ESAIM: Probability and Statistics Will be set by the publisher
URL: http://www.emath.fr/ps/

ASYMPTOTIC NORMALITY AND EFFICIENCY OF TWO SOBOL INDEX ESTIMATORS

Alexandre Janon[1], Thierry Klein[2], Agnès Lagnoux[2], Maëlle Nodet[1] and Clémentine Prieur[1]

Abstract. Many mathematical models involve input parameters, which are not precisely known. Global sensitivity analysis aims to identify the parameters whose uncertainty has the largest impact on the variability of a quantity of interest (output of the model). One of the statistical tools used to quantify the influence of each input variable on the output is the Sobol sensitivity index. We consider the statistical estimation of this index from a finite sample of model outputs: we present two estimators and state a central limit theorem for each. We show that one of these estimators has an optimal asymptotic variance. We also generalize our results to the case where the true output is not observable, and is replaced by a noisy version.

Résumé. De nombreux modèles mathématiques font intervenir plusieurs paramètres qui ne sont pas tous connus précisément. L'analyse de sensibilité globale se propose de sélectionner les paramètres d'entrée dont l'incertitude a le plus d'impact sur la variabilité d'une quantité d'intérêt (sortie du modèle). Un des outils statistiques pour quantifier l'influence de chacune des entrées sur la sortie est l'indice de sensibilité de Sobol. Nous considérons l'estimation statistique de cet indice à l'aide d'un nombre fini d'échantillons de sorties du modèle: nous présentons deux estimateurs de cet indice et énonçons un théorème central limite pour chacun d'eux. Nous démontrons que l'un de ces deux estimateurs est optimal en terme de variance asymptotique. Nous généralisons également nos résultats au cas où la vraie sortie du modèle n'est pas observée, mais où seule une version dégradée (bruitée) de la sortie est disponible.

1991 Mathematics Subject Classification. 62G05, 62G20.

The dates will be set by the publisher.

Introduction

Many mathematical models encountered in applied sciences involve a large number of poorly-known parameters as inputs. It is important for the practitioner to assess the impact of this uncertainty on the model output. An aspect of this assessment is sensitivity analysis, which aims to identify the most sensitive parameters, that is, parameters having the largest influence of the output. In global stochastic sensitivity analysis (see for example [22] and references therein) the input variables are assumed to be independent random variables. Their probability distributions account for the practitioner's belief about the input uncertainty. This turns the model output into a random variable, whose total variance can be split down into different partial variances (this is the so-called Hoeffding decomposition, see [32]). Each of these partial variances measures the uncertainty on the output induced by each input

Keywords and phrases: sensitivity analysis, Sobol indices, asymptotic efficiency, asymptotic normality, confidence intervals, metamodelling, surface response methodology

[1] Laboratoire Jean Kuntzmann, Université Joseph Fourier, INRIA/MOISE, 51 rue des Mathématiques, BP 53, 38041 Grenoble cedex 9, France

[2] Laboratoire de Statistique et Probabilités, Institut de Mathématiques Université Paul Sabatier (Toulouse 3) 31062 Toulouse Cedex 9, France

variable uncertainty. By considering the ratio of each partial variance to the total variance, we obtain a measure of importance for each input variable that is called the *Sobol index* or *sensitivity index* of the variable [27]; the most sensitive parameters can then be identified and ranked as the parameters with the largest Sobol indices.

Once the Sobol indices have been defined, the question of their effective computation or estimation remains open. In practice, one has to estimate (in a statistical sense) those indices using a finite sample (of size typically in the order of hundreds of thousands) of evaluations of model outputs [8]. Indeed, many Monte Carlo or quasi Monte Carlo approaches have been developed by the experimental sciences and engineering communities. This includes the FAST methods (see for example [4], [31] and references therein) and the Sobol pick-freeze (SPF) scheme (see [27, 28]). In SPF a Sobol index is viewed as the regression coefficient between the output of the model and its pick-freezed replication. This replication is obtained by holding the value of the variable of interest (frozen variable) and by sampling the other variables (picked variables). The sampled replications are then combined to produce an estimator of the Sobol index. In this paper we study very deeply this Monte Carlo method in the general framework where one or more variables can be frozen. This allows to define sensitivity indices with respect to a general random input living in a probability space (groups of variables, random vectors, random processes...). In this work, we study and compare two Sobol index estimators based on the SPF scheme; the first estimator, denoted by S_N^X, is well-known, the second, denoted by T_N^X has been introduced in [17]. For both estimators, we show convergence and give the rate of convergence; we also show that T_N^X is optimal (in terms of asymptotic variance) amongst regular estimators which are functions of the pick-freezed replications – this feature is called *asymptotic efficiency* and is a generalization of the notion of minimum variance unbiased estimator (see [32] chapters 8 and 25 or [10] for more details).

The SPF method requires many (typically, around one thousand times the number of input variables) evaluations of the model output. In many interesting cases, an evaluation of the model output is made by a complex computer code (for instance, a numerical partial differential equation solving algorithm) whose running time is not negligible (typically in the order of a second or a minute) for one single evaluation. When thousands of such evaluations have to be made, one generally replaces the original *exact* model by a faster-to-run *metamodel* (also known in the literature as *surrogate model* or *response surface* [1]) which is an approximation of the true model. Well-known metamodels include Kriging [24], polynomial chaos expansion [30] and reduced bases [12, 19], to name a few. When a metamodel is used, the estimated Sobol indices are tainted by a twofold error: *sampling error*, due to the replacement of the original, infinite population of all the possible inputs by a finite sample, and *metamodel error*, due to the replacement of the original model by an approximative metamodel.

The goal of this paper is to study the asymptotic behavior of these two errors on Sobol index estimation in the double limit where the sample size goes to infinity and the metamodel converges to the true model. Some work has been done on the non-asymptotic error quantification in Sobol index estimation in earlier papers [13, 16, 29] by means of confidence intervals which account for both sampling and metamodel errors. In this paper, we give necessary and sufficient conditions on the rate of convergence of the metamodel to the exact model for asymptotic normality of a natural Sobol index estimator to hold. The asymptotic normality allows us to produce asymptotic confidence intervals in order to assess the quality of our estimation. We also give sufficient conditions for a metamodel-based estimator to be asymptotically efficient. Asymptotic efficiency of an other Sobol index estimator has already been considered in [5]. In this work, the authors were interested in the asymptotic efficiency for local polynomial estimates of Sobol indices. Our approach proposes an estimator which has a simpler form, is less computationally intensive and is more precise in practice. Moreover, we derive results also in the case where the full model is replaced by a metamodel.

This paper is organized as follows: in the first section, we set up the notation, review the definition of Sobol indices and give two estimators of interest. In the second section, we prove asymptotic normality and asymptotic efficiency when the sample of outputs comes from the true model. These two properties are generalized in the third section where metamodel error is taken into account. The fourth section gives numerical illustrations on benchmark models and metamodels.

1. Definition and estimation of Sobol indices

1.1. Exact model

The output $Y \in \mathbb{R}$ is a function of independent random input variables $X \in \mathbb{R}^{p_1}$ and $Z \in \mathbb{R}^{p_2}$. In other words, Y and (X, Z) are linked by the relation

$$Y = f(X, Z) \tag{1}$$

where f is a deterministic function defined on $\mathcal{P} \subset \mathbb{R}^{p_1+p_2}$. We denote by $p = p_1 + p_2$ the total number of inputs of f.

In the paper X' will denote an independent copy of X. We also write $Y^X = f(X, Z')$.

We assume that Y is square integrable and non deterministic ($\mathrm{Var}(Y) \neq 0$). We are interested in the following Sobol index:

$$S^X = \frac{\mathrm{Var}\left(\mathbb{E}(Y|X)\right)}{\mathrm{Var}(Y)} \in [0, 1]. \tag{2}$$

This index quantifies the influence of the X input on the output Y: a value of S^X that is close to 1 indicates that X is highly influential on Y.

Remark 1.1. *All the results in this paper readily apply when X is multidimensional. In this case, S^X is usually called the* closed sensitivity index *of X (see [23]).*

Note that this separation between the input variables can be made without loss of generality, when one estimates Sobol indices independently. An ongoing work treats the case of joint Sobol index estimation.

1.2. Estimation of S^X

The next lemma shows how to express S^X using covariances. This will lead to a natural estimator which has already been considered in [9].

Lemma 1.2. *Assume that the random variables Y and Y^X are square integrable. Then*

$$\mathrm{Var}(\mathbb{E}(Y|X)) = \mathrm{Cov}(Y, Y^X).$$

In particular

$$S^X = \frac{\mathrm{Cov}\left(Y, Y^X\right)}{\mathrm{Var}(Y)}. \tag{3}$$

Remark 1.3. *Using a classical regression result, we see that*

$$S^X = \operatorname*{argmin}_{a \in \mathbb{R}} \left\{ \mathbb{E}\left((Y^X - \mathbb{E}(Y^X)) - a(Y - \mathbb{E}(Y)) \right)^2 \right\}. \tag{4}$$

A first estimator. In view of Lemma 1.2, we are now able to define a first natural estimator of S^X (all sums are taken for i from 1 to N):

$$S_N^X = \frac{\frac{1}{N}\sum Y_i Y_i^X - \left(\frac{1}{N}\sum Y_i\right)\left(\frac{1}{N}\sum Y_i^X\right)}{\frac{1}{N}\sum Y_i^2 - \left(\frac{1}{N}\sum Y_i\right)^2}, \tag{5}$$

where, for $i = 1, \ldots, N$:

$$Y_i = f(X_i, Z_i), \quad Y_i^X = f(X_i, Z_i'),$$

and $\{(X_i, Z_i)\}_{i=1,\ldots,N}$ and $\{(X_i, Z_i')\}_{i=1,\ldots,N}$ are two independent and identically distributed (i.i.d.) samples of the distribution of (X, Z), with $\{Z_i\}_i$ independent of $\{Z_i'\}_i$.

This estimator has been considered in [9], where it has been shown to be a practically efficient estimator.

