A Bayesian approach for global sensitivity analysis of (multi-fidelity) computer codes

Loic Le Gratiet, Claire Cannamela, Bertrand Iooss

A Bayesian approach for global sensitivity analysis of (multi-fidelity) computer codes

Loic Le Gratiet [†] [‡], Claire Cannamela [‡], Bertrand Iooss [§] [⋆]

[†] Université Paris Diderot 75205 Paris Cedex 13, France

[‡] CEA, DAM, DIF, F-91297 Arpajon, France

[§] EDF R&D, 6 quai Watier 78401 Chatou, France

[⋆] Institut de Mathématiques de Toulouse, 31062 Toulouse, France

July 8, 2013

Abstract

Complex computer codes are widely used in science and engineering to model physical phenomena. Furthermore, it is common that they have a large number of input parameters. Global sensitivity analysis aims to identify those which have the most important impact on the output. Sobol indices are a popular tool to perform such analysis. However, their estimations require an important number of simulations and often cannot be processed under reasonable time constraint. To handle this problem, a Gaussian process regression model is built to surrogate the computer code and the Sobol indices are estimated through it. The aim of this paper is to provide a methodology to estimate the Sobol indices through a surrogate model taking into account both the estimation errors and the surrogate model errors. In particular, it allows us to derive non-asymptotic confidence intervals for the Sobol index estimations. Furthermore, we extend the suggested strategy to the case of multi-fidelity computer codes which can be run at different levels of accuracy. For such simulators, we use an extension of Gaussian process regression models for multivariate outputs.

Keywords: Sensitivity analysis, Gaussian process regression, multi-fidelity model, complex computer codes, Sobol index, Bayesian analysis.

1 Introduction

Complex computer codes commonly have a large number of input parameters for which we want to measure their importance on the model response. We focus on the Sobol indices [26], [23] and [27] which are variance-based importance measures coming from the Hoeffding-Sobol decomposition [8]. We note that the presented sensitivity analysis holds when the input

parameters are independent. For an analysis with dependent inputs, the reader is referred to the articles [16] and [2].

A widely used method to estimate the Sobol indices are the Monte-Carlo based methods. They allow for quantifying the errors due to numerical integrations (with a Bootstrap procedure in a non-asymptotic case [1] and [12] or thanks to asymptotic normality properties in an asymptotic case [11]). However, the estimation of the Sobol indices by sampling methods requires a large number of simulations that are sometimes too costly and time-consuming. A popular method to overcome this difficulty is to build a mathematical approximation of the code input/output relation [18] and [9].

We deal in this paper with the use of kriging and multi-fidelity co-kriging models to surrogate the computer code. The reader is referred to the books [28], [24] and [21] for an overview of kriging methods for computer experiments. A pioneering article dealing with the kriging approach to perform global sensitivity analysis is the one of Oakley and O'Hagan [19]. Their method is also investigated in [17]. The strength of the suggested approach is that it allows for inferring from the surrogate model uncertainty about the Sobol index estimations. However, it does not use Monte-Carlo integrations and it does not take into account the numerical errors due to the numerical integrations. Furthermore, the implementation of the method is complex for general covariance kernels. Another flaw of the method presented in [19] and [17] is that it is not based on the exact definition of Sobol indices (it uses the ratio of two expectations instead of the expectation of a ratio).We note that a bootstrap procedure can also be used to evaluate the impact of the surrogate model uncertainty on the Sobol index estimates as presented in [29]. However, this approach only focuses on the parameter estimation errors.

On the other hand, a method giving confidence intervals for the Sobol index estimations and taking into account both the meta-model uncertainty and the numerical integration errors is suggested in [12]. They consider Monte-Carlo integrations to estimate the Sobol indices (see [26], [25] and [11]) instead of numerical integrations and they infer from the sampling errors thanks to a bootstrap procedure. Furthermore, to deal with the meta-model error, they consider an upper bound on it. In the kriging case they use the kriging variance up to a multiplicative constant as upper bound. Nevertheless, this is a rough upper bound which considers the worst error on a test sample. Furthermore, this method does not allow for inferring from the meta-model uncertainty about the Sobol index estimations.

We propose in this paper a method combining the approaches presented in [19] and [12]. As in [19] we consider the code as a realization of a Gaussian process. Furthermore, we use the method suggested in [12] to estimate the Sobol indices with Monte-Carlo integrations. Therefore, we can use the bootstrap method presented in [1] to infer from the sampling error on the Sobol index estimations. Furthermore, contrary to [19] and [17] we deal with the exact definition of Sobol indices. Consequently, we introduce non-asymptotic certified Sobol index estimations, i.e. with confidence intervals which take into account both the surrogate model error and the numerical integration errors.

Finally, the suggested approach is extended to a multi-fidelity co-kriging model. It allows for approximating a computer code using fast and coarse versions of it. The suggested multi-fidelity models is derived from the original one proposed in [13]. We note that the use of co-kriging model to deal with multi-fidelity codes have been largely investigated during this last decade (see [14], [7], [22] and [20]). A definition of Sobol indices for multi-fidelity computer codes is presented in [10]. However, their approach is based on tabulated biases between fine and coarse codes and does not allow for inferring from the meta-model uncertainty. The co-

kriging model fixes these weakness since it allows for considering general forms for the biases and for inferring from the surrogate model error.

This paper is organized as follows. First we introduce in Section 2 the so-called Sobol indices. Then, we present in Section 3 the kriging-based sensitivity analysis suggested by [19]. Our approach is developed in Section 4. In particular, we give an important result allowing for effectively sampling with respect to the kriging predictive distribution in Subsection 4.3. Finally, we extend in Section 5 the presented approaches to multi-fidelity co-kriging models. We highlight that we present in Subsection 5.2 a method to sampling with respect to the multi-fidelity predictive distribution. In this case the predictive distribution is not anymore Gaussian. Numerical tests are performed in Section 6 and an industrial example is considered in Section 7. A conclusion synthesizes this work in the last section.

2 Global sensitivity analysis: the method of Sobol

We present in this section the method of Sobol for global sensitivity analysis [26]. It is inspired by the book of [23] giving an overview of classical sensitivity analysis methods.

2.1 Sobol variance-based sensitivity analysis

Let us consider the input parameter space $Q \subset \mathbb{R}^d$ such that $(Q, \mathcal{B}(Q))$ is a measurable product space of the form:
$$(Q, \mathcal{B}(Q)) = (Q_1 \times \cdots \times Q_d, \mathcal{B}(Q_1 \times \cdots \times Q_d))$$

where $\mathcal{B}$ is the Borelian σ-algebra and $Q_i \subset \mathbb{R}$ is a nonempty open set, for $i = 1, \ldots, d$. Furthermore, we consider a probability measure μ on $(Q, \mathcal{B}(Q))$, with values in $\mathbb{R}$ and of the form
$$\mu(x) = \mu_1(x^1) \otimes \cdots \otimes \mu_d(x^d)$$

The Hoeffding-Sobol decomposition (see [8]) states that any function $z(x) \in L^2_\mu(\mathbb{R}^d)$ can be decomposed into summands of increasing dimensionality in such way:

$$z(x) = z_0 + \sum_{i=1}^d z_i(x^i) + \sum_{1 \le i < j \le k} z_{ij}(x^i, x^j) + \cdots + z_{1,2,\ldots,d}(x^1, \ldots, x^d) = \sum_{u \in \mathcal{P}} z_u(x^u) \qquad (1)$$

where $\mathcal{P}$ is the collection of all subsets of $\{1, \ldots, d\}$ and x^u is a group of variables such that $x^u = (x^i)_{i \in u}$. Furthermore, the decomposition is unique if we consider the following property for every summand $u = (u_1, \ldots, u_k)_{1 \le k \le d}$, $1 \le u_i \le d$:

$$\int z_u(x^u) \, d\mu_{u_i}(x^{u_i}) = 0, \quad \forall i = 1, \ldots, k. \qquad (2)$$

Now, let us suppose that the inputs are a random vector $X = (X^1, \ldots, X^d)$ defined on the probability space $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ and with measure μ. Sobol [26] showed that the decomposition

(1) can be interpreted as conditional expectations as follows:

$$
\begin{aligned}
z_0 &= \mathbb{E}_X\left[z(X)\right] \\
z_i(X^i) &= \mathbb{E}_X\left[z(X)|X^i\right] - z_0 \\
z_{ij}(X^i, X^j) &= \mathbb{E}_X\left[z(X)|X^i, X^j\right] - z_i(X^i) - z_j(X^j) - z_0 \\
&\vdots \\
z_u(X^u) &= \mathbb{E}_X\left[z(X)|X^u\right] - \sum_{v \subset u} z_v(X^v)
\end{aligned}
$$

with $u \in \mathcal{P}$. From this scheme, we can naturally develop the variance-based sensitivity indices of Sobol. First, let us consider the total variance D of $z(x)$:

$$
D = \text{var}_X\left(z(X)\right) \tag{3}
$$

By squaring and integrating the decomposition (1), we obtain

$$
D = \sum_{i=1}^d D_i + \sum_{1 \leq i < j \leq d} D_{ij} + \cdots + D_{1,2,\dots,d} = \sum_{u \in \mathcal{P}} D_u. \tag{4}
$$

with $D_u = \text{var}_X\left(\mathbb{E}_X\left[z(X)|X^u\right]\right) - \sum_{v \subset u} \text{var}_X\left(\mathbb{E}_X\left[z(X)|X^v\right]\right)$. Finally, the Sobol sensitivity indices are given by

$$
S_u = \frac{D_u}{D} \tag{5}
$$

where $u \in \mathcal{P}$. We note that we have the following useful equality which allows for easily interpreting S_u as the part of variance of $z(x)$ due to x^u and not explained by x^v with $v \subset u$.

$$
1 = \sum_{i=1}^d S_i + \sum_{1 \leq i < j \leq d} S_{ij} + \cdots + S_{1,2,\dots,d} = \sum_{u \in \mathcal{P}} S_u. \tag{6}
$$

In particular, S_i is called the first-order sensitivity index for variable x^i. It measures the main effect of x^i on the output, i.e. the part of variance of $z(x)$ explained by the factor x^i. Furthermore, S_{ij} for $i \neq j$ is the second-order sensitivity index. It measures the part of variance of $z(x)$ due to x^i and x^j and not explained by the individual effects of x^i and x^j.

2.2 Monte-Carlo Based estimations of Sobol indices

Now, let us denote by $Q^{d_1} = Q_{i_1} \times \cdots \times Q_{i_{d_1}}$, $d_1 \leq d$, $\{i_1, \dots, i_{d_1}\} \in \mathcal{P}$ and $Q^{d_2} = Q_{j_1} \times \cdots \times Q_{j_{d_2}}$ such that $\{j_1, \dots, j_{d_2}\} = \{1, \dots, d\} \setminus \{i_1, \dots, i_{d_1}\}$. Analogously, we use the notation $X^{d_1} = (X^i)_{i \in \{i_1, \dots, i_{d_1}\}}$, $X^{d_2} = (X^j)_{j \in \{j_1, \dots, j_{d_2}\}}$, $\mu^{d_1} = \left(\bigotimes_{i \in \{i_1, \dots, i_{d_1}\}} \mu_i\right)$ and $\mu^{d_2} = \left(\bigotimes_{j \in \{j_1, \dots, j_{d_2}\}} \mu_j\right)$ where μ^{d_1} and μ^{d_2} are probability measures on $(Q^{d_1}, \mathcal{B}(Q^{d_1}))$ and $(Q^{d_2}, \mathcal{B}(Q^{d_2}))$. Consequently, we have the equalities $\mu = \mu^{d_1} \otimes \mu^{d_2}$, $Q = Q^{d_1} \times Q^{d_2}$ and $X = (X^{d_1}, X^{d_2})$ with $d = d_1 + d_2$.

We are interested in evaluating the closed sensitivity index:

$$
\mathcal{S}^{X^{d_1}} = \frac{V^{X^{d_1}}}{V} = \frac{\text{var}_X\left(\mathbb{E}_X\left[z(X)|X^{d_1}\right]\right)}{\text{var}_X\left(z(X)\right)} \tag{7}
$$

A first method would be to use d-dimensional numerical integrations to approximate the numerator and denominator of (7) as presented in [19] and [17]. Nonetheless, since d is large in general, this method leads to numerical issues and is computationally expensive. A second approach is to take advantage of the probabilistic interpretation of the Sobol indices and to use a Monte-Carlo procedure to evaluate the different integrals as presented in the forthcoming developments.

Proposition 1. *Let us consider the random vectors $(X, \tilde{X})$ with $X = (X^{d_1}, X^{d_2})$ and $\tilde{X} = (X^{d_1}, \tilde{X}^{d_2})$ where X^{d_1} is a random vector with measure μ^{d_1} on Q^{d_1}, X^{d_2} and $\tilde{X}^{d_2}$ are random vectors with measure μ^{d_2} on Q^{d_2} and $X^{d_2} \perp \tilde{X}^{d_2}$. We have the following equality:*

$$\text{var}_X \left(\mathbb{E}_X \left[z(X) | X^{d_1} \right] \right) = \text{cov}_X \left(z(X), z(\tilde{X}) \right) \tag{8}$$

$\mathcal{S}^{X^{d_1}}$ in equation (7) can thus be estimated by considering two random vectors $(X_i)_{i=1,\ldots,m}$ and $(\tilde{X}_i)_{i=1,\ldots,m}$, $m \in \mathbb{N}^*$ lying in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ such that $X_i \overset{\mathcal{L}}{=} X$ and $\tilde{X}_i \overset{\mathcal{L}}{=} \tilde{X}$ ($\overset{\mathcal{L}}{=}$ stands for an equality in distribution) and by using an estimator for the covariance $\text{cov}_X \left(z(X), z(\tilde{X}) \right)$.

