Statistical inference for Sobol pick freeze Monte Carlo method

Fabrice Gamboa, Alexandre Janon, Thierry Klein, Agnes Lagnoux-Renaudie, Clémentine Prieur

Statistical inference for Sobol pick freeze Monte Carlo method

F. Gamboa[*], A. Janon[†], T. Klein[*], A. Lagnoux[*], C. Prieur[‡]

March 26, 2013

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 study asymptotic and non-asymptotic properties of two estimators of Sobol indices. These properties are applied to significance tests and estimation by confidence intervals.

Contents

[*]Institut de mathématiques de Toulouse, Université Toulouse 3

[†]Laboratoire de Sciences Actuarielle et Financière, ISFA, Université Claude Bernard Lyon 1

[‡]Laboratoire Jean Kuntzmann, MOISE/INRIA, Université Joseph Fourier, Grenoble

1 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 on the output. In global stochastic sensitivity analysis (see for example [14] 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 [17]). Each of these partial variances measures the uncertainty on the output induced by each input 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 [15]; 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 [3]. Indeed, many Monte Carlo or quasi Monte Carlo approaches have been developed by the experimental sciences and engineering communities. This includes the Sobol pick-freeze (SPF) scheme (see [15, 16]). 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 [7], the authors have studied the asymptotic behavior of two pick-freeze estimators of a single Sobol index. The results in this paper can be continued in two directions. The first direction is motivated by the fact that in general, so as to rank input variables according to their importance, the pratictioners jointly estimate the collection of all the first-order as well as the total Sobol indices. As these different estimators are dependent, the asymptotic marginal distributions are not fully informative, and one has to characterize the joint law of the estimators. This joint law allows, for example, to perform significance tests and comparisons between different indices, so as to rigorously rank the input variables, taking into account indices estimation errors. The second direction is motivated by the fact that asymptotic distributions are unattainable in practice, hence, non-asymptotic tools (such as concentration inequalities, and Berry-Esseen-like theorems) about the distribution of the Sobol indices estimators should be investigated. Such results will allow conservative certification for the index estimates.

This paper is organized as follows: in Section 2, we review the Sobol pick-freeze method and give the estimators that are studied in the paper. In Section 3, we prove a central limit theorem which gives the joint asymptotic distribution of any closed Sobol index [14], which in particular can be used to explicit the asymptotic distribution of all first-order and total index estimators. We then apply this central limit theorem to significance and comparison tests on Sobol indices. Sections 4 and 5 are dedicated to non-asymptotic studies of the distribution of a single Sobol index estimator. These two sections, respectively, give concentration inequalities and Berry-Esseen bounds. All our theoretical results are numerically illustrated on model examples.

2 Sobol pick freeze Monte Carlo method

2.1 Black box model and Sobol indices

In the whole paper, we consider a non necessarily linear regression model connecting an output $Y \in \mathbb{R}$ to independent random input vectors $X_1, \dots X_p$ with for $i = 1, \dots p$, X_i belongs to some probability

space $\mathcal{X}_i$. We denote

$$Y = f(X) := f(X_1, \dots, X_p) \tag{1}$$

where f is a deterministic real valued measurable function defined on $\mathcal{X} = \mathcal{X}_1 \times \dots \mathcal{X}_p$. We assume that Y is square integrable and non deterministic ($\mathrm{Var} Y \neq 0$).

Let $\mathbf{u} := (u_1, \dots, u_k)$ be k subsets of $I_p := \{1, \dots, p\}$. The vector of closed Sobol indices (see [14]) is then

$$S^{\mathbf{u}}_{\mathrm{Cl}} := \left(\frac{\mathrm{Var}(\mathbb{E}(Y|X_i, i \in u_1))}{\mathrm{Var}(Y)}, \dots, \frac{\mathrm{Var}(\mathbb{E}(Y|X_i, i \in u_k))}{\mathrm{Var}(Y)} \right).$$

As pointed out and discussed in the Introduction, Sobol indices are useful quantities widely used in engineering and applied sciences in the context of prioritisation of influent input variables of a complicated computer simulation code (see for example [14], [2]) and our paper gives a rigourous statistical analysis of these quantities. Notice that considering the whole vector $S^{\mathbf{u}}_{\mathrm{Cl}}$ allows estimation of asymptotic confidence regions and tests for joint significance (see Section 3).