A second estimator. We can take into account the observation of $\{Y_i^X\}_{1\le i\le N}$ to make an estimation of $\mathbb{E}(Y)$ and $\mathrm{Var}(Y)$ which is expected to perform better than any other based on $\{Y_i\}_{1\le i\le N}$ only. We propose the following estimator:

$$T_N^X = \frac{\frac{1}{N}\sum Y_i Y_i^X - \left(\frac{1}{N}\sum\left[\frac{Y_i+Y_i^X}{2}\right]\right)^2}{\frac{1}{N}\sum\left[\frac{Y_i^2+(Y_i^X)^2}{2}\right] - \left(\frac{1}{N}\sum\left[\frac{Y_i+Y_i^X}{2}\right]\right)^2}. \tag{6}$$

This estimator has been introduced in [17]. We will clarify what we mean when saying that T_N^X performs better than S_N^X in Proposition 2.3, Section 2.2 and Subsection 4.1.

Remark 1.4. *Note that the empirical variances in S_N^X and T_N^X can be rewritten as:*

$$S_N^X = \frac{\sum(Y_i-\overline{Y})(Y_i^X-\overline{Y^X})}{\sum(Y_i-\overline{Y})^2} \tag{7}$$

$$T_N^X = \frac{\sum(Y_i-\overline{Y_2})(Y_i^X-\overline{Y_2})}{\sum\left(\frac{Y_i+Y_i^X}{2}-\overline{Y_2}\right)^2} \tag{8}$$

where:

$$\overline{Y} = \frac{1}{N}\sum Y_i,\ \overline{Y^X} = \frac{1}{N}\sum Y_i^X,\ \overline{Y_2} = \frac{\overline{Y}+\overline{Y^X}}{2}.$$

The use of these formulae enables greater numerical stability (ie., less error due to round-offs). The Kahan compensated summation algorithm [14] may also be used on these sums. However, we will use definitions (5) and (6) for the mathematical analysis of S_N^X and T_N^X. This analysis is of course independent of the way the estimators are numerically computed in practice.

2. Asymptotic properties: exact model

2.1. Consistency and asymptotic normality

Throughout all the paper, we denote by $\mathcal{N}_k(\mu,\Sigma)$ the k-dimensional Gaussian distribution with mean μ and covariance matrix Σ, and, given any sequence of random variables $\{R_n\}_{n\in\mathbb{N}}$, we note

$$\overline{R}_N = \frac{1}{N}\sum_{n=1}^{N} R_n.$$

Proposition 2.1 (Consistency)**.** *We have:*

$$S_N^X \xrightarrow[N\to\infty]{a.s.} S^X \tag{9}$$

$$T_N^X \xrightarrow[N\to\infty]{a.s.} S^X. \tag{10}$$

Proof. The result is a straightforward application of the strong law of large numbers and that $\mathbb{E}(Y) = \mathbb{E}(Y^X)$ and $\mathrm{Var}(Y) = \mathrm{Var}(Y^X)$. □

Proposition 2.2 (Asymptotic normality)**.** *Assume that $\mathbb{E}(Y^4) < \infty$. Then*

$$\sqrt{N}\left(S_N^X - S^X\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1\left(0,\sigma_S^2\right) \tag{11}$$

and

$$\sqrt{N}\left(T_N^X - S^X\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1\left(0,\sigma_T^2\right) \tag{12}$$

where

$$\sigma_S^2 = \frac{\mathrm{Var}\left((Y-\mathbb{E}(Y))\left[(Y^X-\mathbb{E}(Y)) - S^X(Y-\mathbb{E}(Y))\right]\right)}{\left(\mathrm{Var}(Y)\right)^2},$$

$$\sigma_T^2 = \frac{\mathrm{Var}\left((Y-\mathbb{E}(Y))(Y^X-\mathbb{E}(Y)) - S^X/2\left((Y-\mathbb{E}(Y))^2 + (Y^X-\mathbb{E}(Y))^2\right)\right)}{(\mathrm{Var}(Y))^2}.$$

Proposition 2.3. *The asymptotic variance of T_N^X is always less than or equal to the asymptotic variance of S_N^X, with equality if and only if $S^X = 0$ or $S^X = 1$.*

To prove this Proposition, we need the following immediate Lemma:

Lemma 2.4. *Y and Y^X are exchangeable random variables, ie. $(Y, Y^X) \overset{\mathcal{L}}{=} (Y^X, Y)$.*

2.2. Asymptotic efficiency

In this section we study the asymptotic efficiency of S_N^X and T_N^X. This notion (see [32], Section 25 for its definition) extends the notion of Cramér-Rao bound to the semiparametric setting and enables to define a criteria of optimality for estimators, called asymptotic efficiency.

Let $\mathcal{P}$ be the set of all cumulative distribution functions (cdf) of exchangeable random vectors in $L^2(\mathbb{R}^2)$. It is clear that the cdf Q of a random vector of $L^2(\mathbb{R}^2)$ is in $\mathcal{P}$ if and only if Q is symmetric:

$$Q(a,b) = Q(b,a) \quad \forall (a,b) \in \mathbb{R}^2.$$

Let P be the cdf of (Y, Y^X). We have $P \in \mathcal{P}$ thanks to Lemma 2.4.

Proposition 2.5 (Asymptotic efficiency)**.** *$\{T_N^X\}_N$ is asymptotically efficient for estimating S^X for $P \in \mathcal{P}$.*

We will use the following Lemma, which is also of interest in its own right:

Lemma 2.6 (Asymptotic efficiency in $\mathcal{P}$)**.**

(1) *Let $\Phi_1 : \mathbb{R} \to \mathbb{R}$ be a function in $L^2(P)$. The sequence of estimators $\left\{\Phi_N^1\right\}_N$ given by:*

$$\Phi_N^1 = \frac{1}{N} \sum \frac{\Phi_1(Y_i) + \Phi_1(Y_i^X)}{2}$$

is asymptotically efficient for estimating $\mathbb{E}(\Phi_1(Y))$ for $P \in \mathcal{P}$.

(2) *Let $\Phi_2 : \mathbb{R}^2 \to \mathbb{R}$ be a symmetric function in $L^2(P)$. The sequence $\left\{\Phi_N^2\right\}_N$ given by:*

$$\Phi_N^2 = \frac{1}{N} \sum \Phi_2\left(Y_i, Y_i^X\right)$$

is asymptotically efficient for estimating $\mathbb{E}(\Phi_2(Y, Y^X))$ for $P \in \mathcal{P}$.

3. Asymptotic properties: metamodel

3.1. Metamodel-based estimation

As said in the introduction, we often are in a situation where the exact output f is too costly to be evaluated numerically (thus, Y and Y^X are not observable variables in our estimation problem) and has to be replaced by a metamodel $\widetilde{f}$, which is a faster to evaluate approximation of f. We view this approximation as a perturbation of the exact model by some function δ:

$$\widetilde{Y} = \widetilde{f}(X,Z) = f(X,Z) + \delta,$$

where the perturbation $\delta = \delta(X, Z, \xi)$ is also a function of a random variable ξ independent from X and Z.

We also define, as before

$$\widetilde{Y}^X = \widetilde{f}(X, Z').$$

Assuming again that $\widetilde{Y}$ is non deterministic and in L^2, we can consider the following Sobol index, with respect to the metamodel:

$$\widetilde{S}^X = \frac{\mathrm{Var}(\mathbb{E}(\widetilde{Y}|X))}{\mathrm{Var}(\widetilde{Y})} \tag{13}$$

and its estimators:

$$\widetilde{S}_N^X = \frac{\frac{1}{N}\sum \widetilde{Y}_i \widetilde{Y}_i^X - \left(\frac{1}{N}\sum \widetilde{Y}_i\right)\left(\frac{1}{N}\sum \widetilde{Y}_i^X\right)}{\frac{1}{N}\sum \widetilde{Y}_i^2 - \left(\frac{1}{N}\sum \widetilde{Y}_i\right)^2} \tag{14}$$

$$\widetilde{T}_N^X = \frac{\frac{1}{N}\sum \widetilde{Y}_i \widetilde{Y}_i^X - \left(\frac{1}{N}\sum \left[\frac{\widetilde{Y}_i + \widetilde{Y}_i^X}{2}\right]\right)^2}{\frac{1}{N}\sum \left[\frac{\widetilde{Y}_i^2 + (\widetilde{Y}_i^X)^2}{2}\right] - \left(\frac{1}{N}\sum \left[\frac{\widetilde{Y}_i + \widetilde{Y}_i^X}{2}\right]\right)^2}. \tag{15}$$

The goal of this section is to give sufficient conditions on the perturbation δ for $\widetilde{S}_N^X$ and $\widetilde{T}_N^X$ to satisfy asymptotic normality (Subsection 3.2), and $\widetilde{T}_N^X$ to be asymptotically efficient (Subsection 3.3), with respect to the Sobol index of the *true* model S^X.

3.2. Consistency and asymptotic normality

In the first Subsection (3.2.1) we suppose that the error term δ does not depend on N. In this case, if the Sobol index of the exact model is different from the Sobol index of the metamodel, then neither consistency nor asymptotic normality are possible. In the second subsection (3.2.2), we let δ depend on N and we give conditions for consistency and asymptotic normality to hold.

3.2.1. *First case : δ does not depend on N*

Remark 3.1. *If $\widetilde{S}^X - S^X \neq 0$ then neither $\widetilde{S}_N^X$ nor $\widetilde{T}_N^X$ are consistent for estimating S^X. Indeed, we have*

$$\widetilde{S}_N^X - S^X = (\widetilde{S}_N^X - \widetilde{S}^X) + (\widetilde{S}^X - S^X).$$

The first term converges to 0 almost surely by Proposition 2.1 applied to $\widetilde{S}_N^X$. However, the second is nonzero by assumption. The same holds for $\widetilde{T}_N^X$.

This remark shows that a naive consideration of the metamodel error (ie., with fixed metamodel) is not satisfactory for an asymptotic justification of the use of a metamodel. More specifically, it is impossible to have asymptotic normality for $\widetilde{S}_N^X$ and $\widetilde{T}_N^X$ in any nontrivial case if δ does not vanish (in some sense) asymptotically. This justifies the consideration of cases where δ depends on N, and this is the object of the next subsection.

3.2.2. *Second case :* Var δ_N *converges to 0 as $N \to \infty$*

We now assume that the perturbation δ is a function of the sample size N. This entails that $\widetilde{f}$, as well as $\widetilde{Y}$, $\widetilde{Y}^X$ and $\widetilde{S}^X$ depend on N. We emphasize this dependence by using the notations δ_N, $\widetilde{f}_N$, $\widetilde{Y}_N$, $\widetilde{Y}_N^X$. We keep, however, using the notations $\widetilde{S}_N^X$ and $\widetilde{T}_N^X$ for the estimators of $\widetilde{S}^X$ defined at (14) and (15).