Following this principle, Sobol [26] suggests the following estimator for the ratio in equation (7):

$$\frac{V_m^{X^{d_1}}}{V_m} = \frac{\frac{1}{m} \sum_{i=1}^m z(X_i) z(\tilde{X}_i) - \frac{1}{m} \sum_{i=1}^m z(X_i) \frac{1}{m} \sum_{i=1}^m z(\tilde{X}_i)}{\frac{1}{m} \sum_{i=1}^m z(X_i)^2 - \left(\frac{1}{m} \sum_{i=1}^m z(X_i) \right)^2} \tag{9}$$

This estimation is improved by [11] who propose the following estimator:

$$\frac{V_m^{X^{d_1}}}{V_m} = \frac{\frac{1}{m} \sum_{i=1}^m z(X_i) z(\tilde{X}_i) - \left(\frac{1}{2m} \sum_{i=1}^m z(X_i) + z(\tilde{X}_i) \right)^2}{\frac{1}{m} \sum_{i=1}^m z(X_i)^2 - \left(\frac{1}{2m} \sum_{i=1}^m z(X_i) + z(\tilde{X}_i) \right)^2} \tag{10}$$

In particular they demonstrate that the asymptotic variance in (10) is better than the one in (9) and they show that the estimator (10) is asymptotically efficient for the first order indices. The main weakness of the estimators (9) and (10) is that they are sometimes not accurate for small values of $V^{X^{d_1}}/V$ in (7). To tackle this issue, [25] propose the following estimator

$$\frac{V_m^{X^{d_1}}}{V_m} = \frac{\frac{1}{m} \sum_{i=1}^m z(X_i) z(\tilde{X}_i) - \frac{1}{m} \sum_{i=1}^m z(X_i) z(\tilde{\tilde{X}}_i)}{\frac{1}{m} \sum_{i=1}^m z(X_i)^2 - \left(\frac{1}{m} \sum_{i=1}^m z(X_i) \right)^2} \tag{11}$$

where $\tilde{\tilde{X}} = (\tilde{\tilde{X}}^{d_1}, \tilde{\tilde{X}}^{d_2})$, $\tilde{\tilde{X}}^{d_1} \overset{\mathcal{L}}{=} X^{d_1}$, $\tilde{\tilde{X}}^{d_1} \perp X^{d_1}$ and $(\tilde{\tilde{X}}_i)_{i=1,\ldots,m}$ is such that $\tilde{\tilde{X}}_i \overset{\mathcal{L}}{=} \tilde{\tilde{X}}$ for all $i = 1, \ldots, m$.

3 Kriging-based sensitivity analysis: a first approach

We present in this Section the approach suggested in [19] and [17] to perform global sensitivity analysis using kriging surrogate models. Then, we present an alternative method that allows for avoiding complex numerical integrations. Nevertheless, we will see that this approach does not provide a correct representations of the Sobol indices. We handle this problem in the next section.

3.1 A short introduction to kriging model

The principle of the kriging model is to consider that our prior knowledge about the code $z(x)$ can be modelled by a Gaussian process $Z(x)$ with mean $\mathbf{f}'(x)\boldsymbol{\beta}$ and covariance kernel $\sigma^2 r(x, \tilde{x})$ (see for example [24]). Then, the code $z(x)$ is approximated by a Gaussian process $Z_n(x)$ having the predictive distribution $[Z(x)|Z(\mathbf{D}) = \mathbf{z}^n, \sigma^2]$ where $\mathbf{z}^n$ are the known values of $z(x)$ at points in the experimental design set $\mathbf{D} = \{x^1, \ldots, x^n\}$, $x^i \in Q$, and σ^2 is the variance parameter:

$$Z_n(x) \sim \text{GP}\left(m_n(x), s_n^2(x, \tilde{x})\right) \tag{12}$$

where the mean $m_n(x)$ and the variance $s_n^2(x, \tilde{x})$ are given by:

$$m_n(x) = \mathbf{f}'(x)\hat{\boldsymbol{\beta}} + \mathbf{r}'(x)\mathbf{R}^{-1}\left(\mathbf{z}^n - \mathbf{F}\hat{\boldsymbol{\beta}}\right)$$

where $\mathbf{R} = [r(x_i, x_j)]_{i,j=1,\ldots,n}$, $\mathbf{r}'(x) = [r(x, x_i)]_{i=1,\ldots,n}$, $\mathbf{F} = [\mathbf{f}'(x_i)]_{i=1,\ldots,n}$ and

$$s_n^2(x, \tilde{x}) = \sigma^2 \left(1 - \begin{pmatrix} \mathbf{f}'(x) & \mathbf{r}'(x) \end{pmatrix} \begin{pmatrix} 0 & \mathbf{F}' \\ \mathbf{F} & \mathbf{R} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{f}(\tilde{x}) \\ \mathbf{r}(\tilde{x}) \end{pmatrix}\right)$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{F}'\mathbf{R}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{R}^{-1}\mathbf{z}^n$. The variance parameter σ^2 can be estimated with a restricted maximum likelihood method, i.e. $\hat{\sigma}^2 = (\mathbf{z}^n - \hat{\boldsymbol{\beta}}\mathbf{F})'\mathbf{R}^{-1}(\mathbf{z}^n - \hat{\boldsymbol{\beta}}\mathbf{F})/(n - p)$ where p is the size of $\boldsymbol{\beta}$.

3.2 Kriging-based Sobol index

The idea suggested in [19] and [17] is to substitute $z(x)$ with $Z_n(x)$ in equation (7):

$$\mathcal{S}_n^{X^{d_1}} = \frac{V_n^{X^{d_1}}}{V_n} = \frac{\text{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)}{\text{var}_X\left(Z_n(X)\right)} \tag{13}$$

Therefore, if we denote by $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$ the probability space where the Gaussian process $Z(x)$ lies, then the estimator $\mathcal{S}_n^{X^{d_1}}$ lies in $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$ (it is hence random). We note that $Z_n(X)$ is defined on the product probability space $(\Omega_X \times \Omega_Z, \sigma(\mathcal{F}_X \times \mathcal{F}_Z), \mathbb{P}_X \otimes \mathbb{P}_Z)$.

Nevertheless, the distribution of $\mathcal{S}_n^{X^{d_1}}$ is intractable and [19] and [17] focus on its mean and variance. More precisely, in order to derive analytically the Sobol index estimations they consider the following quantity:

$$\tilde{\mathcal{S}}_n^{X^{d_1}} = \frac{\mathbb{E}_Z\left[\text{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)\right]}{\mathbb{E}_Z\left[\text{var}_X\left(Z_n(X)\right)\right]} \tag{14}$$

where $\mathbb{E}_Z[.]$ stands for the expectation in the probability space $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$. Furthermore, the uncertainty on $\tilde{\mathcal{S}}_n^{X^{d_1}}$ is evaluated with the following quantity:

$$\sigma^2(\tilde{\mathcal{S}}_n^{X^{d_1}}) = \frac{\text{var}_Z\left(\text{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)\right)}{\mathbb{E}_Z\left[\text{var}_X\left(Z_n(X)\right)\right]^2} \tag{15}$$

As shown in [19] and [17], the equations (14) and (15) can be derived analytically through multi-dimensional integrals for the cases $d_1 = i$, $i = 1, \ldots, d$, i.e. for the first-order indices. Furthermore, with some particular formulations of $\mathbf{f}(x)$, $\mu(x)$ and $r(x, \tilde{x})$, these multi-dimensional integrals can be written as product of one-dimensional ones. We note that a method is suggested in [17] to generate samples of the numerator $\text{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)$ in (13). It allows for estimating the uncertainty of $\tilde{\mathcal{S}}_n^{X^{d_1}}$ in (14) without processing the complex numerical integrations involved in (15).

Discussions: The method suggested in [19] and [17] provides an interesting tool to perform sensitivity analysis of complex models. Nevertheless, in our opinion it suffers from the following flaws:

1. For general choice of $\mathbf{f}(x)$, $\mu(x)$ and $r(x, \tilde{x})$, the numerical evaluations of (14) and (15) can be very complex since it requires multi-dimensional integrals.

2. The method is derived for first-order sensitivity indices and cannot easily be extended to higher order indices.

3. The method allows for inferring from the surrogate model uncertainty about the sensitivity indices but does not allow for taking into account the numerical errors due to the multi-dimensional integral estimations.

4. The considered index expectation and deviation do not correspond to the real Sobol index ones since we obviously have

$$\frac{\mathbb{E}_Z\left[\text{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)\right]}{\mathbb{E}_Z\left[\text{var}_X\left(Z_n(X)\right)\right]} \neq \mathbb{E}_Z\left[\frac{\text{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)}{\text{var}_X\left(Z_n(X)\right)}\right]$$

and

$$\frac{\text{var}_Z\left(\text{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)\right)}{\mathbb{E}_Z\left[\text{var}_X\left(Z_n(X)\right)\right]^2} \neq \text{var}_Z\left(\frac{\text{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)}{\text{var}_X\left(Z_n(X)\right)}\right)$$

In the next subsection, we deal with the points 1, 2 and 3 by suggesting a Monte-Carlo sampling method to evaluate (14) and (15) instead of quadrature integrations. Nonetheless, we do not tackle the issue of point 4. To handle it, we suggest another method in Section 4.

3.3 Monte-Carlo estimations for the first approach

We present in this Subsection, another approach to deal with the evaluation of $\tilde{\mathcal{S}}_n^{X^{d_1}}$ in (14). Its principle simply consists in using the estimation methods suggested in Subsection 2.2 instead of quadrature integrations to compute $\mathbb{E}_Z\left[\text{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)\right]$ and $\mathbb{E}_Z\left[\text{var}_X\left(Z_n(X)\right)\right]$. We present the method with the estimator presented in [26]. The extension to those presented in [12] and [25] is straightforward. Let us substitute in the estimator presented in equation (9) the code $z(x)$ by the Gaussian process $Z_n(x)$:

$$\frac{V_{m,n}^{X^{d_1}}}{V_{m,n}} = \frac{\frac{1}{m}\sum_{i=1}^m Z_n(X_i)Z_n(\tilde{X}_i) - \frac{1}{m}\sum_{i=1}^m Z_n(X_i)\frac{1}{m}\sum_{i=1}^m Z_n(\tilde{X}_i)}{\frac{1}{m}\sum_{i=1}^m Z_n(X_i)^2 - \left(\frac{1}{m}\sum_{i=1}^m Z_n(X_i)\right)^2} \tag{16}$$

where the samples $(X_i)_{i=1,\dots,m}$ and $(\tilde{X}_i)_{i=1,\dots,m}$ are those introduced in Subsection 2.2. Therefore, $V_{m,n}^{X^{d_1}}/V_{m,n}$ is an estimator of $V^{X^{d_1}}/V$ (7) when we replace the true function $z(x)$ by its approximation $Z_n(x)$ built from n observations $\mathbf{z}^n$ of $z(x)$ and when we estimate the variances and the expectation involved in (7) by a Monte-Carlo method with m particles. To be clear in the remainder of this paper, we name as Monte-Carlo error the one due to the Monte-Carlo estimation and we name as meta-model error the one due to the substitution of $z(x)$ by a surrogate model. Furthermore, m will always denote the number of Monte-Carlo particles and n the number of observations used to build the surrogate model.

The strength of this formulation is that it gives closed form formulas for the evaluation of (14) for any choice of $\mathbf{f}(x)$, $\mu(x)$ and $r(x, \tilde{x})$ contrary to [19] and [17]. Furthermore, this method can directly be used for any order of Sobol indices which contrasts with the one presented in Subsection (3.2). Finally, unlike quadrature integrations, Monte-Carlo integrations allow for taking into account the numerical errors due to the integral evaluations. In particular, as presented in [1], the bootstrap method can be directly used to obtain confidence intervals on the Sobol indices.

We give in the following equation the Monte-Carlo estimation of $\tilde{\mathcal{S}}_n^{X^{d_1}}$ (14) corresponding to the kriging-based sensitivity indices presented in [19] and [17].

$$
\begin{aligned}
\tilde{\mathcal{S}}_{m,n}^{X^{d_1}} &= \frac{\mathbb{E}_Z\left[V_{m,n}^{X^{d_1}}\right]}{\mathbb{E}_Z\left[V_{m,n}\right]} \\
&= \frac{\frac{1}{m}\sum_{i=1}^m s_n^2(X_i, \tilde{X}_i) + m_n(X_i)m_n(\tilde{X}_i) - \frac{1}{m^2}\sum_{i,j=1}^m s_n^2(X_i, \tilde{X}_j) + m_n(X_i)m_n(\tilde{X}_j)}{\frac{1}{m}\sum_{i=1}^m s_n^2(X_i, X_i) + m_n(X_i)m_n(X_i) - \frac{1}{m^2}\sum_{i,j=1}^m s_n^2(X_i, X_j) + m_n(X_i)m_n(X_j)}
\end{aligned}
\tag{17}
$$

We note that the expression of $\tilde{\mathcal{S}}_{m,n}^{X^{d_1}}$ is different from the one obtained by estimating $V_m^{X^{d_1}}/V_m$ in (9) by replacing $z(x)$ by the predictive mean $m_n(x)$. In $\tilde{\mathcal{S}}_{m,n}^{X^{d_1}}$ we take into account the kriging predictive covariance through the terms $s_n^2(X_i, \tilde{X}_j)$ and $s_n^2(X_i, X_j)$.

4 Kriging-based sensitivity analysis: a second approach

We have highlighted at the end of Subsection 3.2 that one of the main flaws of the method presented by [19] is that it does not care about the exact definition of Sobol indices. We present in Subsection 4.1 another approach which deals with this issue. Then, in Subsection 4.3 we present an efficient method to compute it.

