1365	0.1216	0.1330	0.1124	0.1419	0.1453	0.1425	0.1431	0.1562
H=2	0.1028	0.1197	0.1212	0.1291	0.1317	0.1171	0.1222	0.1627	0.1143
H=10	0.0813	0.0891	0.1010	0.1081	0.1077	0.1256	0.0893	0.0831	0.1001
H=20	0.1027	0.0246	0.1041	0.0620	0.0942	0.1030	0.0913	0.0091	0.0329

Table 1: Sensitivity indices for the plume model (17)

4.3 Singular value decomposition in partial differential equation

In this section, we study the sensitivity of the solution (numerical approximation) of a partial differential equation, when the parameters of the equation (inputs) vary. In particular, we analyze the sensitivity of the subspaces generated by the singular value decomposition of the numerical grid output matrix solution. Many problems can be modelled by an elliptical differential partial equation. For instance, in physics, electric potential, potential flow, structural mechanics are all studied, see [32]. In biology, the reaction–diffusion–advection equation is used to model chemotaxis observed in bacteria, population migration and evolutionary adaptation to changing environments, see [37].

In this setting, it is usual to compact the information through the singular value decomposition of the solution matrix, that is, the numerical solution of the differential equation. Furthermore, it can also be useful to analyze the influence of the parameters in this information compactification.

In this section, an elliptical differential partial equation of type diffusive transport is considered:

$$B\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left[D\frac{\partial C}{\partial x}\right] + \frac{\partial}{\partial y}\left[D\frac{\partial C}{\partial y}\right] - rC + p_{xy}, \tag{19}$$

with production rate p_{xy} at location (x, y), consumption rate r, and diffusive transport D of a substance C in time t and spatial dimensions (x, y). The boundaries are prescribed as zero-gradient (default value). The parameter p_{xy} is zero everywhere except in 50 randomly positioned spots denoted by (x_i, y_i), for $i = 1, \ldots, 50$. We assume that the production rate is the same at any of the 50 locations and equal to p and we consider that the function f in (1) defining the output C is given by

$$f\colon \begin{array}{ccc} \mathbb{R}^4 & \to & L^2(\mathbb{R}_+ \times \mathbb{R}^2) \\ (B, D, r, p) & \mapsto & f(B, D, r, p) = (C(t, x, y))_{(t,x,y)\in\mathbb{R}_+\times\mathbb{R}^2} \end{array}. \tag{20}$$

All the input parameters are then assumed to be uniformly distributed:

$$B \sim \mathcal{U}([1 - \beta, 1 + \beta]),$$
$$D \sim \mathcal{U}([2 - \delta, 2 + \delta]),$$
$$r \sim \gamma \cdot \mathcal{U}([1, 2]),$$
$$p \sim \mathcal{U}([0, 1]).$$

Let $C(0, x, y)$ be the solution of (19) at time $t = 0$. We compute the matrix A that has the first two principal component scores along its columns. Note that these two columns represent a rank-2 approximation of the matrix solution. This matrix A is a way to embed the approximated solution on a Stiefel manifold S_t. That is, $A \in S_t = \{M \in \mathcal{M}_{50,2} : M^\top M = Id\}$, where $\mathcal{M}_{n,k}$ stands for the set of matrices of size $n \times k$. We consider the Stiefel manifold as an embedded one into the Euclidean space, and we choose the standard inner product in this space (Frobenius product) as metric in the Riemannian manifold. Notice that it is also possible to select another metric in S_t. Therefore, the similarity between two matrices is given by the Frobenius distance, that is, for any matrices A_1 and $B_2 \in \mathcal{M}_{n,k}$,

$$d(A_1, A_2) = \sqrt{\mathrm{tr}((A_1 - A_2)^\top (A_1 - A_2))},$$

where $\mathrm{tr}(A)$ represents the trace of the matrix A. We consider the parametric family of functions given by

$$T_{(A_1, A_2)}(\cdot) = \mathbb{1}_{\cdot \in B(A_1, A_2) \cap B(A_2, A_1)},$$

where the parameters A_1 and A_2 of the test functions are now matrices (thus are written with capital letters) and $B(A_1, A_2)$ (resp. $B(A_2, A_1)$) still stands for the ball centered at A_1 (resp. A_2) with radius $\overline{A_1 A_2}$. In Table 2, the sensitivity indices are calculated for different values of β, δ, and γ and the high influence of the parameter r is observed in all cases. As expected, this influence increases with γ and decreases as the value of δ increases. The simulations have been generated using the R language [28]. In particular, the discretized solution of the differential equation has been computed with the `ReacTran` package [33].