Assumption. We suppose that $\widetilde{f}_N - f = \delta_N \underset{N\to+\infty}{\xrightarrow{L^2}} c$ for some constant c.

Proposition 3.2. *We have $\widetilde{S}^X \underset{N\to+\infty}{\longrightarrow} S^X$.*

Proposition 3.3. *Assume there exist $s > 0$ and $C > 0$ such that*

$$\forall N, \;\; \mathbb{E}\left(\left|\widetilde{Y}_N\right|^{4+s}\right) < C. \tag{16}$$

Then

$$\sqrt{N}\left(\widetilde{S}_N^X - \widetilde{S}^X\right) \underset{N\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}_1\left(0, \sigma_S^2\right) \tag{17}$$

$$\sqrt{N}\left(\widetilde{T}_N^X - \widetilde{S}^X\right) \underset{N\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}_1\left(0, \sigma_T^2\right) \tag{18}$$

where σ_S^2 and σ_T^2 are the asymptotic variances of S_N^X and T_N^X given, respectively, in (11) and (12).

We are actually interested in the asymptotic distribution of $\sqrt{N}\left(\widetilde{S}_N^X - S^X\right)$. In the remaining of the subsection, we will show that this convergence depends on the rate of convergence to 0 of $\mathrm{Var}(\delta_N)$.

Theorem 3.4. *Let:*

$$C_{\delta,N} = 2\mathrm{Var}(Y)^{1/2}\left[\mathrm{Corr}(Y,\delta_N^X) - \mathrm{Corr}(Y,Y^X)\mathrm{Corr}(Y,\delta_N)\right] + \mathrm{Var}(\delta_N)^{1/2}\left[\mathrm{Corr}(\delta_N,\delta_N^X) - \mathrm{Corr}(Y,Y^X)\right],$$

for $\delta_N^X = \delta_N(X,Z')$, and, given any L^2 random variables A and B of nonzero variance:

$$\mathrm{Corr}(A,B) = \frac{\mathrm{Cov}(A,B)}{\sqrt{\mathrm{Var}A\,\mathrm{Var}B}}.$$

Assume that $C_{\delta,N}$ does not converge to 0.

(1) *If $\mathrm{Var}(\delta_N) = o\left(\frac{1}{N}\right)$, then asymptotic normalities of $\widetilde{S}_N^X$ and $\widetilde{T}_N^X$ for S^X hold, i.e.*

$$\sqrt{N}(\widetilde{S}_N^X - S^X) \underset{N\to+\infty}{\longrightarrow} \mathcal{N}(0,\sigma_S^2) \tag{19}$$

and:

$$\sqrt{N}(\widetilde{T}_N^X - S^X) \underset{N\to+\infty}{\longrightarrow} \mathcal{N}(0,\sigma_T^2). \tag{20}$$

(2) *If $N\mathrm{Var}(\delta_N) \to \infty$, then (19) and (20).*

(3) *If $C_{\delta,N}$ converges to a nonzero constant C and $\gamma \in \mathbb{R}$ so that $\mathrm{Var}(\delta_N) = \frac{\gamma}{CN} + o\left(\frac{1}{N}\right)$, then:*

$$\sqrt{N}(\widetilde{S}_N^X - S^X) \underset{N\to+\infty}{\longrightarrow} \mathcal{N}(\gamma,\sigma_S^2),$$

and:

$$\sqrt{N}(\widetilde{T}_N^X - S^X) \underset{N\to+\infty}{\longrightarrow} \mathcal{N}(\gamma,\sigma_T^2).$$

Remark 3.5. *Obviously, if $C_{\delta,N}$ converges to 0, then asymptotic normalities of $\widetilde{S}_N^X$ and $\widetilde{T}_N^X$ hold under weaker assumptions on $\mathrm{Var}(\delta_N)$.*

3.3. **Asymptotic efficiency**

Proposition 3.6 (Asymptotic efficiency for the metamodel)**.** *Assume*

(1) $\exists s>0, C>0$ *s.t.* $\forall N$, $\mathbb{E}\left(\left|Y\right|^{4+s}\right) < C$ *and* $\mathbb{E}\left(\left|\widetilde{Y}\right|^{4+s}\right) < C$;
(2) $N\mathrm{Var}(\delta_N) \to 0$;
(3) $\sqrt{N}\mathbb{E}(\delta_N) \to 0$.

Then $\left\{\widetilde{T}_N^X\right\}$ is asymptotically efficient for estimating S^X.

Remark 3.7. *By Minkowski inequality, the first hypothesis implies $\mathbb{E}(\delta_N^{4+s}) < 2C^{\frac{1}{4+s}}$ and the asymptotic normality by Lemma 3.3 and Theorem 3.4.*

4. Numerical illustrations

In this section, we illustrate the asymptotic results of Sections 2.1 and 3.2 when the exact model is the Ishigami function [11]:

$$f(X_1,X_2,X_3) = \sin X_1 + 7\sin^2 X_2 + 0.1X_3^4 \sin X_1 \tag{21}$$

for $(X_j)_{j=1,2,3}$ are i.i.d. uniform random variables in $[-\pi;\pi]$. In this case, all the integrability conditions are satisfied (we even have $Y \in L^\infty$).

The Sobol index of f with respect to input variable X_1 is S^X defined in (2) for $X = X_1$ and $Z = (X_2,X_3)$; we denote it by S^1. Similarly, S^2 (resp. S^3) is S^X obtained taking $X = X_2$ and $Z = (X_1,X_3)$ (resp. $X = X_3$ and $Z = (X_1,X_2)$).

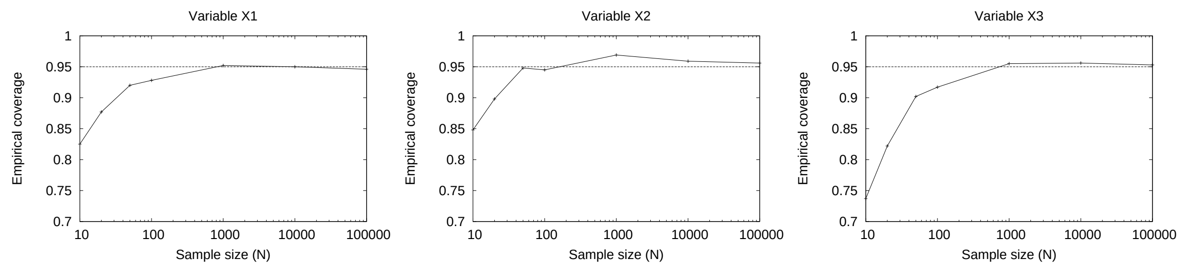

Figure 1. Empirical coverages of asymptotic confidence intervals for S^1 (left), S^2 (center) and S^3 (right), as a function of the sample size. The S_N estimator is used.

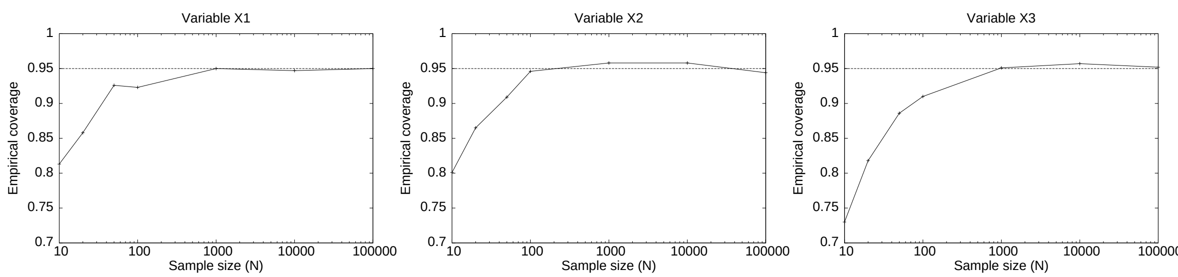

Figure 2. Empirical coverages of asymptotic confidence intervals for S^1 (left), S^2 (center) and S^3 (right), as a function of the sample size (for the exact model). The T_N estimator is used.

Exact values of these indices are analytically known:

$$S^1 = 0.3139, \quad S^2 = 0.4424, \quad S^3 = 0.$$

For a sample size N, a risk level $\alpha \in]0;1[$ and for each input variable, a confidence interval for S^X (S^X being one of S^1, S^2 or S^3) of confidence level $1-\alpha$ can be estimated – using evaluations of the true model f – by approximating the distribution of S_N^X (or T_N^X) by its Gaussian distribution given in Proposition 11, using empirical estimators of the asymptotic variances stated in this Proposition.

In the case where only a perturbated model (metamodel) $\widetilde{f}_N = f + \delta_N$ is available, a confidence interval can still be estimated by using the $\widetilde{S}_N^X$ (or $\widetilde{T}_N^X$) estimator.

Thanks to Proposition 3.3, the level of the resulting confidence interval should be close to $1-\alpha$ for sufficiently large values of N if (and only if) $\mathrm{Var}\delta_N$ decreases sufficiently quickly with N.

The levels of the obtained confidence interval can be estimated by computing a large number R of confidence interval replicates, and by considering the empirical coverage, that is, the proportion of intervals containing the true index value; it is well known that this empirical coverage strongly converges to the level of the interval as R goes to infinity.

In the next subsections, we present the estimations of the levels of the confidence interval for the Ishigami model (21) using the true model (Subsection 4.1), and, with various synthetic model perturbations (Subsections 4.2 and 4.3), as well as RKHS (Kriging) metamodels (Subsection 4.4) and nonparametric regression metamodels (Subsection 4.5). We begin by comparing S_N and T_N on the exact model (Subsection 4.1), then we illustrate the generalization to the metamodel case on the widespread estimator S_N; the condition to ensure asymptotic normality in the metamodel is the same for S_N and T_N. All simulations have been made with $R = 1000$ and $\alpha = 0.05$.

4.1. **Exact model**

Figure 1 shows the empirical coverage of the asymptotic confidence interval built using the S_N^X estimator, plotted as a function of the sample size N. The theoretical level 0.95 is represented with a dotted line. Figure 2 does the same using the T_N^X estimator.