4.1 Kriging-based Sobol index estimation

First of all, in the previous section we have considered the variance of the main effects $V^{X^{d_1}}$ and the total variance V separately in equation (7). That is why the ratio of the expectations is considered as a sensitivity index in equation (14). In fact, in a Sobol index framework, we are interested in the ratio between $V^{X^{d_1}}$ and V. Therefore, we suggest to deal directly with the following estimator (see equation (16)):

$$
\mathcal{S}_{m,n}^{X^{d_1}} = \frac{V_{m,n}^{X^{d_1}}}{V_{m,n}}
\tag{18}
$$

which corresponds to the ratio $V^{X^{d_1}}/V$ after substituting the code $z(x)$ by the Gaussian process $Z_n(x)$ and estimating the terms $\mathrm{var}_X\left(\mathbb{E}_X\left[Z_n(X)|X^{d_1}\right]\right)$ and $\mathrm{var}_X\left(Z_n(X)\right)$ with a Monte-Carlo procedure as presented in [26]. We note that we can naturally adapt the presented estimator with the ones suggested by [25] and [11]. Nevertheless, we cannot obtain closed form expressions for the mean or the variance of this estimator. We thus have to numerically estimate them. We present in Algorithm 1 the suggested method to compute the distribution of $\mathcal{S}_{m,n}^{X^{d_1}}$.

Algorithm 1 Evaluation of the distribution of $\mathcal{S}_{m,n}^{X^{d_1}}$.

1: Build $Z_n(x)$ from the n observations $\mathbf{z}^n$ of $z(x)$ at points in $\mathbf{D}$ (see equation (12)).
2: Generate two samples $(x_i)_{i=1,\ldots,m}$ and $(\tilde{x}_i)_{i=1,\ldots,m}$ of the random vectors $(X_i)_{i=1,\ldots,m}$ and $(\tilde{X}_i)_{i=1,\ldots,m}$ with respect to the probability measure μ (see Proposition 1).
3: Set N_Z the number of samples for $Z_n(x)$ and B the number of bootstrap samples for evaluating the uncertainty due to Monte-Carlo integrations.
4: **for** $k = 1, \ldots, N_Z$ **do**
5:     Sample a realization $z_n(\mathbf{x})$ of $Z_n(\mathbf{x})$ with $\mathbf{x} = \{(x_i)_{i=1,\ldots,m}, (\tilde{x}_i)_{i=1,\ldots,m}\}$
6:     Compute $\hat{\mathcal{S}}_{m,n,k,1}^{X^{d_1}}$ thanks to the equation (16) from $z_n(\mathbf{x})$.
7:     **for** l=2,...,B **do**
8:         Sample with replacements two samples $\mathbf{u}$ and $\tilde{\mathbf{u}}$ from $\{(x_i)_{i=1,\ldots,m}\}$ and $\{(\tilde{x}_i)_{i=1,\ldots,m}\}$.
9:         Compute $\hat{\mathcal{S}}_{m,n,k,l}^{X^{d_1}}$ from $z_n(\mathbf{x}^B)$ with $\mathbf{x}^B = \{\mathbf{u}, \tilde{\mathbf{u}}\}$.
10:     **end for**
11: **end for**
    return $\left(\hat{\mathcal{S}}_{m,n,k,l}^{X^{d_1}}\right)_{\substack{k=1,\ldots,N_Z \\ l=1,\ldots,B}}$

The output $\left(\hat{\mathcal{S}}_{m,n,k,l}^{X^{d_1}}\right)_{\substack{k=1,\ldots,N_Z \\ l=1,\ldots,B}}$ of Algorithm 1 is a sample of size $N_Z \times B$ of $\mathcal{S}_{m,n}^{X^{d_1}}$ defined on $(\Omega_X \times \Omega_Z, \sigma(\mathcal{F}_X \times \mathcal{F}_Z), \mathbb{P}_X \times \mathbb{P}_Z)$ (i.e. $\mathcal{S}_{m,n}^{X^{d_1}}$ takes both into account the uncertainty of the metamodel and the one of the Monte-Carlo integrations). Then, we can deduce the following estimate $\bar{\mathcal{S}}_{m,n}^{X^{d_1}}$ for $\mathcal{S}_{m,n}^{X^{d_1}}$:

$$\bar{\mathcal{S}}_{m,n}^{X^{d_1}} = \frac{1}{N_Z B} \sum_{\substack{k=1,\ldots,N_Z \\ l=1,\ldots,B}} \hat{\mathcal{S}}_{m,n,k,l}^{X^{d_1}} \tag{19}$$

Furthermore, we can estimate the variance of $\mathcal{S}_{m,n}^{X^{d_1}}$ with

$$\hat{\sigma}^2(\mathcal{S}_{m,n}^{X^{d_1}}) = \frac{1}{N_Z B - 1} \sum_{\substack{k=1,\ldots,N_Z \\ l=1,\ldots,B}} \left(\hat{\mathcal{S}}_{m,n,k,l}^{X^{d_1}} - \bar{\mathcal{S}}_{m,n}^{X^{d_1}}\right)^2 \tag{20}$$

We note that the computational limitation of the algorithm is the sampling of the Gaussian process $Z_n(x)$ on $\mathbf{x} = \{(x_i)_{i=1,\ldots,m}, (\tilde{x}_i)_{i=1,\ldots,m}\}$. For that reason, we use a bootstrap procedure to evaluate the uncertainty of the Monte-Carlo integrations instead of sampling different realizations of the random vectors $(X_i)_{i=1,\ldots,m}$ and $(\tilde{X}_i)_{i=1,\ldots,m}$. Furthermore, the same bootstrap samples are used for the N_Z realizations of $Z_n(x)$.

Nevertheless, the number of Monte-Carlo particles m is very large in general - it is often around $m = 5000d$ - and it thus can be an issue to compute realizations of $Z_n(x)$ on $\mathbf{x}$. We present in the Subsection 4.3 an efficient method to deal with this point for any choice of $\mu(x)$, $\mathbf{f}(x)$ and $r(x, \tilde{x})$ and any index order. The idea to carry out an estimation of (18) from realizations of conditional Gaussian processes has already been suggested in [6]. The main contribution of this section is the procedure to balance the Monte-Carlo and the meta-model errors (see Subsection 4.2).

4.2 Determining the minimal number of Monte-Carlo particles m

We are interested here in quantifying the uncertainty of the considered estimator $\mathcal{S}_{m,n}^{X^{d_1}}$ (18). This estimator integrates two sources of uncertainty, the first one is related to the meta-model approximation and the second one is related to the Monte-Carlo integration. Therefore, we can decompose the variance of $\mathcal{S}_{m,n}^{X^{d_1}}$ as follows:

$$\mathrm{var}\left(\mathcal{S}_{m,n}^{X^{d_1}}\right) = \mathrm{var}_Z\left(\mathbb{E}_X\left[\mathcal{S}_{m,n}^{X^{d_1}}|Z_n(x)\right]\right) + \mathrm{var}_X\left(\mathbb{E}_Z\left[\mathcal{S}_{m,n}^{X^{d_1}}|(X_i,\tilde{X}_i)_{i=1,\dots,m}\right]\right)$$

where $\mathrm{var}_Z\left(\mathbb{E}_X\left[\mathcal{S}_{m,n}^{X^{d_1}}|Z_n(x)\right]\right)$ is the contribution of the meta-model on the variability of $\mathcal{S}_{m,n}^{X^{d_1}}$ and $\mathrm{var}_X\left(\mathbb{E}_Z\left[\mathcal{S}_{m,n}^{X^{d_1}}|(X_i,\tilde{X}_i)_{i=1,\dots,m}\right]\right)$ is the one of the Monte-Carlo integration. Furthermore, we have the following equalities:

$$\begin{cases} \mathrm{var}_Z\left(\mathbb{E}_X\left[\mathcal{S}_{m,n}^{X^{d_1}}|Z_n(x)\right]\right) & = & \mathbb{E}_X\left[\mathrm{var}_Z\left(\mathcal{S}_{m,n}^{X^{d_1}}|(X_i,\tilde{X}_i)_{i=1,\dots,m}\right)\right] \\ \mathrm{var}_X\left(\mathbb{E}_Z\left[\mathcal{S}_{m,n}^{X^{d_1}}|(X_i,\tilde{X}_i)_{i=1,\dots,m}\right]\right) & = & \mathbb{E}_Z\left[\mathrm{var}_X\left(\mathcal{S}_{m,n}^{X^{d_1}}|Z_n(x)\right)\right] \end{cases}$$

Therefore, from the sample $\left(\hat{\mathcal{S}}_{m,n,k,l}^{X^{d_1}}\right)_{\substack{k=1,\dots,N_Z \\ l=1,\dots,B}}$ we can estimate the part of variance of the estimator $\mathcal{S}_{m,n}^{X^{d_1}}$ related to the meta-modelling as follows:

$$\hat{\sigma}_{Z_n}^2(\mathcal{S}_{m,n}^{X^{d_1}}) = \frac{1}{B}\sum_{l=1}^{B}\frac{1}{N_Z-1}\sum_{k=1}^{N_Z}\left(\hat{\mathcal{S}}_{m,n,k,l}^{X^{d_1}}-\bar{\hat{\mathcal{S}}}_{m,n,l}^{X^{d_1}}\right)^2 \tag{21}$$

where $\bar{\hat{\mathcal{S}}}_{m,n,l}^{X^{d_1}} = \left(\sum_{i=1}^{N_Z}\mathcal{S}_{m,n,i,l}^{X^{d_1}}\right)/N_Z$. Furthermore, we can evaluate the part of variance of $\mathcal{S}_{m,n}^{X^{d_1}}$ related to the Monte-Carlo integrations as follows:

$$\hat{\sigma}_{MC}^2(\mathcal{S}_{m,n}^{X^{d_1}}) = \frac{1}{N_Z}\sum_{i=1}^{N_Z}\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\mathcal{S}}_{m,n,k,i}^{X^{d_1}}-\bar{\bar{\hat{\mathcal{S}}}}_{m,n,k}^{X^{d_1}}\right)^2 \tag{22}$$

where $\bar{\bar{\hat{\mathcal{S}}}}_{m,n,k}^{X^{d_1}} = \left(\sum_{i=1}^{B}\mathcal{S}_{m,n,k,i}^{X^{d_1}}\right)/B$.

Therefore, we have three different cases:

1. $\hat{\sigma}_{Z_n}^2(\mathcal{S}_{m,n}^{X^{d_1}}) \gg \hat{\sigma}_{MC}^2(\mathcal{S}_{m,n}^{X^{d_1}})$: the estimation error of $\mathcal{S}_{m,n}^{X^{d_1}}$ is essentially due to the meta-model error.

2. $\hat{\sigma}_{Z_n}^2(\mathcal{S}_{m,n}^{X^{d_1}}) \ll \hat{\sigma}_{MC}^2(\mathcal{S}_{m,n}^{X^{d_1}})$: the estimation error of $\mathcal{S}_{m,n}^{X^{d_1}}$ is essentially due to the Monte-Carlo error.

3. $\hat{\sigma}_{Z_n}^2(\mathcal{S}_{m,n}^{X^{d_1}}) \approx \hat{\sigma}_{MC}^2(\mathcal{S}_{m,n}^{X^{d_1}})$: the metamodel and the Monte-Carlo errors have the same contribution on the estimation error of $\mathcal{S}_{m,n}^{X^{d_1}}$.

Considering that the number of observations n is fixed, the minimal number of Monte-Carlo particles m is the one such that $\hat{\sigma}_{Z_n}^2(\mathcal{S}_{m,n}^{X^{d_1}}) \approx \hat{\sigma}_{MC}^2(\mathcal{S}_{m,n}^{X^{d_1}})$. We call it "minimal" since it is the one from which the Monte-Carlo error no longer dominates. Therefore, it should be the minimum number of required particles in practical applications. In practice, to determine it, we start with a small value of m and we increase it while the inequality $\hat{\sigma}_{Z_n}^2(\mathcal{S}_{m,n}^{X^{d_1}}) > \hat{\sigma}_{MC}^2(\mathcal{S}_{m,n}^{X^{d_1}})$ is true.

4.3 Sampling with respect to the kriging predictive distribution on large data sets

We saw in the previous subsection in Algorithm 1 that in a kriging framework, we can assess the distribution of the Sobol index estimators from realizations of the conditional Gaussian process $Z_n(x)$ at points in $\mathbf{x}$. Nevertheless, the size of the corresponding random vector could be important since it equals twice the number of Monte-Carlo particles m. Therefore, computing such realizations could lead to numerical issues such as ill-conditioned matrix or huge computational cost, especially if we use a Cholesky decomposition. Indeed, Cholesky decomposition complexity is $\mathcal{O}((2m)^3)$ and it often leads to ill-conditioned matrix since the predictive variance of $Z_n(x)$ is close to zero around the experimental design points.

Let us introduce the following unconditioned Gaussian process:

$$\tilde{Z}(x) \sim \mathrm{GP}(0, \sigma^2 r(x, \tilde{x})) \tag{23}$$

We have the following proposition [3]:

Proposition 2 (Sampling $Z_n(x)$ by kriging conditioning). *Let us consider the following Gaussian process:*

$$\tilde{Z}_n(x) = m_n(x) - \tilde{m}_n(x) + \tilde{Z}(x) \tag{24}$$

where $m_n(x)$ is the predictive mean of $Z_n(x)$ (12),

$$\tilde{m}_n(x) = \mathbf{f}'(x)\tilde{\boldsymbol{\beta}} + \mathbf{r}'(x)\mathbf{R}^{-1}\left(\tilde{Z}(\mathbf{D}) - \mathbf{F}\tilde{\boldsymbol{\beta}}\right) \tag{25}$$

and $\tilde{\boldsymbol{\beta}} = \left(\mathbf{F}'\mathbf{R}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{R}^{-1}\tilde{Z}(\mathbf{D})$. Then, we have

$$\tilde{Z}_n(x) \stackrel{\mathcal{L}}{=} Z_n(x)$$

where $Z_n(x)$ has the distribution of the Gaussian process $Z(x)$ of mean $\mathbf{f}'(x)\boldsymbol{\beta}$ and covariance kernel $\sigma^2 r(x, \tilde{x})$ conditioned by $\mathbf{z}^n$ at points in $\mathbf{D}$ (12). We note that we are in a Universal kriging case, i.e. we infer from the parameter $\boldsymbol{\beta}$. In a simple kriging case, the proposition remains true by setting $\tilde{\boldsymbol{\beta}} = 0$.