	$\delta = 0.1$			$\delta = 0.5$		
$\gamma = 0.001$	B	D	r	B	D	r
$\beta = 0.1$	0.001	0.011	0.546	0.020	0.071	0.119
$\beta = 0.5$	0.010	0.007	0.491	0.000	0.041	0.102
	$\delta = 0.1$			$\delta = 0.5$		
$\gamma = 0.01$	B	D	r	B	D	r
$\beta = 0.1$	0.000	0.001	0.664	0.010	0.053	0.168
$\beta = 0.5$	0.013	0.006	0.621	0.008	0.041	0.132
	$\delta = 0.1$			$\delta = 0.5$		
$\gamma = 0.1$	B	D	r	B	D	r
$\beta = 0.1$	0.005	0.005	0.794	0.020	0.051	0.179
$\beta = 0.5$	0.000	0.006	0.721	0.000	0.043	0.171

Table 2: Sensitivity indices for the partial differential equation (19)

5 Conclusion

In this paper, we explain how to construct a large variety of sensibility indices when the output space of the black-box model is a general metric space. This construction encompasses the classical Sobol indices [15] and their vectorial generalization [10] as well as some indices based on the whole distribution, namely the Cramér-von-Mises indices [12]. In addition, we propose an estimation procedure that ensures strong consistency and asymptotic normality at a cost of $2N$ calls to the computer code with a rate of convergence $\sqrt{N}$. As soon as $m \geqslant 1$, this new methodology appears to be more efficient than the so-called Pick-Freeze estimation procedure.

References

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

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

[3] E. Borgonovo, G. B. Hazen, V. R. R. Jose, and E. Plischke. Probabilistic sensitivity measures as information value. *European Journal of Operational Research*, 289(2):595–610.

[4] E. Borgonovo, G. B. Hazen, and E. Plischke. A common rationale for global sensitivity measures and their estimation. *Risk Analysis*, 36(10):1871–1895, 2016.

[5] M. Carrascal, M. Puigcerver, and P. Puig. Sensitivity of gaussian plume model to dispersion specifications. *Theoretical and Applied Climatology*, 48(2-3):147–157, 1993.

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

[7] E. De Rocquigny, N. Devictor, and S. Tarantola. *Uncertainty in industrial practice*. Wiley Chisterter England, 2008.

[8] J.-C. Fort, T. Klein, and N. Rachdi. New sensitivity analysis subordinated to a contrast. *Comm. Statist. Theory Methods*, 45(15):4349–4364, 2016.

[9] R. Fraiman, F. Gamboa, and L. Moreno. Sensitivity indices for output on a riemannian manifold. *International Journal for Uncertainty Quantification*, 10(4), 2020.

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

[11] 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.

[12] F. Gamboa, T. Klein, and A. Lagnoux. Sensitivity analysis based on Cramér–von Mises distance. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):522–548, 2018.

[13] T. Goda. Computing the variance of a conditional expectation via non-nested Monte Carlo. *Operations Research Letters*, 45(1):63 − 67, 2017.

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

[15] 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.

[16] W. Kahan. Pracniques: further remarks on reducing truncation errors. *Communications of the ACM*, 8(1):40, 1965.

[17] Z. Kala. Quantile-oriented global sensitivity analysis of design resistance. *Journal of Civil Engineering and Management*, 25(4):297–305, 2019.

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

[19] 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.

[20] L. Luyi, L. Zhenzhou, F. Jun, and W. Bintuan. Moment-independent importance measure of basic variable and its state dependent parameter solution. *Structural Safety*, 38:40–47, 2012.

[21] S. Mahanta, R. Chutia, D. Datta, and H. K. Baruah. Sensitivity analysis with reference to emission concentration of gaussian plume model. *International Journal of Energy, Information and Communications*, 3(2):45–52, 2012.

[22] A. Marrel, N. Saint-Geours, and M. De Lozzo. Sensitivity analysis of spatial and/or temporal phenomena. *Springer Handbook on Uncertainty Quantification, Springer*, pages 1327–1357, 2017.

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

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

[25] A. B. Owen. Better estimation of small Sobol' sensitivity indices. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 23(2):1–17, 2013.

[26] K. Pearson. On the partial correlation ratio. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 91(632):492–498, 1915.

[27] S. Pouget, M. Bursik, P. Singla, and T. Singh. Sensitivity analysis of a one-dimensional model of a volcanic plume with particle fallout and collapse behavior. *Journal of Volcanology and Geothermal Research*, 326:43–53, 2016.

[28] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

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

[30] 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.

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

[32] S. L. Sobolev. *Partial Differential Equations of Mathematical Physics: International Series of Monographs in Pure and Applied Mathematics*. Elsevier, 2016.

[33] K. Soetaert and F. Meysman. R-package reactran: Reactive transport modelling in r. *Environ. Model. Softw.*, 32:49–60, 2012.

[34] J. M. Stockie. The mathematics of atmospheric dispersion modeling. *Siam Review*, 53(2):349–372, 2011.

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

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

[37] V. A. Volpert. *Elliptic partial differential equations*, volume 1. Springer, 2011.