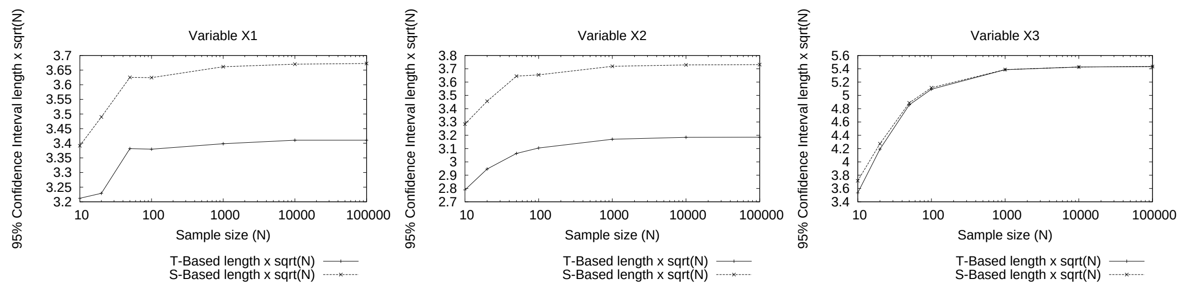

FIGURE 3. Lengths (rescaled by $\sqrt{N}$) of the estimated 95% confidence intervals for S^1 (left), S^2 (center) and S^3 (right), as functions of the sample size (for the exact model). In solid line: length of the interval built from T_N estimator; in dotted line: length of the interval built from S_N estimator.

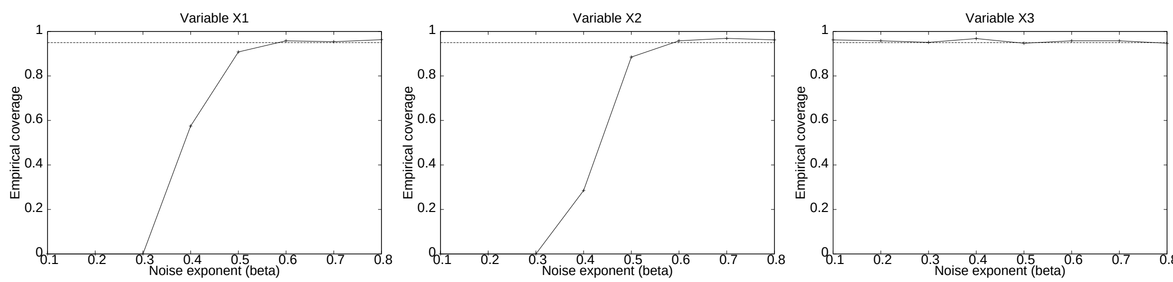

FIGURE 4. Empirical coverages of the asymptotic confidence intervals for S^1, S^2 and S^3, as a function of β (for the Gaussian-perturbated model).

We see that the coverages get closer to the target level 0.95 as N increases, thereby assessing the reliability of the asymptotic confidence interval.

Figure 3 compares the efficiency of S_N^X and T_N^X by plotting the confidence interval lengths for the two estimators, as functions of the sample size. As the lengths for both estimators are $O(1/\sqrt{N})$, we plot the lengths multiplied by $\sqrt{N}$. We see that T_N^X always produce smaller confidence intervals, except for X_3 where the lengths are sensibly the same; this conclusion fully agrees with Proposition 2.3.

4.2. **Gaussian-perturbated model**

We consider a perturbation $\widetilde{f}_N$ of the original output f:

$$\widetilde{f}_N = f + \frac{5\xi}{N^{\beta/2}}$$

where $\beta > 0$ and ξ is a standard Gaussian.

The perturbation $\delta_N = 5\frac{\xi}{N^{\beta/2}}$ leads to $\mathrm{Var}\delta_N \propto N^{-\beta}$. Since:

$$C_\delta = O\left(\mathrm{Var}(\delta_N)^{1/2}\right) = O\left(N^{-\beta/2}\right),$$

the proof of Theorem 3.4 shows that $\widetilde{S}_N$ is asymptotically normal for S if $\beta > 1/2$. For indices relative to X_1 and X_2, this sufficient condition is also necessary, as C_δ is actually equivalent to $N^{-\beta/2}$. For X_3, we have $C_\delta = 0$ so that $\widetilde{S}_N$ is asymptotically normal for S for any positive β.

This is illustrated for $N = 50000$ in Figure 4. We see that the empirical coverages of the confidence interval for S^1 and S^2 jump to 0.95 near $\beta = 1/2$, while, for S^3, this coverage is always close to 0.95.

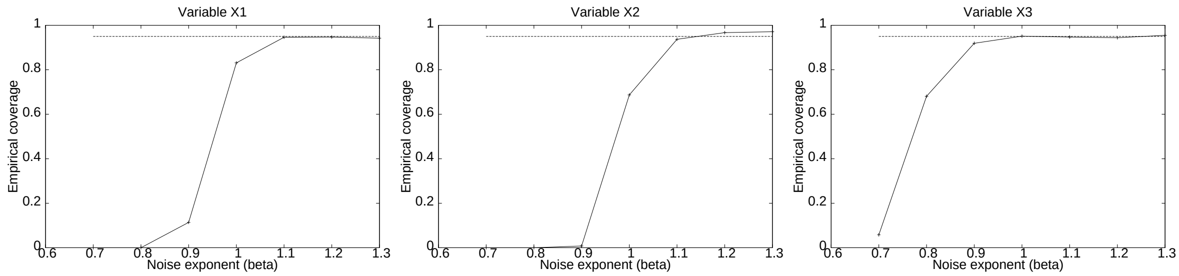

FIGURE 5. Empirical coverages of the asymptotic confidence intervals for S^1, S^2 and S^3, as a function of β (for the Weibull-perturbated model).

4.3. **Weibull-perturbated model**

We now take a different perturbation of the output:

$$\widetilde{f}_N = f + \frac{5WX_3^2}{N^{\beta/2}}$$

where W is Weibull-distributed with scale parameter $\lambda = 1$ and shape parameter $k = 1/2$. Here, the perturbation depends on the inputs and, as for every input variable, $C_{\delta,N}$ does not converge to zero, Theorem 3.4 states in particular that $\widetilde{S}_N$ is asymptotically normal for S for $\beta > 1$. Again, this property is suggested for $N = 50000$ by the plot in Figure 5.

4.4. **RKHS metamodel**

In this part, we discuss the use of a reproducing kernel Hilbert space (RKHS) interpolator [24, 25, 26] as metamodel $\widetilde{f}$. Such metamodels (also known as Kriging, or Gaussian process metamodels) are widely used when performing sensitivity analysis of time-expensive computer codes [16]. Note that, according to [3], analytical formulae are in some cases (e.g., uniform or gaussian distributions for the inputs) available for Sobol indices computation, avoiding the necessity to use a Monte-Carlo scheme. In this paper, we chose to perform Monte-Carlo estimation on an RKHS metamodel so as to illustrate our theoretical results. Moreover, the Monte-Carlo approach is more flexible and can be applied for complex inputs' distributions. The interpolator depends on a learning sample $\{(d_1, f(d_1)), \ldots, (d_n, f(d_n))\}$, where the design points $\mathcal{D} = \{d_i\}_{i=1,\ldots,n} \subset \mathcal{P}$ are generally chosen according to a space-filling design, for instance the so-called maximin LHS (latin hypercube sampling) designs. Increasing the learning sample size n will increase the necessary number of evaluations of the true model f (each evaluation being potentially very computationally demanding) to build the learning sample, but will also enhance the quality of the interpolation (i.e. reduce metamodel error).

The error analysis of the RKHS method [15, 25] shows that there exist positive constants $\mathcal{C}$ and $\mathcal{K}$, depending on f, so that:

$$\forall u \in \mathcal{P}, \quad \left| f(u) - \widetilde{f}(u) \right| \leq \mathcal{C} e^{-\mathcal{K}/h_{\mathcal{D},\mathcal{P}}}$$

where:

$$h_{\mathcal{D},\mathcal{P}} = \sup_{u \in \mathcal{P}} \min_{d \in \mathcal{D}} \|d - u\|$$

for a given norm $\|\cdot\|$ on $\mathcal{P}$.

The quantity $h_{\mathcal{D},\mathcal{P}}$ can be linked to the number of points $n^*(\epsilon)$ in an optimal covering of $\mathcal{D}$:

$$n^*(\epsilon) = \min\{p \in \mathbb{N}^* \mid \exists (d_1, \ldots, d_p) \in \mathcal{P} \text{ s.t. } \forall u \in \mathcal{P}, \exists i \in \{1, \ldots, p\} \text{ satisfying } \|u - d_i\| \leq \epsilon\}.$$

In other words, $n^*(\epsilon)$, known as the *covering number of* $\mathcal{P}$, is the smallest size of a design $\mathcal{D}$ satisfying $h_{\mathcal{D},\mathcal{P}} \leq \epsilon$.

It is known that, when $\mathcal{P}$ is a compact subset of $\mathbb{R}^p$ (in our context, $p = p_1 + p_2$ is the number of input parameters), there exist constants A and B so that:

$$A\epsilon^{-p} \leq n^*(\epsilon) \leq B\epsilon^{-p}.$$

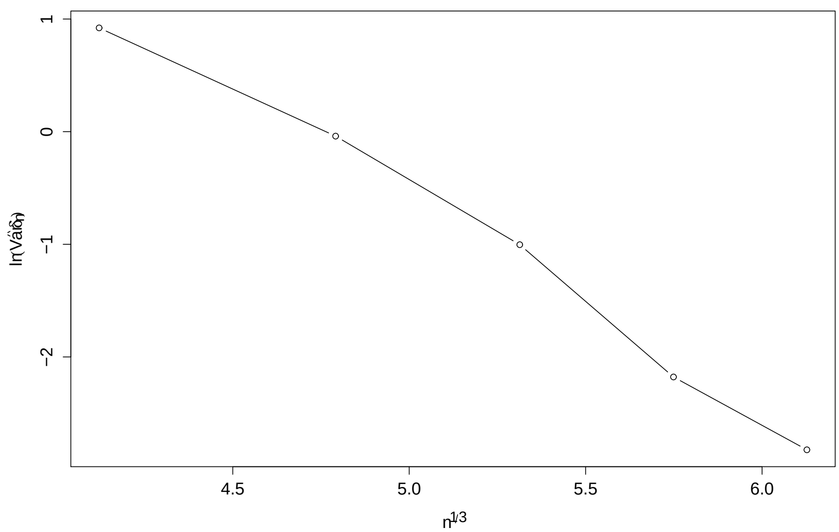

FIGURE 6. Estimation of the Kriging metamodel error variance (log. scale) as function of the learning sample size n.