The strength of Proposition 2 is that it allows for sampling with respect to the distribution of $Z_n(x)$ by sampling an unconditioned Gaussian process $\tilde{Z}(x)$. The first consequence is that the conditioning of the covariance matrix is better since the variance of $\tilde{Z}(x)$ is not close to zero around points in $\mathbf{D}$. The second important consequence is that it allows for using efficient algorithms to compute realizations of $\tilde{Z}(x)$. For example, if $r(x, \tilde{x})$ is a stationary kernel, one can use the Bochner's theorem ([28] p.29) and the Fourier representation of $\tilde{Z}(x)$ to compute realizations of $\tilde{Z}(x)$ as presented in [28]. Furthermore, another efficient method is to use the Mercer's representation of $r(x, \tilde{x})$ (see [15] and [5]) and the Nyström procedure to approximate the Karhunen-Loeve decomposition of $\tilde{Z}(x)$ as presented in [21] p.98. One of the main advantage of the Karhunen-Loeve decomposition of $Z(x)$ is that it allows for sequentially adding new points to $\mathbf{x}$ without re-estimating the decomposition. Therefore, we can easily obtain the values of a given realization $z_n(x)$ of $Z_n(x)$ at new points not in $\mathbf{x}$. This interesting property will allow us to efficiently estimate the number m of Monte-Carlo particles such that the metamodel error and the Monte-Carlo estimation one are equivalent (see Subsection 4.2).

11

5 Multi-fidelity co-kriging based sensitivity analysis

Now let us suppose that we have s levels of code $(z_t(x))_{t=1,\ldots,s}$ from the less accurate one $z_1(x)$ to the most accurate one $z_s(x)$ and that we want to perform a Global sensitivity analysis for $z_s(x)$. We consider that, conditioning on the model parameters, $(z_t(x))_{t=1,\ldots,s}$ are realizations of Gaussian processes $(Z_t(x))_{t=1,\ldots,s}$. Furthermore, we consider the following multi-fidelity model $t = 2, \ldots, s$:

$$\begin{cases} Z_t(x) = \rho_{t-1} Z_{t-1}^*(x) + \delta_t(x) \\ Z_{t-1}^*(x) \perp \delta_t(x) \\ Z_{t-1}^*(x) \sim [Z_{t-1}(x)|\mathbf{Z}^{(t-1)} = \mathbf{z}^{(t-1)}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\sigma}^2] \end{cases} \tag{26}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_t)_{t=1,\ldots,s}$, $\boldsymbol{\rho} = (\rho_{t-1})_{t=2,\ldots,s}$, $\boldsymbol{\sigma}^2 = (\sigma_t^2)_{t=1,\ldots,s}$, $\mathbf{Z}^{(l-1)} = (Z_1(\mathbf{D}_1), \ldots, Z_{t-1}(\mathbf{D}_{t-1}))$, $\mathbf{z}^{(t-1)} = (z_1(\mathbf{D}_1), \ldots, z_{t-1}(\mathbf{D}_{t-1}))$ and $(\mathbf{D}_t)_{t=1,\ldots,s}$ are the experimental design sets at level t with n_t points and such that $\mathbf{D}_s \subseteq \mathbf{D}_{s-1} \subseteq \cdots \subseteq \mathbf{D}_1$. Further, conditioning on $\boldsymbol{\beta}_t$ and σ_t^2, $\delta_t(x)$ is a Gaussian process of mean $\mathbf{f}_t'(x)\boldsymbol{\beta}_t$ and covariance $\sigma_t^2 r_t(x, \tilde{x})$ and we use the convention $Z_1(x) = \delta_1(x)$. This model is analogous to the one presented in [13] except that $Z_{t-1}^*(x)$ has a conditional distribution.

We propose a Bayesian formulation of the model which allows to consider non-informative prior distributions for the the regression parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_t)_{t=1,\ldots,s}$ and the adjustment parameters $\boldsymbol{\rho} = (\rho_{t-1})_{t=2,\ldots,s}$. This leads to the following predictive distribution which integrates the posterior distributions of the parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_t)_{t=1,\ldots,s}$ and $\boldsymbol{\rho} = (\rho_{t-1})_{t=2,\ldots,s}$.

$$[Z_s(x)|\mathbf{Z}^{(s)} = \mathbf{z}^{(s)}, \boldsymbol{\sigma}^2] \tag{27}$$

The predictive distribution (27) is not Gaussian. Nevertheless, we can have closed form expressions for its mean $\mu_{n_s}^s(x)$ and covariance $k_{n_s}^s(x, \tilde{x})$:

$$\mu_{n_s}^s(x) = \hat{\rho}_{s-1} \mu_{n_{s-1}}^{s-1}(x) + \mu_{\delta_s}(x) \tag{28}$$

and:

$$k_{n_s}^s(x, \tilde{x}) = \widehat{\rho_{s-1}^2} k_{n_{s-1}}^{s-1}(x, \tilde{x}) + k_{\delta_s}(x, \tilde{x}) \tag{29}$$

where for $t = 1, \ldots, s$, $\begin{pmatrix} \hat{\rho}_{t-1} \\ \hat{\boldsymbol{\beta}}_t \end{pmatrix} = (\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{H}_t)^{-1}\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{z}_t$, $\mathbf{H}_t = [z_{t-1}(\mathbf{D}_t) \quad \mathbf{F}_t]$, $\mathbf{F}_t = \mathbf{f}_t'(\mathbf{D}_t)$, $\hat{\rho}_0 = 0$, $\mathbf{H}_1 = \mathbf{F}_1$, $\widehat{\rho_{t-1}^2} = \hat{\rho}_{t-1}^2 + \left[(\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{H}_t)^{-1}\right]_{[1,1]}$, $\widehat{\rho_0^2} = 0$,

$$\mu_{\delta_t}(x) = \mathbf{f}_t'(x)\hat{\boldsymbol{\beta}}_t + \mathbf{r}_t'(x)\mathbf{R}_t^{-1}(\mathbf{z}_t - \mathbf{F}_t\hat{\boldsymbol{\beta}}_t - \hat{\rho}_{t-1}z_{t-1}(\mathbf{D}^t)) \tag{30}$$

and

$$k_{\delta_t}(x, \tilde{x}) = \sigma_t^2 \left(r_t(x, \tilde{x}) - \begin{pmatrix} \mathbf{h}_t'(x) & \mathbf{r}_t'(x) \end{pmatrix} \begin{pmatrix} 0 & \mathbf{H}_t' \\ \mathbf{H}_t & \mathbf{R}_t \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{h}_t(\tilde{x}) \\ \mathbf{r}_t(\tilde{x}) \end{pmatrix} \right) \tag{31}$$

with $\mathbf{h}_t'(x) = [\mu_{n_{t-1}}^{t-1}(x) \quad \mathbf{f}_t'(x)]$ and $\mathbf{h}_1'(x) = \mathbf{f}_1'(x)$. We note that, in the mean of the predictive distribution, the regression and adjustment parameters have been replaced by their posterior means. Furthermore, the predictive variance integrates the uncertainty due to the regression and adjustment parameters.

We note that for each $t = 1, \ldots, s$, the variance parameter σ_t^2 is estimated with a restricted maximum likelihood method. Thus, its estimation is given by $\hat{\sigma}_t^2 = (\mathbf{z}_t - \mathbf{H}_t\hat{\boldsymbol{\beta}}_t)'\mathbf{R}_t^{-1}(\mathbf{z}_t - \mathbf{H}_t\hat{\boldsymbol{\beta}}_t)/(n_t - p_t - 1)$ where p_t is the size of $\boldsymbol{\beta}_t$.

We present in Subsection 5.1 the extension in a multi-fidelity framework of the first kriging-based Sobol index estimations presented in [19]. Then, we present in Subsection 5.2 the extension of our approach to perform co-kriging-based multi-fidelity sensitivity analysis.

5.1   Extension of the first approach for multi-fidelity co-kriging models

Let us denote by $\tilde{\mathcal{S}}_{m,s}^{X^{d_1}}$ the estimation of $V^{X^{d_1}}/V$ when we substitute $z_s(x)$ by $Z_{n,s}(x) \sim [Z_s(x)|\mathbf{Z}^{(s)} = \mathbf{z}^{(s)}, \sigma^2]$ and when we use the Sobol procedure to perform Monte-Carlo estimations (see [26] and Subsection 2.2). Then, the estimator suggested in [19] and [17] becomes in a multi-fidelity framework:

$$
\begin{aligned}
\tilde{\mathcal{S}}_{m,s}^{X^{d_1}} &= \frac{\frac{1}{m}\sum_{i=1}^m k_{n_s}^s(X_i, \tilde{X}_i) + \mu_{n_s}^s(X_i)\mu_{n_s}^s(\tilde{X}_i) - \frac{1}{m^2}\sum_{i,j=1}^m k_{n_s}^s(X_i, \tilde{X}_j) + \mu_{n_s}^s(X_i)\mu_{n_s}^s(\tilde{X}_j)}{\frac{1}{m}\sum_{i=1}^m k_{n_s}^s(X_i, X_i) + \mu_{n_s}^s(X_i)\mu_{n_s}^s(X_i) - \frac{1}{m^2}\sum_{i,j=1}^m k_{n_s}^s(X_i, X_j) + \mu_{n_s}^s(X_i)\mu_{n_s}^s(X_j)} \\
&= \frac{U}{D}
\end{aligned}
$$

where

$$
\begin{aligned}
U &= \frac{1}{m}\sum_{i=1}^m \left(\sum_{t=1}^s \left(\prod_{j=t}^{s-1} \widehat{\rho_j^2}\right) k_{\delta_t}(X_i, \tilde{X}_i) + \sum_{t,\tilde{t}=1}^s \left(\prod_{j=t}^{s-1} \hat{\rho}_j\right)\left(\prod_{j=\tilde{t}}^{s-1} \hat{\rho}_j\right) \mu_{\delta_t}(X_i)\mu_{\delta_{\tilde{t}}}(\tilde{X}_i) \right) \\
&\quad - \frac{1}{m^2}\sum_{i,j=1}^m \left(\sum_{t=1}^s \left(\prod_{j=t}^{s-1} \widehat{\rho_j^2}\right) k_{\delta_t}(X_i, \tilde{X}_j) + \sum_{t,\tilde{t}=1}^s \left(\prod_{j=t}^{s-1} \hat{\rho}_j\right)\left(\prod_{j=\tilde{t}}^{s-1} \hat{\rho}_j\right) \mu_{\delta_t}(X_i)\mu_{\delta_{\tilde{t}}}(\tilde{X}_j) \right)
\end{aligned}
$$

$$
\begin{aligned}
D &= \frac{1}{m}\sum_{i=1}^m \left(\sum_{t=1}^s \left(\prod_{j=t}^{s-1} \widehat{\rho_j^2}\right) k_{\delta_t}(X_i, X_i) + \sum_{t,\tilde{t}=1}^s \left(\prod_{j=t}^{s-1} \hat{\rho}_j\right)\left(\prod_{j=\tilde{t}}^{s-1} \hat{\rho}_j\right) \mu_{\delta_t}(X_i)\mu_{\delta_{\tilde{t}}}(X_i) \right) \\
&\quad - \frac{1}{m^2}\sum_{i,j=1}^m \left(\sum_{t=1}^s \left(\prod_{j=t}^{s-1} \widehat{\rho_j^2}\right) k_{\delta_t}(X_i, X_j) + \sum_{t,\tilde{t}=1}^s \left(\prod_{j=t}^{s-1} \hat{\rho}_j\right)\left(\prod_{j=\tilde{t}}^{s-1} \hat{\rho}_j\right) \mu_{\delta_t}(X_i)\mu_{\delta_{\tilde{t}}}(X_j) \right)
\end{aligned}
$$

and with the conventions $\hat{\rho}_0 = 0$, $\prod_{i=s}^{s-1} \hat{\rho}_i = 1$, $\widehat{\rho_0^2} = 0$, $\prod_{i=s}^{s-1} \widehat{\rho_i^2} = 1$, $\mu_{\delta_1}(x) = \mu_{n_1}^1(x)$ and $k_{\delta_1}(x, \tilde{x}) = k_{n_1}^1(x, \tilde{x})$.

We note that $\tilde{\mathcal{S}}_{m,s}^{X^{d_1}}$ is the analogous of $\tilde{\mathcal{S}}_{m,n}^{X^{d_1}}$ presented in Subsection 3.3. Furthermore, the developed expression of $\tilde{\mathcal{S}}_{m,s}^{X^{d_1}}$ allows for identifying the contribution of each code level t to the sensitivity index and the one of the covariance between the bias and the code at level t. We note that the covariance here is with respect to the distribution of the input parameters X. Nevertheless, as pointed out in previous sections, this estimator is based on a ratio of expectations and thus does not correspond to the true Sobol indices.

5.2   Extension of the second approach for multi-fidelity co-kriging models

We present here the extension of the approach presented in Section 4 to the multi-fidelity co-kriging model. Therefore, we aim to sample with respect to the distribution of

$$
\mathcal{S}_{m,s}^{X^{d_1}} = \frac{\frac{1}{m}\sum_{i=1}^m Z_{n,s}(X_i)Z_{n,s}(\tilde{X}_i) - \frac{1}{m}\sum_{i=1}^m Z_{n,s}(X_i)\frac{1}{m}\sum_{i=1}^m Z_{n,s}(\tilde{X}_i)}{\frac{1}{m}\sum_{i=1}^m Z_{n,s}(X_i)^2 - \left(\frac{1}{m}\sum_{i=1}^m Z_{n,s}(X_i)\right)^2}
\tag{32}
$$

13

which is the analog of $\mathcal{S}_{m,n}^{X^{d_1}}$ (16) in an univariate case when we substitute $z(x)$ with $Z_{n,s}(x) \sim [Z_s(x)|\mathbf{Z}^{(s)} = \mathbf{z}^{(s)}, \sigma^2]$. In fact, we can directly use Algorithm 1 by sampling realizations of $Z_{n,s}(x)$ instead of $Z_n(x)$. Moreover, the procedure presented in Subsection 4.2 to determine the optimal number of Monte-Carlo particles m is straightforward.

However, the distribution of $Z_{n,s}(x)$ is not Gaussian and thus the method presented in Subsection 4.3 cannot be used directly. In order to handle this problem, we consider the conditional distribution $[Z_s(x)|\mathbf{Z}^{(s)} = \mathbf{z}^{(s)}, \sigma^2, \boldsymbol{\rho}, \boldsymbol{\beta}]$, with $\boldsymbol{\sigma}^2 = (\sigma_t^2)_{t=1,\dots,s}$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_t)_{t=2,\dots,s}$ and $\boldsymbol{\rho} = (\rho_{t-1})_{t=2,\dots,s}$ which is Gaussian (note that we infer from $\boldsymbol{\beta}_1$). It corresponds to the distribution (27) conditioning by $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$. Furthermore, the Bayesian estimation of $(\rho_{t-1}, \boldsymbol{\beta}_t)$ gives us for all $t = 2, \dots, s$:

$$\begin{pmatrix} \rho_{t-1} \\ \boldsymbol{\beta}_t \end{pmatrix} \sim \mathcal{N}\left((\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{H}_t)^{-1}\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{z}_t, \sigma_t^2(\mathbf{H}_t'\mathbf{R}_t^{-1}\mathbf{H}_t)^{-1}\right) \tag{33}$$

From the recursive formulation given in (26), we can define the following Gaussian process having the desired distribution $[Z_s(x)|\mathbf{Z}^{(s)} = \mathbf{z}^{(s)}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \boldsymbol{\beta}]$:

$$Z_{n,s,\rho,\boldsymbol{\beta}}(x) = \left(\prod_{j=1}^{s-1} \rho_j\right) Z_{n,1}(x) + \sum_{t=2}^{s-1}\left(\prod_{j=t}^{s-1} \rho_j\right)\delta_{t,\rho_{t-1},\boldsymbol{\beta}_t}(x) + \delta_{s,\rho_{s-1},\boldsymbol{\beta}_s}(x) \tag{34}$$

where (see equations (30) and (31)):

$$Z_{n,1}(x) \sim \mathrm{GP}\left(\mu_{\delta_1}(x), k_{\delta_1}(x, \tilde{x})\right) \tag{35}$$

and for $t = 2, \dots, s$:

$$\delta_{t,\rho_{t-1},\boldsymbol{\beta}_t}(x) \sim \mathrm{GP}\left(\mu_{t,\rho_{t-1},\boldsymbol{\beta}_t}(x), k_{t,\rho_{t-1},\boldsymbol{\beta}_t}(x, \tilde{x})\right) \tag{36}$$

with $\mu_{t,\rho_{t-1},\boldsymbol{\beta}_t}(x) = \mathbf{r}_t'(x)\mathbf{R}_t^{-1}(\mathbf{z}_t - \mathbf{F}_t\boldsymbol{\beta}_t - \rho_{t-1}z_{t-1}(\mathbf{D}^t))$, $\left((\delta_{t,\rho_{t-1},\boldsymbol{\beta}_t}(x))_{t=2,\dots,s}, Z_{n,1}(x)\right)$ independent and

$$k_{t,\rho_{t-1},\boldsymbol{\beta}_t}(x, \tilde{x}) = \sigma_t^2\left(r_t(x, \tilde{x}) - \mathbf{r}_t'(x)\mathbf{R}_t^{-1}\mathbf{r}_t(\tilde{x})\right).$$

Therefore, we can deduce the following algorithm to compute a realization $z_{n,s}(x)$ of $Z_{n,s}(x) \sim [Z_s(x)|\mathbf{Z}^{(s)} = \mathbf{z}^{(s)}, \sigma^2]$.

Algorithm 2 provides an efficient tool to sample with respect to the distribution $[Z_s(x)|\mathbf{Z}^{(s)} = \mathbf{z}^{(s)}, \sigma^2]$. Then, from each sample we can estimate the Sobol indices with a Monte-Carlo procedure. Naturally, we can easily use a bootstrap procedure to take into account the uncertainty due to the Monte-Carlo scheme. Furthermore, we see in Algorithm 2 that once a sample of $[Z_s(x)|\mathbf{Z}^{(s)} = \mathbf{z}^{(s)}, \sigma^2]$ is available, a sample for each distribution $[Z_t(x)|\mathbf{Z}^{(t)} = \mathbf{z}^{(t)}, \sigma^2]$, $t = 1, \dots, s-1$ is also available. Therefore, we can directly quantify the difference between the Sobol indices at a level t and the ones at another level $\tilde{t}$.

6 Numerical illustrations on an academic example

We illustrate here the kriging-based sensitivity analysis suggested in Section 4. We remind that the aim of this approach is to perform a sensitivity index taking into account both the uncertainty related to the surrogate modeling and the one related to the Monte-Carlo integrations. Let us consider the Ishigami function:

$$z(x_1, x_2, x_3) = \sin(x_1) + 7\sin(x_2)^2 + 0.1x_3^4\sin(x_1),$$

14

Algorithm 2 Sampling with respect to the predictive distribution $[Z_s(x)|\mathbf{Z}^{(s)} = \mathbf{z}^{(s)}, \sigma^2]$.

1: Generate a sample $z_{n,1}(x)$ with respect to (35) thanks to the method presented in Proposition 2 in the universal kriging case.

2: Set $z_{n,s}(x) = z_{n,1}(x)$.

3: **for** t=2,...,s **do**

4:    Generate a sample $\begin{pmatrix} \rho^*_{t-1} \\ \boldsymbol{\beta}^*_t \end{pmatrix}$ with respect to (33).

5:    Conditionally to $\begin{pmatrix} \rho^*_{t-1} \\ \boldsymbol{\beta}^*_t \end{pmatrix}$, generate a sample $\delta^*_{t,\rho^*_{t-1},\boldsymbol{\beta}^*_t}(x)$ with respect to (36) thanks to the method presented in Proposition 2 in the simple kriging case.

6:    Set $z_{n,s}(x) = \rho^*_{t-1}z_{n,s}(x) + \mu^*_{t,\rho^*_{t-1},\boldsymbol{\beta}^*_t}(x)$.

7: **end for**

   return $z_{n,s}(x)$.

where μ_i is uniform on $[-\pi, \pi]$, $i = 1, 2, 3$. We are interested in the first order sensitivity indices theoretically given by

$$(S_1, S_2, S_3) = (0.314, 0.442, 0).$$

This section is organized as follows. First, in Subsection 6.1 we compare the Sobol index estimator $\hat{\mathcal{S}}^{X^{d_1}}_{m,n}$ (17) proposed by [19], the suggested one given by the mean of $\mathcal{S}^{X^{d_1}}_{m,n}$ (18) and the usual one which consists in substituting $z(x)$ by the predictive mean $m_n(x)$ (12) in (10). Then, in sections 6.3, 6.4 and 6.5 we deal with the approach presented in Section 4. In particular, we show that this approach is relevant to perform an uncertainty quantification taking into account both the uncertainty of the meta-modeling and the one of the Monte-Carlo integrations. We note that the construction of the surrogate models used in sections 6.3, 6.4 and 6.5 is presented in Section 6.2.

6.1 Comparison between the different methods

The aim of this subsection is to perform a numerical comparison between $\tilde{\mathcal{S}}^{X^{d_1}}_{m,n}$ (17), the empirical mean of $\mathcal{S}^{X^{d_1}}_{m,n}$ given in Equation (19) and the following estimator (see (10)):

$$\check{\mathcal{S}}^{X^{d_1}}_{m,n} = \frac{\frac{1}{m}\sum_{i=1}^m m_n(X_i)m_n(\tilde{X}_i) - \left(\frac{1}{2m}\sum_{i=1}^m m_n(X_i) + m_n(\tilde{X}_i)\right)^2}{\frac{1}{m}\sum_{i=1}^m m_n(X_i)^2 - \left(\frac{1}{2m}\sum_{i=1}^m m_n(X_i) + m_n(\tilde{X}_i)\right)^2}. \tag{37}$$

We note that the empirical mean $\bar{\mathcal{S}}^{X^{d_1}}_{m,n}$ of $\mathcal{S}^{X^{d_1}}_{m,n}$ is evaluated thanks to Algorithm 1, with $N_Z = 500$ and $B = 1$:

$$\bar{\mathcal{S}}^{X^{d_1}}_{m,n} = \frac{1}{N_Z} \sum_{k=1,...,N_Z} \hat{\mathcal{S}}^{X^{d_1}}_{m,n,k,1},$$

and for $\tilde{\mathcal{S}}^{X^{d_1}}_{m,n}$ and $\mathcal{S}^{X^{d_1}}_{m,n}$ we use the Monte-Carlo estimator (10) suggested in [11] (it is the one used in (37)). Then for the comparison, different sizes of the learning sample are considered ($n = 40, 50, 60, 70, 90, 120, 150, 200$ observations) and we randomly build 100 Latin Hypercube Samples (LHS) for each size of the learning sample. From these experimental design sets, we

build kriging models with a constant trend β and a tensorised 5/2-Matérn kernel. Furthermore, the characteristic length scales $(\theta_i)_{i=1,2,3}$ are estimated with a maximum likelihood procedure for each design set. The Nash-Sutcliffe model efficiency coefficient (sometimes called the predictivity coefficient Q^2),

$$Eff_n = 1 - \frac{\sum_{x \in T}(m_n(x) - z(x))^2}{\sum_{x \in T}(m_n(x) - \bar{z}(x))^2}, \quad \bar{z}(x) = \frac{1}{\#T}\sum_{x \in T} z(x),$$

of the different kriging models are evaluated on a test set T composed of 1,000 points uniformly spread on the input parameter space $[-\pi, \pi]^3$. The values of Eff_n are presented in Figure 1. The closer Eff is to 1, the more accurate is the model $m_n(x)$.
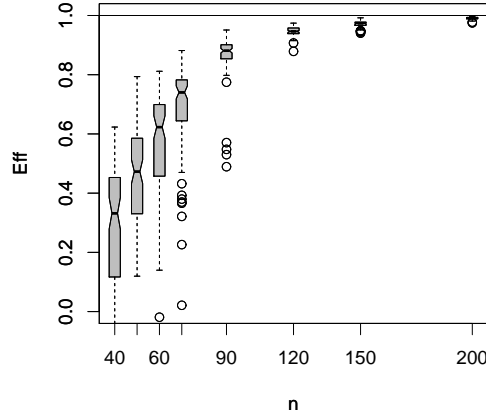
Figure 1: Convergence of the model efficiency when the number n of observations increases. 100 LHS are randomly sampled for each number of observations n.

Figure 2 illustrates the Sobol index estimates obtained with the three methods. We see in Figure 2 that the suggested estimator $\bar{\mathcal{S}}_{m,n}^{X^{d_1}}$ performs as well as the usual estimator $\check{\mathcal{S}}_{m,n}^{X^{d_1}}$ (37). In fact, as we will see in the next subsections, the strength of the suggested estimator is to provide more relevant uncertainty quantification. Finally, we see in Figure 2c that the estimator $\tilde{\mathcal{S}}_{m,n}^{X^{d_1}}$ (17) suggested in [19] seems to systematically underestimate the true value of the Sobol index for non-negligible index and when the model efficiency is low.

6.2 Model building and Monte-Carlo based estimator

For the numerical illustrations in sections 6.3 and 6.4, we use different kriging models built from different experimental design sets (optimized-LHS with respect to the centered L_2-discrepancy criterion, [4]) of size $n = 30, \ldots, 200$. Furthermore, for all kriging models, we consider a constant trend β and a tensorised 5/2-Matérn kernel (see [21]).