Hence, assuming that an optimal design of size n is chosen, we have, for a constant B':

$$h_{\mathcal{D},\mathcal{P}} \leq B' n^{-1/p}$$

and we have the following pointwise metamodel error bound, for constants C and K':

$$\forall u \in \mathcal{P}, \quad \left| f(u) - \widetilde{f}(u) \right| \leq \mathcal{C} e^{-\mathcal{K}' n^{1/p}}$$

which obviously leads to an integrated error bound on the variance of the metamodel error:

$$\mathrm{Var}\delta \leq C e^{-k n^{1/p}}$$

for suitable constants C and k.

Numerical illustration

We illustrate the properties of the RKHS-based sensitivity analysis using the Ishigami function (21) as true model, maximin LHSes for design points selection. RKHS interpolation also depends on the choice of a kernel, which we choose Gaussian all the way through. All simulations have been made with the `R` software [20], together with the `lhs` package [2] for design sampling and the `mlegp` package [6] for Kriging.

Figure 6, which shows an estimation (based on a sample of 1000 metamodel errors) of the (logarithm of) variance of metamodel error, plotted against the cubed root of the learning sample size $n^{1/3}$. Using an exponential regression, we find that:

$$\mathrm{Var}(\delta) \approx \widehat{C} e^{-\widehat{k} n^{1/3}} \tag{22}$$

where:

$$\widehat{k} = 1.91$$

Now, if we let the learning sample size n depend on the Monte-Carlo sample size N by the relation:

$$n = (a \ln N)^3$$

a	N	n	Coverage for S^1	Cov. for S^2	Cov. for S^3
.4	3000	33	0.1	0	0.7
.4	4000	37	0.08	0	0.78
.4	6000	43	0.26	0.3	0.88
.4	10000	51	0.28	0.18	0.78
.4	20000	77	0.28	0.1	0.59
.6	3000	111	0.79	0.37	0.9
.6	4000	124	0.8	0.7	0.94
.6	10000	169	0.92	0.82	0.94
.6	20000	210	0.93	0.85	0.95
.7	3000	177	0.93	0.88	0.93
.7	4000	196	0.9	0.91	0.94
.7	6000	226	0.94	0.93	0.97
.8	4000	293	0.95	0.95	0.95

Table 1. Estimation of the asymptotic coverages for the RKHS Ishigami metamodel. Empirical coverages are obtained using 100 confidence interval replicates. Theoretical coverage is 0.95.

for $a > 0$, Theorem 3.4 suggests that the metamodel-based estimators of the sensitivity indices are asymptotically normal if and only if $N^{-a\widehat{k}+1} \to 0$ when $N \to +\infty$, that is $a > \frac{1}{\widehat{k}}$, or

$$a > 0.52, \tag{23}$$

according to our numerical value for $\widehat{k}$.

Even if it has not been rigorously proved that this condition is necessary and sufficient (due to the estimation of k and the fact that (22) provably holds, possibly with different constants, as an upper bound), one should observe in practice that the behavior of the empirical confidence intervals for large values of N changes as this critical value of a is crossed. Table 1 below shows the results obtained for different subcritical and supercritical values of a (i.e., (23) does not hold, or hold, respectively), and provides a clear illustration of this fact.

4.5. **Nonparametric regression**

In this section, we consider the case where the true model f is not directly observable, but is only available through a finite set of *noisy* realisations of:

$$f_{\text{noisy}}(D_i) = f(D_i) + \epsilon_i, \quad i = 1, \dots, n$$

where $\mathcal{D} = (D_i = (X_i, Z_i))_{i=1,\dots,n}$ are independent copies of (X, Z), and $\{\epsilon_i\}_{i=1,\dots,n}$ are independent, identically distributed centered random variables.

As discussed in Section 3.2.1, one should expect that the Sobol index estimator computed on f_{noisy} are not asymptotically normal for the estimation of the Sobol indices of f (as $\operatorname{Var}(\epsilon_i)$ is fixed). This motivates the use of a smoothed estimate of f, which we will take as our perturbated model $\widetilde{f} = \widetilde{f}_{\mathcal{D}}$. We consider the Nadaraya-Watson estimator:

$$\widetilde{f}_{\mathcal{D}}(u) = \begin{cases} \dfrac{\sum_{i=1}^n K_h(u - D_i) f_{\text{noisy}}(D_i)}{\sum_{i=1}^n K_h(u - D_i)} & \text{if } \displaystyle\sum_{i=1}^n K_h(u - D_i) \neq 0 \\ 0 \text{ else.} \end{cases}$$

where K_h is a smoothing kernel of window $h \in \mathbb{R}^p$; for instance K_h is a Gaussian kernel:

$$K_h(v) = \exp\left(- \sum_{i=1}^p \frac{\|v_i\|^2}{h_i^2} \right) \tag{24}$$

where the norm $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^p$.

It is known that, under regularity conditions on f, and a n-dependent appropriate choice of h, the mean integrated square error (MISE) of $\widetilde{f}$ satisfies:

$$\int \mathbb{E}_{\mathcal{D}} \left(\left(f(u) - \widetilde{f}_{\mathcal{D}}(u) \right)^2 \right) \mathrm{d}u \leq C' n^{-\gamma}, \tag{25}$$

for a positive constant C' and a positive γ (which depends only on the dimension p and the regularity of f), and where $\mathbb{E}_{\mathcal{D}}$ denotes expectation with respect to the random "design" $\mathcal{D}$.

Now, by Fubini-Tonelli's theorem, we have:

$$\int \mathbb{E}_{\mathcal{D}} \left(\left(f(u) - \widetilde{f}_{\mathcal{D}}(u) \right)^2 \right) \mathrm{d}u = \mathbb{E}_{\mathcal{D}} \left(\int \left(f(u) - \widetilde{f}_{\mathcal{D}}(u) \right)^2 \mathrm{d}u \right). \tag{26}$$

By using (26), (25) and applying Markov's inequality to the positive random variable $\int \left(f(u) - \widetilde{f}_{\mathcal{D}}(u) \right)^2 \mathrm{d}u$, we have that, for any $\epsilon > 0$,

$$\mathbb{P} \left(\left\{ \mathcal{D} \ / \ \int \left(f(u) - \widetilde{f}_{\mathcal{D}}(u) \right)^2 \mathrm{d}u \leq \frac{C'}{\epsilon} n^{-\gamma} \right\} \right) \geq 1 - \epsilon.$$

Hence, for a fixed risk $\epsilon > 0$, there exist $C > 0$ and $\gamma > 0$ so that:

$$\int \left(\widetilde{f}_{\mathcal{D}}(u) - f(u) \right)^2 \mathrm{d}u \leq C n^{-\gamma} \tag{27}$$

holds with probability greater than $1 - \epsilon$ (with respect to the choice of $\mathcal{D}$).

We recall that the quantity we have to consider in order to study asymptotic normality of Sobol index estimator on the metamodel is:

$$\mathrm{Var}(\delta) = \int \left(f(u) - \widetilde{f}_{\mathcal{D}}(u) \right)^2 \mathrm{d}u - \left(\int \left(f(u) - \widetilde{f}_{\mathcal{D}}(u) \right) \mathrm{d}u \right)^2$$

and that, obviously,

$$\mathrm{Var}(\delta) \leq \int \left(f(u) - \widetilde{f}_{\mathcal{D}}(u) \right)^2 \mathrm{d}u.$$

This gives, by making use of (27):

$$\mathrm{Var}\,(\delta) \leq C n^{-\gamma} \tag{28}$$

with probability greater than $1 - \epsilon$.

In most cases of application, the design $\mathcal{D}$ is fixed. In view of (28), it is reasonable to suppose that there exist $C > 0$ and $\beta > 0$ so that:

$$\mathrm{Var}\,(\delta) \leq C n^{-\beta}$$

and we make n depend on N by the following relation:

$$n = N^a,$$

for $a > 0$. By Theorem 3.4, the estimator sequence $\{\widetilde{S}_N\}$ is asymptotically normal provided that $N\mathrm{Var}(\delta_N) \to 0$, that is: $a > \frac{1}{\beta}$.

Numerical illustration

We now illustrate this property using the Ishigami function (21) as true model, and a Gaussian white noise ϵ_i of standard deviation 0.3 (yielding to a signal-to-noise ratio of 90%).

The nonparametric regressions are carried using a Gaussian kernel (24), the R package `np` [7], together with the extrapolation method of [21] for window selection and the `FIGtree` [18] C++ library for efficient Nadaraya-Watson evaluation based on fast gaussian transform.

Figure 7, which shows an estimation (based on a test sample of size 3000) of $\mathrm{Var}(\delta)$ in function of n, and a power regression shows that:

$$\mathrm{Var}\,(\delta) \approx C n^{-\widehat{\beta}}$$

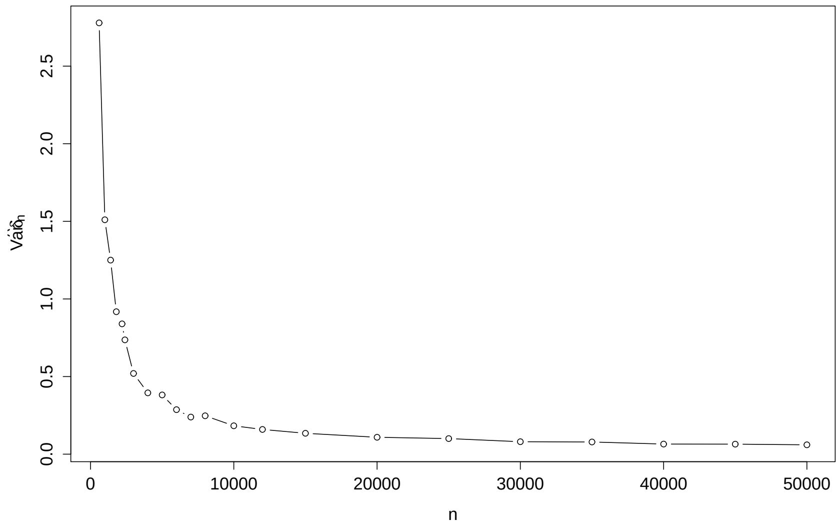

FIGURE 7. Estimation of the nonparametric regression error variance (log. scale) as function of the learning sample size n (Subsection 4.5).

a	N	n	Coverage for S^1	Cov. for S^2	Cov. for S^3
0.8	1000	252	0.25	0.01	0.94
0.8	2000	438	0.05	0.02	0.86
1.1	1000	1996	0.95	0.97	0.96
1.1	2000	4277	0.95	0.93	0.96
1.2	1000	3982	0.93	0.95	0.96
1.2	2000	9147	0.96	0.97	0.95
1.3	1000	7944	0.95	0.99	0.94
1.3	2000	19559	0.95	0.95	0.96

TABLE 2. Estimation of the asymptotic coverages for the Ishigami nonparametric regression. Empirical coverages are obtained using 100 confidence interval replicates. Theoretical coverage is 0.95.

with $\widehat{\beta} = 0.86$. This gives an estimate of 1.16 as the critical a for asymptotic normality.