The characteristic length scales $(\theta_i)_{i=1,2,3}$ are estimated for each experimental design set by maximizing the marginal likelihood. Furthermore, the variance parameter σ^2 and the trend parameter β are estimated with a maximum likelihood method for each experimental
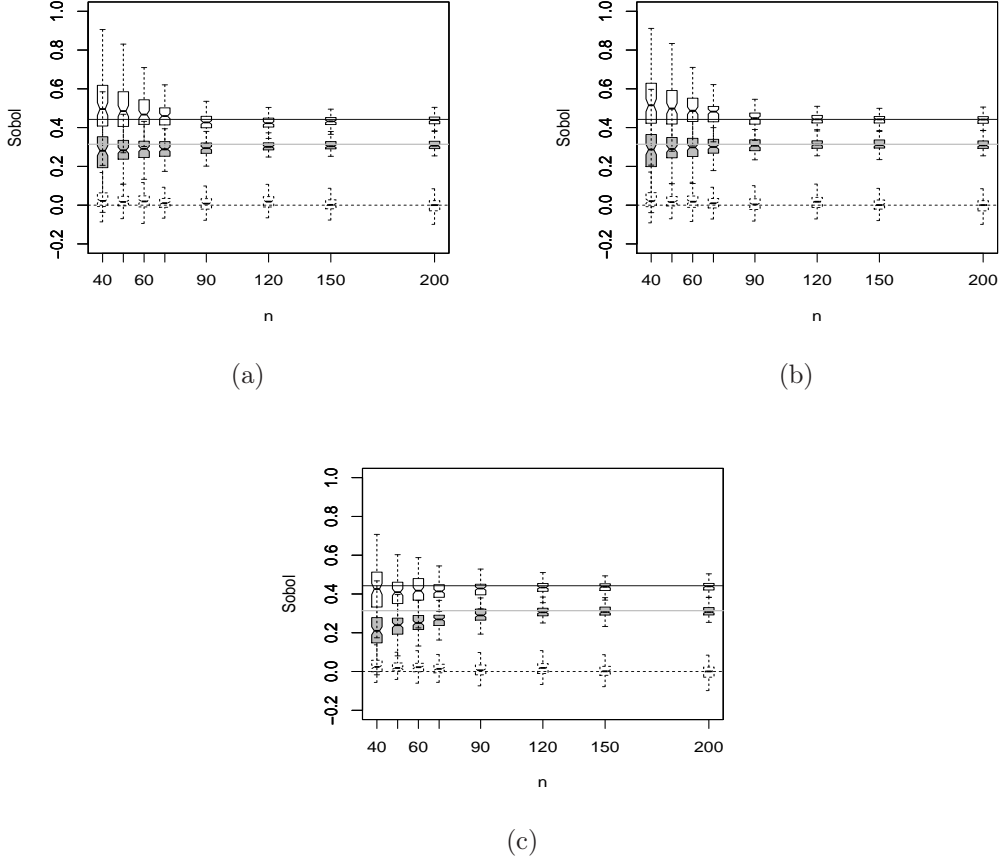
16

(a)

(b)

(c)

Figure 2: Comparison between three Sobol index estimators. The comparison are performed from 100 random LHS for each number of observations n. Figure (a) corresponds to the suggested Sobol estimator $\bar{\mathcal{S}}_{m,n}^{X^{d_1}}$ (see Section 4), Figure (b) corresponds to the usual (metamodel predictor only) estimator $\check{\mathcal{S}}_{m,n}^{X^{d_1}}$ (see Equation (37)) and Figure (c) corresponds to the estimator $\tilde{\mathcal{S}}_{m,n}^{X^{d_1}}$ suggested in [19]. The horizontal lines represent the true values of the Sobol indices (solid gray line: S_1; solid black line: S_2 and dashed black line: S_3)

design set too. Then for each n, the Nash-Sutcliffe model efficiency is evaluated on a test set composed of 10,000 points uniformly spread on the input parameter space $[-\pi, \pi]^3$. Figure 3 illustrates the estimated values of Eff_n with respect to the number of observations n.

Then, for estimating the Sobol indices, we use the Monte-Carlo based estimator given by (10). It has the strength to be asymptotically efficient for the first order indices (see [11]).

6.3 Sensitivity index estimates when n increases

Let us consider a fixed number of Monte-Carlo particles $m = 10,000$. The aim of this subsection is to quantify the part of the index estimator uncertainty related to the Monte-Carlo integrations and the one related to the surrogate modeling.

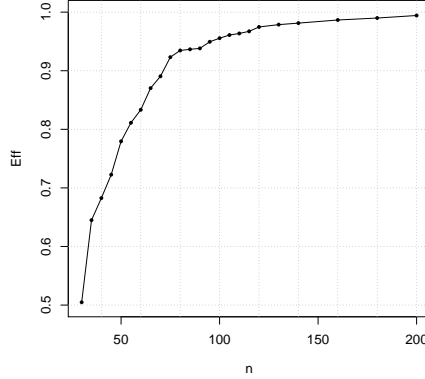To perform such analysis we use the procedure presented in Algorithm 1 with $B = 300$

Figure 3: Convergence of the model efficiency when the number n of observations increases. For each number of observations n, the experimental design set is an optimized-LHS with respect to the centered L_2-discrepancy.

bootstrap samples and $N_Z = 500$ realizations of $Z_n(x)$ (12). It results for each $i = 1, 2, 3$ a sample $\left(\hat{S}^i_{m,n,k,l}\right)$, $k = 1, \dots, N_Z$, $l = 1, \dots, B$, with respect to the distribution of the estimator obtained by substituting $z(x)$ with $Z_n(x)$ in (10).

Then, we estimate the 0.05 and 0.95 quantiles of $\left(\hat{S}^i_{m,n,k,1}\right)$, $k = 1, \dots, N_Z$ for each $i = 1, 2, 3$ with a bootstrap procedure. The resulting quantiles represent the uncertainty related to the surrogate modeling. Furthermore, we estimate the 2.50% and 97.50% quantiles of $\left(\hat{S}^i_{m,n,k,l}\right)$, $k = 1, \dots, N_Z$, $l = 1, \dots, B$ with a bootstrap procedure too. These quantiles represent the total uncertainty of the index estimator. Figure 4 illustrates the result of this procedure for different numbers of observations n. We see in Figure 4 that for small values of n, the error related to the surrogate modeling dominates. Then, when n increases, this error decreases and it is the one related to the Monte-Carlo integrations which is the largest. This emphasizes that it is worth to adapt the number of Monte-Carlo particles m to the number of observations n. Finally, we highlight that the equilibrium between the two types of uncertainty does not occur for the same n for the three indices. Indeed, it is around $n = 100$ for S_1, $n = 150$ for S_2 and around $n = 75$ for S_3. We observe that the smaller the index is, the larger its Monte-Carlo estimation error is.

6.4 Optimal Monte-Carlo resource when n increases

We saw in the previous subsection that the equilibrium between the error related to the Monte-Carlo integrations and the one related to the surrogate modeling depends on the considered sensitivity index. The purpose of this subsection is to determine this equilibrium for each index. To perform such analysis, we use the method presented in Subsection 4.2.

Let us consider a sample $\left(\hat{S}^i_{m,n,k,l}\right)$, $m = 30, \dots, 200$, $k = 1, \dots, N_Z$, $l = 1, \dots, B$, $i = 1, 2, 3$, generated with Algorithm 1 and using the Monte-Carlo estimator presented in (10). For each pair (m, n) we can evaluate the variance $\hat{\sigma}^2_{Z_n}\left(S^i_{m,n}\right)$, $i = 1, 2, 3$, related to the meta-modeling with Equation (21) and the variance $\hat{\sigma}^2_{MC}\left(S^i_{m,n}\right)$, $i = 1, 2, 3$, related to the

(a)   (b)
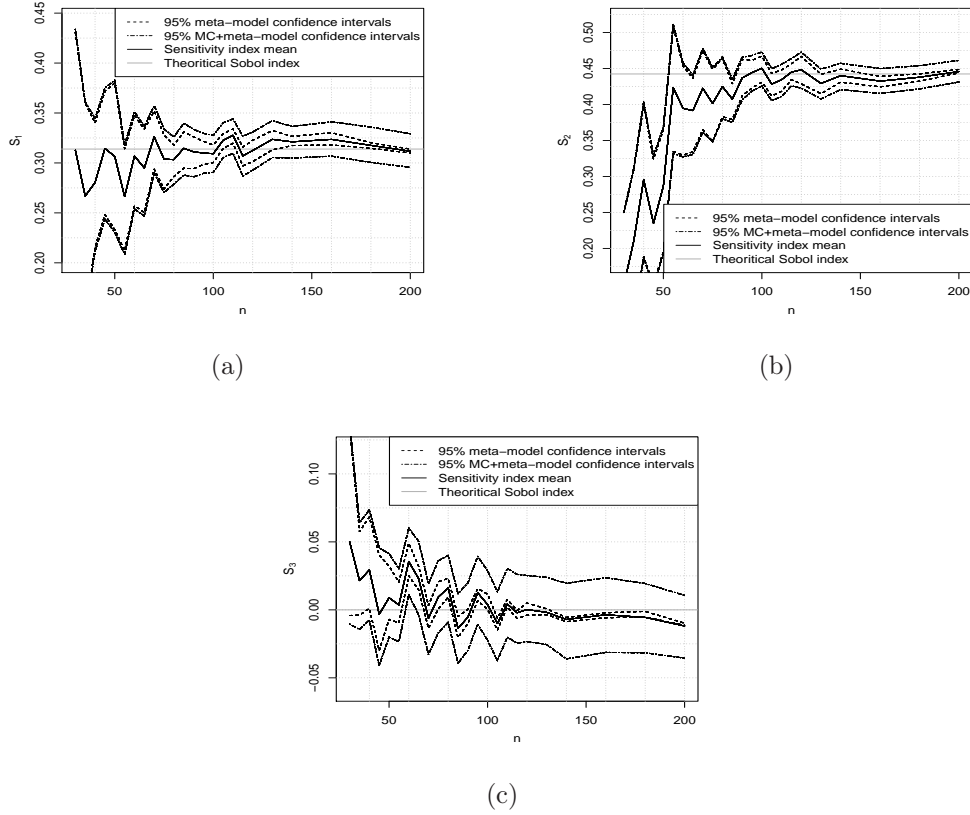

(c)

Figure 4: Sensitivity index estimates when n increases. The solid lines represent the means of the sensitivity index estimators. The dashed lines represent the 2.50% and 97.50% confidence intervals taking into account only the uncertainty related to the surrogate modeling. The dashed-dotted lines represent the 2.50% and 97.50% confidence intervals taking into account both the uncertainty related to the surrogate modeling and the one related to the Monte-Carlo integrations. The horizontal gray lines represent the true values of S_1 (a), S_2 (b) and S_3 (c).

Monte-Carlo integrations with Equation (22). We state that the equilibrium between the two types of uncertainty corresponds to the case

$$\hat{\sigma}^2_{Z_n}\left(S^i_{m,n}\right) = \hat{\sigma}^2_{MC}\left(S^i_{m,n}\right). \tag{38}$$

We present in Figure 5 the pairs (m, n) such that the equality (38) is satisfied. We see that the smaller is the sensitivity index, the more important is the number of particles m required to have the equilibrium. Furthermore, we note that the curve increases extremely quickly for the index $S_3 = 0$. Therefore, it could be unrealistic to consider the equilibrium for this case, especially when n is important (i.e. $n > 100$).

The presented analysis is of practical interest since it provides the appropriate number of Monte-Carlo particles m for the sensitivity index estimation in function of the number of observations n. Furthermore, in the framework of computer experiments, the observations are often time-consuming and n cannot be large. Therefore, we look for a number of particles m such that the variance $\hat{\sigma}^2_{Z_n}\left(S^i_{m,n}\right)$ related to the meta-modeling is smaller than the one of the
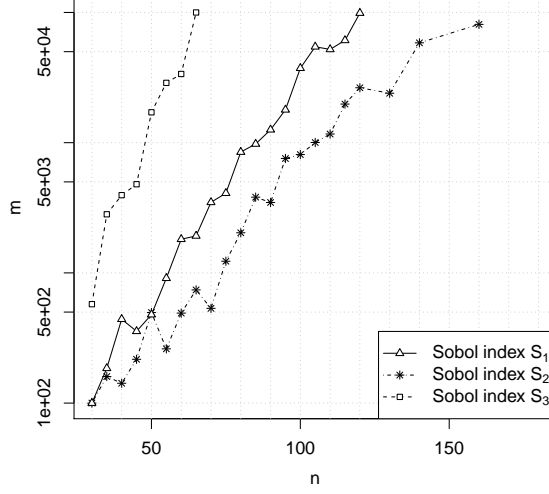
Figure 5: Relation between the number of observations n and the number of Monte-Carlo particles m such that the error related to the meta-modeling and the one related to the Monte-Carlo scheme have the same order of magnitude.

Monte-Carlo integration $\hat{\sigma}^2_{MC}\left(S^i_{m,n}\right)$. However, we saw that it could be unfeasible for some values of sensitivity index. In this case a compromise must necessarily be done.

6.5 Coverage rate of the suggested Sobol index estimator

Algorithm 1 in Subsection 4.1 allows for obtaining a sample $\left(\hat{S}^i_{m,n,k,l}\right)$, $k = 1,\ldots,N_Z$, $l = 1,\ldots,B$ of the estimator of S_i for each $i = 1,2,3$. The purpose of this subsection is to verify the relevance of the confidence intervals provided by $\left(\hat{S}^i_{m,n,k,l}\right)$. To perform such analysis, we generate 200 random LHS $(\mathbf{D}_{n,j})_{j=1,\ldots,200}$ for different numbers of observations n. For each $\mathbf{D}_{n,j}$, we build a kriging model with the procedure presented in Subsection 6.2 and we generate a sample $\left(\hat{S}^i_{m,n,k,l}\right)$, $k = 1,\ldots,N_Z$, $l = 1,\ldots,B$, with $B = 200$ and $N_Z = 300$. The efficiency of the different kriging models with respect to the number of observation n is presented in Figure 6. From this sample, we evaluate the 2.50% and 97.50% quantiles with a bootstrap procedure and we check if the true value of S_i is covered by these two quantiles. At the end of the procedure, the ratio between the number of confidence intervals covering the true value of S_i and the total number of confidence intervals (i.e. 200) has to be close to 95% for each n.