As in the RKHS case, we performed estimations of the coverages of the asymptotic confidence interval for several values of a and N; the results are gathered in Table 2. We see that, first, the condition $a > 1.16$ implies correct coverages, and, second, the condition also seems to be near-necessary to have asymptotic normality. We also remark that, for the asymptotic normality to hold, the necessary number of noisy model evaluations is asymptotically comparable to the Monte-Carlo sample size (while, in the RKHS case, the necessary number of true model evaluations was asymptotically negligible with respect to the Monte-Carlo sample size): this shows that the nonparametric regression is suitable in the case of noisy but abundant model evaluations, while RKHS interpolation is clearly preferable when the true model output is costly to evaluate (i.e. few model outputs are available).

5. Appendix: Proofs

Proof of Lemma 1.2. On one hand, since $Y \overset{\mathcal{L}}{=} Y^X$ (that is, Y and Y^X have the same distribution), we have

$$\mathrm{Cov}(Y, Y^X) = \mathbb{E}(YY^X) - \mathbb{E}(Y)\mathbb{E}(Y^X) = \mathbb{E}(YY^X) - \mathbb{E}(Y)^2.$$

On the other hand, Y and Y^X are independent conditionally on X, so that

$$\mathbb{E}(YY^X) = \mathbb{E}(\mathbb{E}(YY^X|X)) = \mathbb{E}(\mathbb{E}(Y|X)\mathbb{E}(Y^X|X)) = \mathbb{E}(\mathbb{E}(Y|X)^2).$$

$\square$

Proposition 2.2: Proof of (11). We begin by noticing that S_N^X is invariant by any centering (translation) of the Y_i and Y_i^X. To simplify the next calculations, we suppose that they have been recentred by $-\mathbb{E}(Y)$. By setting:

$$U_i = \left((Y_i - \mathbb{E}(Y))(Y_i^X - \mathbb{E}(Y)), \quad Y_i - \mathbb{E}(Y), \quad Y_i^X - \mathbb{E}(Y), \quad (Y_i - \mathbb{E}(Y))^2\right)^T, \tag{29}$$

this implies that:

$$S_N^X = \Psi_S(\overline{U}_N)$$

with:

$$\Psi_S(x,y,z,t) = \frac{x - yz}{t - y^2}$$

The central limit theorem gives that:

$$\sqrt{N}\left(\overline{U}_N - \mu\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_4\left(0,\Gamma\right)$$

where Γ is the covariance matrix of U_1 and:

$$\mu = \begin{pmatrix} \mathrm{Cov}(Y, Y^X) \\ 0 \\ 0 \\ \mathrm{Var}(Y) \end{pmatrix}.$$

The so-called Delta method [32] (Theorem 3.1) gives:

$$\sqrt{N}\left(S_N^X - S^X\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0, g^T\Gamma g)$$

where:

$$g = \nabla\Psi_S(\mu).$$

Note that since by assumption $\mathrm{Var}(Y) \neq 0$, Ψ_S is differentiable at μ and we will see that $g^T\Gamma g \neq 0$, so that the application of the Delta method is justified. By differentiation, we get that, for any x, y, z, t so that $t \neq y^2$:

$$\nabla\Psi_S(x,y,z,t) = \left(\frac{1}{t-y^2}, \quad \frac{-z(t-y^2) + (x-yz)\cdot 2y}{(t-y^2)^2}, \quad -\frac{y}{t-y^2}, \quad -\frac{x-yz}{(t-y^2)^2}\right)^T$$

so that, by using (3):

$$g = \left(\frac{1}{\mathrm{Var}(Y)}, \quad 0, \quad 0, \quad -\frac{S^X}{\mathrm{Var}(Y)}\right)^T.$$

Hence

$$\begin{aligned}
g^T\Gamma g &= \frac{\mathrm{Var}\left((Y-\mathbb{E}(Y))(Y^X - \mathbb{E}(Y))\right)}{(\mathrm{Var}(Y))^2} + \frac{(S^X)^2}{(\mathrm{Var}(Y))^2}\mathrm{Var}\left((Y-\mathbb{E}(Y))^2\right) \\
&\quad -2\frac{S^X}{(\mathrm{Var}(Y))^2}\mathrm{Cov}\left((Y-\mathbb{E}(Y))(Y^X-\mathbb{E}(Y)), (Y-\mathbb{E}(Y))^2\right) \\
&= \frac{1}{(\mathrm{Var}(Y))^2}\Bigg(\mathrm{Var}\left((Y-\mathbb{E}(Y))(Y^X-\mathbb{E}(Y))\right) + \mathrm{Var}\left(S^X\left((Y-\mathbb{E}(Y))^2\right)\right) \\
&\quad -2\mathrm{Cov}\left((Y-\mathbb{E}(Y))(Y^X-\mathbb{E}(Y)), S^X(Y-\mathbb{E}(Y))^2\right)\Bigg) \\
&= \frac{\mathrm{Var}\left((Y-\mathbb{E}(Y))\left[(Y^X-\mathbb{E}(Y)) - S^X(Y-\mathbb{E}(Y))\right]\right)}{(\mathrm{Var}(Y))^2},
\end{aligned}$$

which is the announced result.

Proof of (12). As in the previous point, it is easy to check that T_N^X is invariant with respect to translations of Y_i and Y_i^X by $-\mathbb{E}(Y)$. Thus, $T_N^X = \Psi\left(\overline{W}_N\right)$ with:

$$\Psi_T(x,y,z) = \frac{x-(y/2)^2}{z/2-(y/2)^2}$$

and:

$$W_i = \left((Y_i - \mathbb{E}(Y))(Y_i^X - \mathbb{E}(Y)), \quad (Y_i - \mathbb{E}(Y)) + (Y_i^X - \mathbb{E}(Y)), \quad (Y_i - \mathbb{E}(Y))^2 + (Y_i^X - \mathbb{E}(Y))^2\right)^T. \tag{30}$$

The result follows from the delta method.

□

Proof of Proposition 2.3. We have that the expressions in (11) and (12) of σ_S^2 and σ_T^2 are translation-invariant, so that we assume without loss of generality that $\mathbb{E}(Y)=0$. By expanding the variances and using the exchangeability of Y and Y^X, we have that $(\mathrm{Var}(Y))^2\left(\sigma_S^2-\sigma_T^2\right)$ is equal to:

$$(\mathrm{Var}(Y))^2\left(\sigma_S^2-\sigma_T^2\right) = \frac{(S^X)^2}{2}\left(\mathrm{Var}(Y^2) - \mathrm{Cov}\left(Y^2,(Y^X)^2\right)\right).$$

We now use Cauchy-Schwarz inequality to see that:

$$\mathrm{Cov}\left(Y^2,(Y^X)^2\right) \le \sqrt{\mathrm{Var}\left(Y^2\right)\mathrm{Var}\left((Y^X)^2\right)} = \mathrm{Var}\left(Y^2\right)$$

so the second term is always non-negative. This proves that the asymptotic variance of S_N^X is greater than the asymptotic variance of T_N^X.

For the equality case, we notice that $S^X=0$ implies the equality of the asymptotic variances. If $S^X \neq 0$, equality holds if and only if there is equality in Cauchy-Schwarz, ie. there exists $k\in\mathbb{R}$ so that:

$$Y^2 = k(Y^X)^2 \text{ almost surely}$$

by taking expectations and using $\mathrm{Var}(Y)=\mathrm{Var}(Y^X)$ we see that $k=1$ necessarily, hence $Y=Y^X$ almost surely, and $S^X=1$ thanks to (3). □

Proof of Lemma 2.6. Let, for $g\in L^2(P)$ and $t\in\mathbb{R}$, P_t^g be the cdf satisfying:

$$\mathrm{d}P_t^g = (1+tg)\mathrm{d}P.$$

It is clear that the tangent set of $\mathcal{P}$ at P is the closure of:

$$\dot{\mathcal{P}}_P = \{g \text{ bounded}, \mathbb{E}(g(Y,Y^X))=0 \text{ and } g(a,b)=g(b,a)\ \forall(a,b)\in\mathbb{R}^2\}.$$

Let, for $Q\in\mathcal{P}$:

$$\Psi_1(Q)=\mathbb{E}_Q\left(\Phi_1(Y)\right) \quad \text{and} \quad \Psi_2(Q)=\mathbb{E}_Q\left(\Phi_2(Y,Y^X)\right).$$

We recall that $\mathbb{E}_Q$ denotes the expectation obtained by assuming that the random vector (Y,Y^X) follows the Q distribution.

Following [32] Section 25.3, we compute the efficient influence functions of Ψ_1 and Ψ_2 with respect to $\mathcal{P}$ and the tangent set $\dot{\mathcal{P}}_P$. These empirical influence functions are related to the minimal asymptotic variance of a regular estimator sequence whose observations lie in $\mathcal{P}$ (op.cit., Theorems 25.20 and 25.21). Let $g\in\dot{\mathcal{P}}_P$.