Furthermore, to perform the analysis we use different values of m according to the procedure presented in Subsection 4.2 for S_1 and S_2 (i.e. such that the variance related to the meta-modeling has the same order of magnitude than the one related to the Monte-Carlo integrations). For S_3, the number of Monte-Carlo particles m increases too quickly with respect to n to use the method presented in Subsection 4.2. Therefore we fix m to the values presented in Table 1. We note that the values of m for S_3 are larger than the ones for S_1 and S_2.
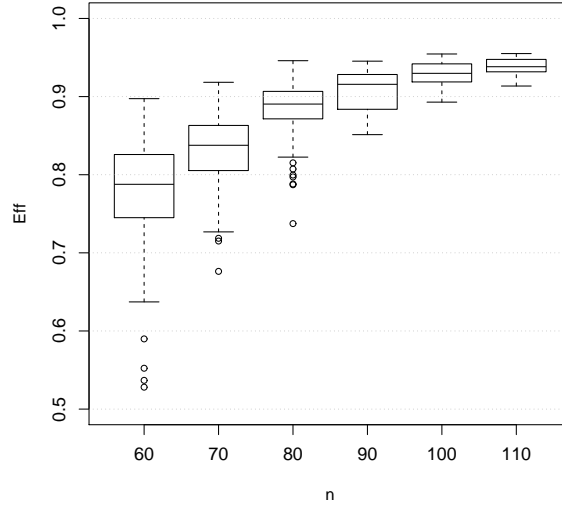
Figure 6: Convergence of the model efficiency when the number n of observations increases. For each number of observations n, 200 LHS are randomly sampled.

n	60	70	80	90	100	110
m	1,000	3,000	5,000	10,000	40,000	60,000

Table 1: Numbers of Monte-Carlo particles m for different values of the number of observations n for the estimation of S_3.

The empirical 95%-confidence intervals as a function of the number of observations n are presented in Figure 7. We study three cases:

1. The confidence intervals are built from $\left(\hat{S}^i_{m,n,k,l} \right)$, $k = 1, \ldots, N_Z$, $l = 1, \ldots, B$. Therefore, it takes into account both the uncertainty related to the meta-model and the one related to the Monte-Carlo estimations.

2. The confidence intervals are built from $\left(\hat{S}^i_{m,n,k,1} \right)$, $k = 1, \ldots, N_Z$. In this case, we do not use the bootstrap procedure to evaluate the uncertainty due to the Monte-Carlo procedure. Therefore, we only take into account the one due to the meta-model.

3. The confidence intervals are built from the estimator $\tilde{\mathcal{S}}^{X^{d_1}}_{m,n}$ (37) with a bootstrap procedure. Here, we estimate the Sobol indices with the kriging mean and we do not infer from the uncertainty of the meta-model. Therefore, we only take into account the uncertainty related to the Monte-Carlo estimations.

We see in Figure 7 that the confidence intervals provided by the approach presented in Section 4 are well evaluated for indices S_1 and S_3. Furthermore, they are underestimated when we take into account only the meta-model or the Monte-Carlo uncertainty. This highlights the

21

relevance of the suggested approach to perform uncertainty quantification on the Sobol index estimates. However, the coverage rate is underestimated for index S_2. This is even worst if we only consider the meta-model error. This may be due to a poor learning in the direction x_2 for the the surrogate model. This emphasizes that the suggested method is valid only if the kriging variance well represents the modeling error.
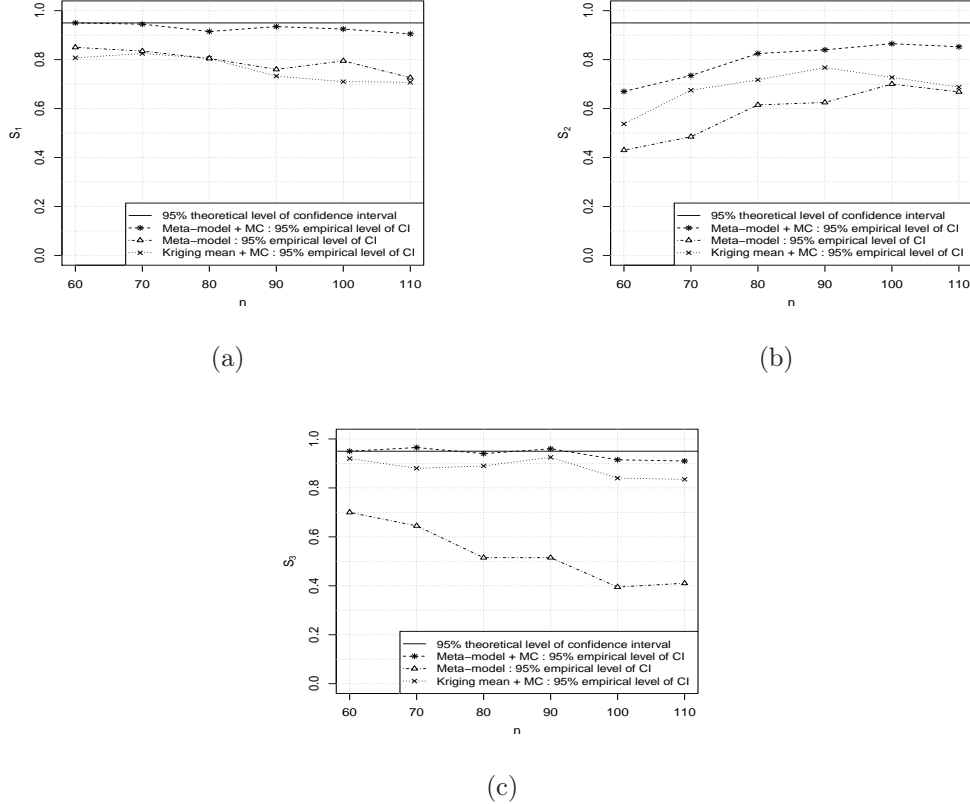

(a)   (b)

(c)

Figure 7: Empirical 95% confidence intervals with respect to the number of observations n for S_1 (a), S_2 (b) and S_3 (c). The empirical coverage rates are evaluated from 200 kriging models built from different random LHS.

7   Application of multi-fidelity sensitivity analysis

In this section, we illustrate the multi-fidelity co-kriging based sensitivity analysis presented in Section 5 on an example about a spherical tank under internal pressure.

7.1   Presentation of the problem

The scheme of the considered tank is presented in Figure 8. We are interested in the von Mises stress on the point labeled 1 in Figure 8. It corresponds to the point where the stress is maximal. The von Mises stresses are of interest since the material yielding occurs when they reach the critical yield strength.
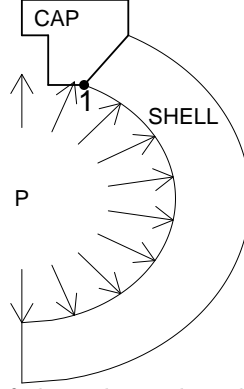
Figure 8: Scheme of the spherical tank under pressure.

The system illustrated in Figure 8 depends on 8 parameters:

- $P\,(MPa) \in [30, 50]$: the value of the internal pressure.

- $R_{int}\,(mm) \in [1500, 2500]$: the length of the internal radius of the shell.

- $T_{shell}\,(mm) \in [300, 500]$: the thickness of the shell.

- $T_{cap}\,(mm) \in [100, 300]$: the thickness of the cap.

- $E_{shell}\,(GPa) \in [63, 77]$: the Young's modulus of the shell material.

- $E_{cap}\,(GPa) \in [189, 231]$: the Young's modulus of the cap material.

- $\sigma_{y,shell}\,(MPa) \in [200, 300]$: the yield stress of the shell material.

- $\sigma_{y,cap}\,(MPa) \in [400, 800]$: the yield stress of the cap material.

The von Mises stress $z_2(x)$, $x = (P, R_{int}, T_{shell}, T_{cap}, E_{shell}, E_{cap}, \sigma_{y,shell}, \sigma_{y,cap})$ is provided by an finite elements code, called Aster, modelling the tank under pressure. The material properties of the shell correspond to high quality aluminums and the ones of the cap corresponds to steels from classical to high quality.

The cheaper version $z_1(x)$ of $z_2(x)$ is obtained by the 1D simplification of the tank corresponding to a perfect spherical tank, i.e. without the cap:

$$z_1(x) = \frac{3}{2} \frac{(R_{int} + T_{shell})^3}{(R_{int} + T_{shell})^3 - R_{int}^3} P$$

7.2 Multi-fidelity model building

We present here the construction of the model presented in Section 5.

First, we build two LHS design sets $\tilde{\mathbf{D}}_1$ and $\mathbf{D}_2$ of size $n_1 \times 8$ and $n_2 \times 8$ optimized with respect to the centered L_2-discrepancy criterion, with $n_1 = 100$ and $n_2 = 20$. We note that the input parameter x is normalized so that the measure $\mu(x)$ of the input parameters is uniform on $[0, 1]^8$. In order to respect the nested property for the experimental design sets, we remove from $\tilde{\mathbf{D}}_1$ the n_2 points that are the closest to those of $\mathbf{D}_2$ and we set that $\mathbf{D}_1$ is

the concatenation of $\mathbf{D}_2$ and $\tilde{\mathbf{D}}_1$. This procedure ensures that $\mathbf{D}_2 \subset \mathbf{D}_1$ without operating any transformation on $\mathbf{D}_2$.

Second, we run the expensive code $z_2(x)$ on points in $\mathbf{D}_2$ and the coarse code $z_1(x)$ on points in $\mathbf{D}_1$. The CPU time of the expensive code is around 1 minute. Furthermore, in order to have a fair illustration, we consider that the CPU time of the coarse code $z_1(x)$ is not negligible and we restrict its runs to $n_1 = 100$.

Third, we use tensorised 5/2-Matérn covariance kernels for $\sigma_1^2 r_1(x, \tilde{x})$ and $\sigma_2^2 r_2(x, \tilde{x})$ with characteristic length scales $(\theta_1^i)_{i=1,\dots,8}$ and $(\theta_2^i)_{i=1,\dots,8}$. Furthermore, we set that the regression functions are constants, i.e. $\mathbf{f}_1(x) = 1$ and $\mathbf{f}_2(x) = 1$.

The estimates of the characteristic length scales are given in Table 2.

$\hat{\boldsymbol{\theta}}_1$	1.71	1.38	1.97	1.98	1.98	1.99	1.95	1.41
$\hat{\boldsymbol{\theta}}_2$	1.83	1.89	0.5	1.93	1.93	0.64	1.89	0.79

Table 2: Maximum likelihood estimates of the characteristic length scales of the tensorised 5/2-Matérn covariance kernels used in the multi-fidelity co-kriging model. $\hat{\boldsymbol{\theta}}_1$ represents the estimates for the code level 1 and $\hat{\boldsymbol{\theta}}_2$ represents the ones for the bias between the code levels 1 and 2.

The estimates of the characteristic length scales given in Table 2 show that the model is very smooth. Then, Table 3 gives the posterior mean of the parameters $(\rho_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{\beta}_1$ and the restricted maximum likelihood estimate of σ_1^2 and σ_2^2.

$\hat{\boldsymbol{\beta}}_1$	148.67	$\hat{\sigma}_1^2$	495.63
$(\hat{\rho}_1, \hat{\boldsymbol{\beta}}_2)$	(0.92, 57.61)	$\hat{\sigma}_2^2$	551.07

Table 3: Posterior means of the trend parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and the adjustment parameter ρ_1 and maximum likelihood estimates of the variance parameters σ_1^2 and σ_2^2.

The parameter estimates presented in Table 3 show that there is an important bias between the cheap code and the expensive code since $\hat{\boldsymbol{\beta}}_2 \approx 58$ whereas the trend of the cheap code is $\hat{\boldsymbol{\beta}}_1 \approx 150$. In particular, it is greater than the standard deviation of the bias which is $\hat{\sigma}_2 \approx 23$. Then, the posterior mean of the adjustment parameter $\hat{\rho}_1 = 0.92$ does not indicate a perfect correlation between the two levels of code. Indeed, the estimated correlation between $z_2(x)$ and $z_1(x)$ is 0.77. Furthermore their estimated variance equals 1514 for $z_2(x)$ and 810 for $z_1(x)$. In fact, the adjustment parameter:

$$\rho_1 = \frac{\text{cov}(Z_2(x), Z_1(x))}{\text{var}(Z_1(x))}$$

represents both the correlation degree and the scale factor between the codes $z_2(x)$ and $z_1(x)$.

Finally, we can estimate the accuracy of the suggested model with a Leave-One-Out cross validation procedure. From the Leave-One-Out errors, we estimate the Nash-Sutcliffe model efficiency $Eff_{LOO} = 83\%$. This means that the suggested multi-fidelity co-kriging model explains 83% of the variability of the model. We note that the closer Eff_{LOO} is to 1, the more accurate is the model. Therefore, we have an excellent model despite the small number

of observations $n_2 = 20$ used for the expensive code $z_2(x)$. In order to strengthen this result, we test the multi-fidelity model on an external test set of $7,000$ points and the estimated efficiency is 86% which is even better.

7.3 Multi-fidelity sensitivity analysis

Now let us perform a multi-fidelity sensitivity analysis using the approach presented in Subsection 5.2. We are interested in the first-order sensitivity indices.

The principle of the method is to sample from the distribution (32) using Algorithm 2. We note that we use the Monte-Carlo estimator (10) instead of (9) since it is asymptotically efficient for the first-order indices We repeat the algorithm 2 to have $N_Z = 200$ realizations of the predictive distribution $[Z_2(x)|\mathbf{Z}^{(2)} = \mathbf{z}^{(2)}, \boldsymbol{\sigma}^2]$ and for each realization we generate $B = 150$ bootstrap samples. Furthermore, we choose $m = 20,000$ for the Monte-Carlo sampling size so that the error due to the Monte-Carlo integrations is negligible compared to the one due to the surrogate modelling (see Subsection 4.2 and 6.4).

Sensitivity analysis for the cheap code.
First, let us present the result of the sensitivity analysis for the cheap code. As emphasized in Subsection 5.2, once samples with respect to the distribution $[Z_2(x)|\mathbf{Z}^{(2)} = \mathbf{z}^{(2)}, \boldsymbol{\sigma}^2]$ are available, samples for $[Z_1(x)|\mathbf{Z}^{(1)} = \mathbf{z}^{(1)}, \sigma_1^2]$ are also available. Therefore, from them we can perform a sensitivity analysis as presented in Section 4. Moreover, from the explicit formula of $z_1(x)$ we expect that only the three variables P, R_{int} and T_{shell} have an impact on the output.

The result of the sensitivity analysis for the cheap code $z_1(x)$ is given in Figure 9. We see in Figure 9 that only the three parameters P, R_{int} and T_{shell} are influent as expected. Furthermore, the internal pressure is the most important parameter whereas the geometrical parameter R_{int} and T_{shell} have equivalent impact on the output. The sum of the first-order sensitivity index means informs us that 97% of the variability of the output is explained by the first-order indices. The interactions between the parameters are thus negligible. Further, we see that the confidence intervals are tight and that the uncertainty on the Sobol index estimator is essentially due to the Monte-Carlo integrations. This means that the model's error on the cheap code is very low.
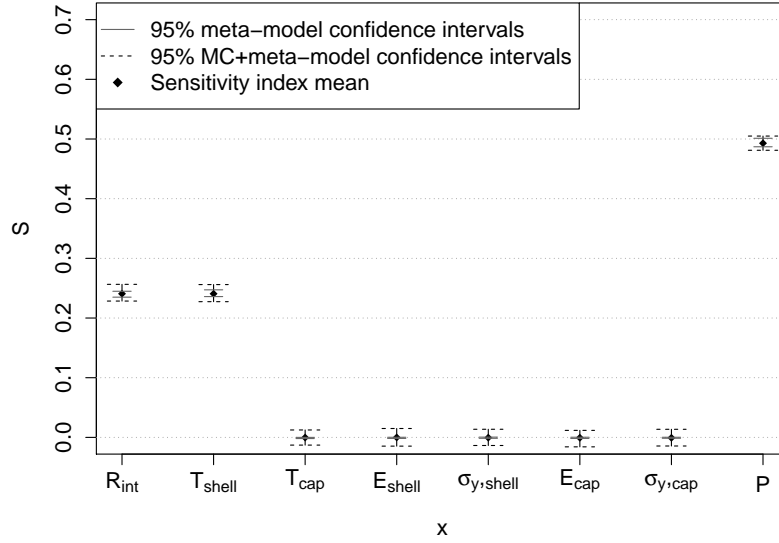
Figure 9: Kriging based sensitivity analysis for the cheap code. The diamonds represent the means of the first-order sensitivity index estimators, the solid gray lines represent the 95% confidence intervals taking into account only the meta-modelling uncertainty and the dashed lines represent the 95% confidence intervals taking into account the uncertainty due to both the Monte-Carlo integrations and the meta-modelling. The means and the confidence intervals are obtained with Algorithm 1.

Sensitivity analysis for the expensive code.

Second, we perform a sensitivity analysis for the expensive code $z_2(x)$ using the predictive distribution $[Z_2(x)|\mathbf{Z}^{(2)} = \mathbf{z}^{(2)}, \boldsymbol{\sigma}^2]$. The result of the analysis is presented in Figure 10.

We see in Figure 10 that the result of the sensitivity analysis for the expensive code is substantially different than the one for the cheap code. First, the importance measure of the parameters P, R_{int} and T_{shell} decreases although the internal pressure P remains the most influent parameter. Second, the material parameters E_{shell}, E_{cap}, $\sigma_{y,shell}$ and $\sigma_{y,cap}$ have still a negligible influence except for the rigidity of the cap E_{cap}. Then, the most noticeable difference is for the thickness of the cap T_{cap} which is now the second most important parameter. Then the sum of the index estimator means equals 96.7%. This means that the first-order indices still explain the main part of the model variability.

The hierarchy between the parameters can be easily interpreted. Indeed, the coarse code corresponds to the approximation of the tank without the cap. Therefore, it is natural that the parameters related to the cap have no influence. On the contrary, for the expensive code, we are interested in the von Mises stress at the junction between the cap and the shell. Consequently, the parameters related to the cap have now an influence. However, it was difficult to have a prior on the impact of the cap. We deduce from this analysis that it is in fact very important.

Influences of material parameters are negligible because the model stands in the regime of elastic deformations. It is thus physically coherent. In fact, they would be more influent in a plastic deformation regime which can occur for more important internal pressure P.

The other important differences between the two sensitivity analysis is the magnitude of

the confidence intervals. Indeed, we see in Figure 10 that, contrary to the cheap code, the confidence intervals for the sensitivity index estimators of the expensive code are very large. Therefore, despite the good multi-fidelity approximation of the expensive code, we have an important uncertainty on it. This is natural since we only use 20 runs of $z_2(x)$ to learn it. Finally, we note that the most important uncertainty is for T_{cap}. This is explained by the fact that this parameter is not considered by the cheap code. Therefore, $z_1(x)$ brings no information about T_{cap} contrary to R_{int}, T_{shell} and P.
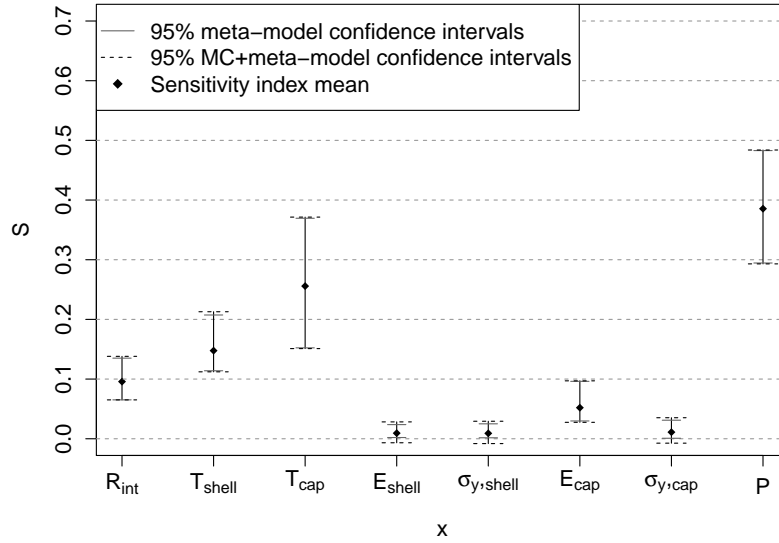

Figure 10: Co-kriging based sensitivity analysis for the expensive code. The diamonds represent the means of the first-order sensitivity index estimators, the solid gray lines represent the 95% confidence intervals taking into account only the meta-modelling uncertainty and the dashed lines represent the 95% confidence intervals taking into account the uncertainty due to both the Monte-Carlo integrations and the meta-modelling.

8 Conclusion

This paper deals with the sensitivity analysis of complex computer codes using Gaussian process regression. The purpose of the paper is to build Sobol index estimators taking into account both the uncertainty due to the surrogate modelling and the one due to the numerical evaluations of the variances and covariances involved in the Sobol index definition. The aim is to provide relevant confidence intervals for the index estimator.

To provide such estimators, we suggest a method which mixes a Gaussian process regression model with Monte-Carlo based integrations. From it, we can quantify the impact of both the Gaussian process regression and the Monte-Carlo procedure on the index estimator variability. In particular, we present a procedure to balance these two sources of uncertainty. Furthermore, we suggest numerical methods to avoid ill-conditioned problems and to easily handle the suggested index estimator.

Then, we propose an extension of the suggested approach for multi-fidelity computer codes. These codes have the characteristic to have coarser but computationally cheaper versions.

They are of practical interest since they allow for dealing with the problem of very expensive simulations. To deal with these codes, we use a multivariate Gaussian process regression model called *Multi-fidelity co-kriging.*

Finally, we perform several numerical tests which confirm the relevance of this new approach. We illustrate the suggested strategy on an academic example for the univariate case and with a real application on a tank under internal pressure for the multi-fidelity analysis.

From this work, two points can naturally be investigated. First, we could improve the uncertainty quantification for the meta-model. Indeed, in this paper, we do not take into account the uncertainty due to the estimation of the hyper-parameters of the covariance kernels. This can imply an underestimation of the predictive variance and thus it can be worth inferring from these parameters. The natural way to perform such analysis is to use a full-Bayesian approach. Second, the meta-model considered is built from a fixed experimental design set. Several methods exist to sequentially add new points on the design in order to perform optimization, to quantify a probability of failure or to improve the accuracy of the meta-model. However, no methods focus on the error reduction of the sensitivity index estimates. It would be of practical interest to develop sequential design strategies for a sensitivity analysis purpose.

9 Acknowledgments

References

[1] G.E.B. ARCHER, A SALTELLI, AND I.M. SOBOL, *Sensitivity measures, ANOVA-like techniques and the use of bootstrap*, Journal of Statistical Computation and Simulation, 58 (1997), pp. 99–120.

[2] G. CHASTAING, F. GAMBOA, AND C. PRIEUR, *Generalized Hoeffding-Sobol decomposition for dependent variables - Application to sensitivity analysis*, Electronic Journal of Statistics, 6 (2012), pp. 2420–2448.

[3] JP CHILÈS AND P DELFINER, *Geostatistics: modeling spatial uncertainty*, Wiley series in probability and statistics (Applied probability and statistics section), (1999).

[4] G. DAMBLIN, M. COUPLET, AND B. IOOSS, *Numerical studies of space filling designs: optimization algorithms and subprojection properties*, Journal of Simulation, submitted, (2013).

[5] JC FERREIRA AND VA MENEGATTO, *Eigenvalues of integral operators defined by smooth positive definite kernels*, Integral Equations and Operator Theory, 64 (2009), pp. 61–81.

[6] ROBERT B GRAMACY AND MATTHEW TADDY, *Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an r package for treed gaussian process models*, Journal of Statistical Software, 33 (2012), pp. 1–48.

[7] D Higdon, M Kennedy, J C Cavendish, J A Cafeo, and R D Ryne, *Combining field data and computer simulation for calibration and prediction*, SIAM Journal on Scientific Computing, 26 (2004), pp. 448–466.

[8] W. Hoeffding, *A class of statistics with asymptotically normal distribution*, The Annals of Mathematical Statistics, 19 (1948), pp. 293–325.

[9] Bertrand Iooss, François Van Dorpe, and Nicolas Devictor, *Response surfaces and sensitivity analyses for an environmental model of dose calculations*, Reliability Engineering & System Safety, 91 (2006), pp. 1241–1251.

[10] J. Jacques, C. Lavergne, and N. Devictor, *Sensitivity analysis in presence of model uncertainty and correlated inputs*, Reliability Engineering and System Safety, 91 (2006), pp. 1126–1134.

[11] Alexandre Janon, Thierry Klein, A. Lagnoux, M. Nodet, and Clementine Prieur, *Asymptotic normality and efficiency of two Sobol index estimators*, To appear in ESAIM Probability and Statistics, (2013).

[12] Alexandre Janon, Maëlle Nodet, Clémentine Prieur, et al., *Uncertainties assessment in global sensitivity indices estimation from metamodels*, To appear in International Journal for Uncertainty Quantification, (2013).

[13] Marc C. Kennedy and Anthony O'Hagan, *Predicting the output from a complex computer code when fast approximations are available*, Biometrika, 87 (2000), pp. 1–13.

[14] M C Kennedy and A O'Hagan, *Bayesian calibration of computer models*, Journal of the Royal Statistical Society, Series B, 63 (2001), pp. 425–464.

[15] Hermann König, *Eigenvalue distribution of compact operators*, Birkhäuser Basel, 1986.

[16] S Kucherenko, S Tarantola, and P Annoni, *Estimation of global sensitivity indices for models with dependent variables*, Computer Physics Communications, 183 (2012), pp. 937–946.

[17] Amandine Marrel, Bertrand Iooss, Beatrice Laurent, and Olivier Roustant, *Calculations of Sobol indices for the Gaussian process metamodel*, Reliability Engineering and System Safety, 94 (2009), pp. 742–751.

[18] Marzio Marseguerra, Riccardo Masini, Enrico Zio, and Giacomo Cojazzi, *Variance decomposition-based sensitivity analysis via neural networks*, Reliability Engineering & System Safety, 79 (2003), pp. 229–238.

[19] Jeremy E. Oakley and Anthony O'Hagan, *Probabilistic sensitivity analysis of complex models a Bayesian approach*, Journal of the Royal Statitistical Society series B, 66 (2004), pp. part 3, 751–769.

[20] Peter Z. G. Qian and C. F. Jeff Wu, *Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments*, Technometrics, 50 (2008), pp. 192–204.

[21] Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, 2006.

[22] C S Reese, A G Wilson, M Hamada, H F Martz, and K J Ryan, *Integrated analysis of computer and physical experiments*, Technometrics, 46 (2004), pp. 153–164.

[23] Andrea Saltelli, K. Chan, and E. M. Scott, *Sensitivity Analysis*, Wiley Series in Probability and Statistics, England, 2000.

[24] Thomas J. Santner, Brian J. Williams, and William I. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, 2003.

[25] IM Sobol, S Tarantola, D Gatelli, SS Kucherenko, and W Mauntz, *Estimating the approximation error when fixing unessential factors in global sensitivity analysis*, Reliability Engineering & System Safety, 92 (2007), pp. 957–960.

[26] I M Sobol, *Sensitivity estimates for non linear mathematical models*, Mathematical Modelling and Computational Experiments, 1 (1993), pp. 407–414.

[27] I. M. Sobol, *Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates*, Mathematics and Computers in Simulations, 55 (2001), pp. 271–280.

[28] Michael L. Stein, *Interpolation of Spatial Data*, Springer Series in Statistics, New York, 1999.

[29] Curtis B Storlie, Laura P Swiler, Jon C Helton, and Cedric J Sallaberry, *Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models*, Reliability Engineering & System Safety, 94 (2009), pp. 1735–1763.