(1) We have

$$\begin{aligned}\frac{\Psi_1(P_t^g)-\Psi_1(P)}{t} &= \mathbb{E}_P\left(\Phi_1(Y)g(Y,Y^X)\right)\\ &= \mathbb{E}_P\left[\left(\frac{\Phi_1(Y)+\Phi_1(Y^X)}{2}-\mathbb{E}(\Phi_1(Y))\right)g(Y,Y^X).\right]\end{aligned}$$

As:

$$\widetilde{\Psi_{1,P}} = \frac{\Phi_1(Y) + \Phi_1(Y^X)}{2} - \mathbb{E}(\Phi_1(Y)) \in \dot{\mathcal{P}_P},$$

it is the efficient influence function of Ψ_1 at P. Hence the efficient asymptotic variance is:

$$\mathbb{E}_P\left(\left(\widetilde{\Psi_{1,P}}\right)^2\right) = \frac{\mathrm{Var}\left(\Phi_1(Y) + \Phi_1(Y^X)\right)}{4}.$$

As, by the central limit theorem, $\left\{\Phi_N^1\right\}$ clearly achieves this efficient asymptotic variance, it is an asymptotically efficient estimator of $\Psi_1(P)$.

(2) We have:

$$\begin{aligned}\frac{\Psi_2(P_t^g) - \Psi_2(P)}{t} &= \mathbb{E}_P\left(\Phi_2(Y, Y^X) g(Y, Y^X)\right) \\ &= \mathbb{E}_P\left[\left(\Phi_2(Y, Y^X) - \mathbb{E}(\Phi_2(Y, Y^X))\right) g(Y, Y^X)\right].\end{aligned}$$

Thanks to the symmetry of Φ_2, we have that

$$\widetilde{\Psi_{2,P}} = \Phi_2(Y, Y^X) - \mathbb{E}(\Phi_2(Y, Y^X))$$

belongs to $\dot{\mathcal{P}_P}$, hence it is the efficient influence function of Ψ_2. So the efficient asymptotic variance is:

$$\mathbb{E}_P\left(\left(\widetilde{\Psi_{2,P}}\right)^2\right) = \mathrm{Var}\left(\Phi_2(Y, Y^X)\right),$$

and this variance is achieved by $\left\{\Phi_N^2\right\}$. $\square$

Proof of Proposition 2.5. By Lemma 2.6, we get that:

$$U_N = \left(\frac{1}{N}\sum_{i=1}^N Y_i Y_i^X, \quad \frac{1}{N}\sum_{i=1}^N \frac{Y_i + Y_i^X}{2}, \quad \frac{1}{N}\sum_{i=1}^N \frac{Y_i^2 + (Y_i^X)^2}{2}\right) \tag{31}$$

is asymptotically efficient, componentwise, for estimating

$$U = \left(\mathbb{E}(YY^X), \quad \mathbb{E}(Y), \quad \mathbb{E}(Y^2)\right) \tag{32}$$

in $\mathcal{P}$.

Using Theorem 25.50 (efficiency in product space) of [32], we can deduce joint efficiency from this componentwise efficiency.

Now, let Ψ be the function defined by:

$$\Psi(x, y, z) = \frac{x - y^2}{z - y^2}$$

and Ψ is differentiable on:

$$\mathbb{R}^3 \setminus \left\{(x, y, z) \,\middle|\, z \neq y^2\right\},$$

Theorem 25.47 (efficiency and Delta method) of [32] implies that $\{\Psi(U_N)\}$ is asymptotically efficient for estimating $\Psi(U)$ for $P \in \mathcal{P}$. The conclusion follows, as $\Psi(U_N) = T_N^X$ and $\Psi(U) = S^X$. $\square$

Proof of Proposition 3.2. We clearly have that $\widetilde{Y}_N \xrightarrow[N\to+\infty]{L^2} Y + c$.

We deduce that:

$$\mathrm{Var}\left(\widetilde{Y}_N\right) \underset{N\to+\infty}{\longrightarrow} \mathrm{Var}(Y + c) = \mathrm{Var}(Y)$$

and

$$\mathbb{E}(\widetilde{Y}_N | Z) \underset{N\to+\infty}{\longrightarrow} \mathbb{E}(Y|Z) + c \text{ in } L^2.$$

From this last convergence we get

$$\mathrm{Var}\left(\mathbb{E}(\widetilde{Y}_N|Z)\right) \underset{N\to+\infty}{\longrightarrow} \mathrm{Var}\left(\mathbb{E}(Y|Z)\right).$$

This proves that $\widetilde{S}^X = \mathrm{Var}\left(\mathbb{E}(\widetilde{Y}_N|Z)\right)/\mathrm{Var}\left(\widetilde{Y}_N\right)$ converges to $S^X = \mathrm{Var}\left(\mathbb{E}(Y|Z)\right)/\mathrm{Var}(Y)$ when N goes to $+\infty$. □

Proof of Proposition 3.3. **Proof of (17).** Let

$$\widetilde{U}_{N,i} = \left((\widetilde{Y}_{N,i} - \mathbb{E}(Y))(\widetilde{Y}^X_{N,i} - \mathbb{E}(Y)), \widetilde{Y}_{N,i} - \mathbb{E}(Y), \widetilde{Y}^X_{N,i} - \mathbb{E}(Y), \left(\widetilde{Y}_{N,i} - \mathbb{E}(Y)\right)^2\right)$$

and

$$\overline{\widetilde{U}}_N := \frac{1}{N}\sum_{i=1}^{N} \widetilde{U}_{N,i}.$$

Using the Lindeberg-Feller central limit theorem (see e.g. [32] 2.27, with $Y_{N,i} = \widetilde{U}_{N,i}/\sqrt{N}$), we get:

$$\sqrt{N}\left(\overline{\widetilde{U}}_N - \mathbb{E}\left(\widetilde{U}_{N,1}\right)\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_4(0,\Gamma)$$

where Γ is the covariance matrix of the U_1 vector defined in (29).
The use of this central limit theorem is justified by the fact that, under assumption (16) of uniform boundedness of moments of $\widetilde{Y}_N$, there are $s' > 0$ and C' such that:

$$\forall N,\ \ \mathbb{E}(||U_{N,i}||^{2+s'}) < C'$$

where $\|\cdot\|$ is the standard Euclidean norm.
This ensures

$$\forall \epsilon > 0,\ \ \mathbb{E}(||\tilde{U}_{N,i}||^2 \mathbf{1}_{||\tilde{U}_{N,i}||>\epsilon\sqrt{N}}) \to 0.$$

Then

$$\mathbb{E}(||\tilde{U}_{N,i}||^2 \mathbf{1}_{||\tilde{U}_{N,i}||>\epsilon\sqrt{N}}) = \mathbb{E}\left(\frac{||\tilde{U}_{N,i}||^{2+s'}}{||\tilde{U}_{N,i}||^{s'}}\mathbf{1}_{||\tilde{U}_{N,i}||>\epsilon\sqrt{N}}\right) \leq \frac{C'}{\epsilon^{s'}N^{s'/2}}.$$

This shows that for each i, $\left\{\left\|\tilde{U}_{N,i}\right\|^2\right\}_N$ is uniformly integrable, hence, the variance-covariance matrix of $\tilde{U}_{N,i}$ converges to Γ when $N \to +\infty$. As $\widetilde{U}_{N,i} \xrightarrow[N\to+\infty]{\mathbb{P}} U_i$, the same convergence holds in L^2 and the covariance matrices of $\widetilde{U}_{N,i}$ converge (as $N \to +\infty$) to Γ, the covariance matrix of U_i.
We conclude the proof by applying the Delta method as for the exact model (cf. the proof of Proposition 2.2).
Proof of (18). We set:

$$\widetilde{W}_{N,i} = \left((\widetilde{Y}_i - \mathbb{E}(Y))(\widetilde{Y}^X_i - \mathbb{E}(Y)),\quad (\widetilde{Y}_i - \mathbb{E}(Y)) + (\widetilde{Y}^X_i - \mathbb{E}(Y)),\quad (\widetilde{Y}_i - \mathbb{E}(Y))^2 + (\widetilde{Y}^X_i - \mathbb{E}(Y))^2\right)^T.$$

As in the previous point, the Lindeberg-Feller theorem can be applied to $\left\{\widetilde{W}_{N,i}\right\}$ to yield the convergence:

$$\sqrt{N}\left(\overline{\widetilde{W}}_N - \mathbb{E}\left(\widetilde{W}_{1,1}\right)\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_3(0,\Sigma)$$

where Σ is the covariance matrix of W_1 defined in (30). The conclusion follows again by an application of the Delta method as in the proof of Proposition 2.2. □

Proof of Theorem 3.4. The following decompositions:

$$\sqrt{N}(\widetilde{S}^X_N - S^X) = \sqrt{N}(\widetilde{S}^X_N - \widetilde{S}^X) + \sqrt{N}(\widetilde{S}^X - S^X) \tag{33}$$

$$\sqrt{N}(\widetilde{T}_N^X - S^X) = \sqrt{N}(\widetilde{T}_N^X - \widetilde{S}^X) + \sqrt{N}(\widetilde{S}^X - S^X) \tag{34}$$

make obvious that if $\sqrt{N}(\widetilde{S}^X - S^X)$ goes to some constant κ then

$$\sqrt{N}(\widetilde{S}_N - S) \xrightarrow[N\to+\infty]{\mathcal{L}} \mathcal{N}(\kappa, \sigma_S^2)$$

and:

$$\sqrt{N}(\widetilde{T}_N - S) \xrightarrow[N\to+\infty]{\mathcal{L}} \mathcal{N}(\kappa, \sigma_T^2).$$

The second point of the theorem is now clear from the proof of Proposition 3.1.
The remaining of the theorem is an immediate consequence of Lemma 5.1 below. □

Lemma 5.1. *We have:*

$$\sqrt{N}\left(\widetilde{S}^X - S^X\right) = \frac{O\left((N\mathrm{Var}(\delta_N))^{1/2}\right)}{\mathrm{Var}(Y) + o(1)}.$$

Proof of 5.1. We have:

$$\begin{aligned}
\widetilde{S}^X - S^X &= \frac{\mathrm{Cov}(\widetilde{Y}_N, \widetilde{Y}_N^X)}{\mathrm{Var}\widetilde{Y}_N} - \frac{\mathrm{Cov}(Y, Y^X)}{\mathrm{Var}(Y)} \\
&= \frac{\mathrm{Cov}(Y, Y^X) + 2\mathrm{Cov}(Y, \delta_N^X) + \mathrm{Cov}(\delta_N, \delta_N^X)}{\mathrm{Var}(Y) + 2\mathrm{Cov}(Y, \delta_N) + \mathrm{Var}(\delta_N)} - \frac{\mathrm{Cov}(Y, Y^X)}{\mathrm{Var}(Y)} \\
&= \frac{\mathrm{Var}(Y)\left(2\mathrm{Cov}(Y, \delta_N^X) + \mathrm{Cov}(\delta_N, \delta_N^X)\right) - \mathrm{Cov}(Y, Y^X)\left(2\mathrm{Cov}(Y, \delta_N) + \mathrm{Var}(\delta_N)\right)}{\mathrm{Var}(Y)\left(\mathrm{Var}(Y) + 2\mathrm{Cov}(Y, \delta_N) + \mathrm{Var}(\delta_N)\right)} \\
&= \frac{\mathrm{Var}(\delta_N)^{1/2} C_{\delta_N}}{\mathrm{Var}(Y) + 2\mathrm{Cov}(Y, \delta_N) + \mathrm{Var}(\delta_N)}
\end{aligned}$$

and:

$$\mathrm{Var}(Y) + 2\mathrm{Cov}(Y, \delta_N) + \mathrm{Var}(\delta_N) = \mathrm{Var}(Y) + o(1).$$

Finally, $C_{\delta,N}$ is uniformly bounded because $\mathrm{Var}(\delta_N)$ goes to 0 and $\mathrm{Var}(Y)$ is a constant. □

Proof of Proposition 3.6. We will use the following lemma.

Lemma 5.2. *For all $N \in \mathbb{N}^*$, let $(Z_{N,i})_{i=1,\dots,N}$ be a sequence of i.i.d variables such that*

1. $\sqrt{N}\mathbb{E}(Z_{N,i}) \underset{N\to+\infty}{\longrightarrow} 0$;
2. $\mathrm{Var}(Z_{N,i}) \underset{N\to+\infty}{\longrightarrow} 0$.

Then

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N Z_{N,i} \xrightarrow[N\to+\infty]{\mathbb{P}} 0.$$

The lemma follows after the following decomposition:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N Z_{N,i} = \sqrt{N}\left(\frac{1}{N}\sum_{i=1}^N Z_{N,i} - \mathbb{E}(Z_{N,1})\right) + \sqrt{N}\mathbb{E}(Z_{N,1}).$$

Let U_N and U be the vectors defined in the proof of Proposition 2.5, in (31) and (32), respectively, and:

$$\widetilde{U}_N = \left(\frac{1}{N}\sum_{i=1}^N \widetilde{Y}_i\widetilde{Y}_i^X, \quad \frac{1}{N}\sum_{i=1}^N \frac{\widetilde{Y}_i + \widetilde{Y}_i^X}{2}, \quad \frac{1}{N}\sum_{i=1}^N \frac{\widetilde{Y}_i^2 + (\widetilde{Y}_i^X)^2}{2}\right).$$

We will show that:

$$\sqrt{N}\left(U_N - \widetilde{U}_N\right) \xrightarrow[N\to\infty]{\mathbb{P}} 0. \tag{35}$$

By Theorem 25.23 of [32] and the fact that (U_N) is asymptotically efficient for U (shown in the proof of Proposition 2.5), this implies that $\left(\widetilde{U}_N\right)$ is asymptotically efficient for U, and the end of the proof of Proposition 2.5 shows the announced result.

To prove (35), it is sufficient to prove componentwise convergence. We will treat the second and the third components, as the result holds in the same way for the other.

For the second component, we have

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\widetilde{Y}_{N,i}-Y_i)=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\delta_{N,i}$$

goes to 0 (in probability) by the previous lemma. The same holds for $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\widetilde{Y}_{N,i}^X-Y_i^X)$.

For the third component, we have

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\widetilde{Y}_{N,i}^2-Y_i^2)=2\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\delta_{N,i}Y_i+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\delta_{N,i}^2.$$

Now by assumption,

$$\sqrt{N}\mathbb{E}(\delta_{N,i}Y_i)\leq\sqrt{N\mathbb{E}(\delta_{N,i}^2)\mathbb{E}(Y_i^2)}=\sqrt{N(\mathrm{Var}(\delta_{N,i})+\mathbb{E}(\delta_{N,i})^2)\mathbb{E}(Y_i^2)}\to 0,$$

and by Cauchy-Schwarz inequality,

$$\mathrm{Var}(\delta_{N,i}Y_i)=\mathbb{E}(\delta_{N,i}^2Y_i^2)-(\mathbb{E}(\delta_{N,i}Y_i))^2\leq\sqrt{\mathbb{E}(\delta_{N,i}^4)\mathbb{E}(Y_i^4)}+\mathbb{E}(\delta_{N,i}^2)\mathbb{E}(Y_i^2)\leq C\mathbb{E}(\delta_N^4)^{1/2}.$$

By assumption, for all i, $\delta_{N,i}\xrightarrow[N\to+\infty]{\mathbb{P}}0$. Hence, the same convergence holds about $\delta_{N,i}^4$. Since δ_N is in L^{4+s}, then $\left\{\delta_N^4\right\}_N$ is uniformly integrable and we get the convergence of $\mathbb{E}(\delta_N^4)$ to 0 when $N\to+\infty$. We conclude by the lemma above. Again, the same convergence occurs for $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}((\widetilde{Y}_{N,i}^X)^2-(Y_i^X)^2)$. □

Conclusion

We have shown that the Sobol index estimator considered in this paper is asymptotically normal and asymptotically efficient. We also proved that these two properties are robust with respect to the use of a perturbated model, provided that the perturbation has a variance which decays fast enough. The asymptotic normality property can be used to produce approximate confidence intervals for the Sobol indices; we have presented numerical experiments asserting the reliability of these confidence intervals.

This work has been partially supported by the French National Research Agency (ANR) through COSINUS program (project COSTA-BRAVA nr. ANR-09-COSI-015).

References

[1] G.E.P. Box and N.R. Draper. *Empirical model-building and response surfaces.* John Wiley & Sons, 1987.

[2] Rob Carnell. *lhs: Latin Hypercube Samples*, 2009. R package version 0.5.

[3] Wei Chen, Ruichen Jin, and Agus Sudjianto. Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty. *TRANSACTIONS-AMERICAN SOCIETY OF MECHANICAL ENGINEERS JOURNAL OF MECHANICAL DESIGN*, 127(5):875, 2005.

[4] RI Cukier, HB Levine, and KE Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of computational physics*, 26(1):1–42, 1978.

[5] S. Da Veiga and F. Gamboa. Efficient estimation of sensitivity indices. *Journal of Nonparametric Statistics, in press*, 2012.

[6] Garrett M. Dancik. *mlegp: Maximum Likelihood Estimates of Gaussian Processes*, 2011. R package version 3.1.2.

[7] Tristen Hayfield and Jeffrey S. Racine. Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5), 2008.

[8] J.C. Helton, J.D. Johnson, C.J. Sallaberry, and C.B. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, 91(10-11):1175–1209, 2006.

[9] T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, 52(1):1–17, 1996.

[10] I.A. Ibragimov and RZ Has' Minskii. *Statistical estimation–asymptotic theory.* Springer-Verlag, New York, 1981. Applications of Mathematics, Vol. 16.

[11] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis for computer models. In *First International Symposium on Uncertainty Modeling and Analysis Proceedings, 1990.*, pages 398–403. IEEE, 1990.

[12] A. Janon, M. Nodet, and C. Prieur. Certified reduced-basis solutions of viscous Burgers equations parametrized by initial and boundary values. Preprint available at http://hal.inria.fr/inria-00524727/en, 2010, *Accepted in Mathematical modelling and Numerical Analysis.*

[13] A. Janon, M. Nodet, and C. Prieur. Uncertainties assessment in global sensitivity indices estimation from metamodels. Preprint available at http://hal.inria.fr/inria-00567977, 2011, *Accepted in International Journal for Uncertainty Quantification.*

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

[15] WR Madych and SA Nelson. Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation. *Journal of Approximation Theory*, 70(1):94–114, 1992.

[16] A. Marrel, B. Iooss, B. Laurent, and O. Roustant. Calculations of sobol indices for the gaussian process metamodel. *Reliability Engineering & System Safety*, 94(3):742–751, 2009.

[17] H. Monod, C. Naud, and D. Makowski. Uncertainty and sensitivity analysis for crop models. In D. Wallach, D. Makowski, and J. W. Jones, editors, *Working with Dynamic Crop Models: Evaluation, Analysis, Parameterization, and Applications*, chapter 4, pages 55–99. Elsevier, 2006.

[18] Vlad I. Morariu, Balaji Vasan Srinivasan, Vikas C. Raykar, Ramani Duraiswami, and Larry S. Davis. Automatic online tuning for fast gaussian summation. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[19] N.C. Nguyen, K. Veroy, and A.T. Patera. Certified real-time solution of parametrized partial differential equations. *Handbook of Materials Modeling*, pages 1523–1558, 2005.

[20] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[21] J. Racine. An efficient cross-validation algorithm for window width selection for nonparametric kernel regression. *Communications in Statistics Simulation and Computation*, 22:1107–1107, 1993.

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

[23] A. Saltelli, S. Tarantola, Campolongo F., and Ratto M. Sensitivity analysis in practice: a guide to assessing scientific models, 2004.

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

[25] R. Schaback. Mathematical results concerning kernel techniques. In *Prep. 13th IFAC Symposium on System Identification, Rotterdam*, pages 1814–1819. Citeseer, 2003.

[26] M. Scheuerer, R. Schaback, and M. Schlather. Interpolation of spatial data – a stochastic or a deterministic problem ? *Preprint, Universität Göttingen. http://num.math.uni-goettingen.de/schaback/research/papers/IoSD.pdf*, 2011.

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

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

[29] C.B. Storlie, L.P. Swiler, J.C. Helton, and C.J. Sallaberry. Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models.

Reliability Engineering & System Safety, 94(11):1735–1763, 2009.

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

[31] J.Y. Tissot and C. Prieur. A bias correction method for the estimation of sensitivity indices based on random balance designs. *Reliability engineering and systems safety*, 2010.

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