2.2 Monte Carlo estimation of S: Sobol pick freeze method

For X and for any subset v of I_p we define X^v by the vector such that $X^v_i = X_i$ if $i \in v$ and $X^v_i = X'_i$ if $i \notin v$ where X'_i is an independent copy of X_i. We then set

$$Y^v := f(X^v).$$

The next lemma [7, Lemma 1.2] shows how to express $S^{\mathbf{u}}_{\mathrm{Cl}}$ in terms of covariances. This will lead to a natural estimator:

Lemma 2.1. *For any $u \subset I_p$, one has*

$$\mathrm{Var}(\mathbb{E}(Y|X_i, i \in u)) = \mathrm{Cov}\left(Y, Y^{\mathbf{u}}\right). \tag{2}$$

An estimator with a close expression has been considered in [5].

Notation

From now on, we will denote $\mathrm{Var}(Y)$ by V, $\mathrm{Cov}(Y, Y^{\mathbf{u}})$ by C_u and $\overline{Z}_N$ the empirical mean of any N-sample $(Z_1, \dots, Z_N)$ of Z.

A first estimation for $S^{\mathbf{u}}_{\mathrm{Cl}}$. In view of Lemma 2.1, we are now able to define a first natural estimator of $S^{\mathbf{u}}_{\mathrm{Cl}}$ (all sums are taken for i from 1 to N):

$$S^{\mathbf{u}}_{N,\mathrm{Cl}} = \left(\frac{\frac{1}{N}\sum Y_i Y_i^{u_1} - \left(\frac{1}{N}\sum Y_i\right)\left(\frac{1}{N}\sum Y_i^{u_1}\right)}{\frac{1}{N}\sum Y_i^2 - \left(\frac{1}{N}\sum Y_i\right)^2}, \dots, \frac{\frac{1}{N}\sum Y_i Y_i^{u_k} - \left(\frac{1}{N}\sum Y_i\right)\left(\frac{1}{N}\sum Y_i^{u_k}\right)}{\frac{1}{N}\sum Y_i^2 - \left(\frac{1}{N}\sum Y_i\right)^2} \right). \tag{3}$$

These estimators have been considered in [5], where it has been showed to be practically efficient estimators.

A second estimation for $S^{\mathbf{u}}_{\mathrm{Cl}}$. Since the observations consist in $(Y_i, Y_i^{u_1}, \dots, Y_i^{u_k})_{(1 \leq i \leq N)}$, a more precise estimation of the first and second moments can be done and we are able to define a second estimator of $S^{\mathbf{u}}_{\mathrm{Cl}}$ taking into account all the available information. Define

$$Z^{\mathbf{u}}_i = \frac{1}{k+1}\left(Y_i + \sum_{j=1}^{k} Y_i^{u_j}\right), \quad M^{\mathbf{u}}_i = \frac{1}{k+1}\left(Y_i^2 + \sum_{j=1}^{k} (Y_i^{u_j})^2\right).$$

The second estimator is then defined as

$$T^{\mathbf{u}}_{N,\mathrm{Cl}} = \left(\frac{\frac{1}{N}\sum Y_i Y_i^{u_1} - \left(\frac{1}{2N}\sum (Y_i + Y_i^{u_1})\right)^2}{\frac{1}{N}\sum M^{\mathbf{u}}_i - \left(\frac{1}{N}\sum Z^{\mathbf{u}}_i\right)^2}, \dots, \frac{\frac{1}{N}\sum Y_i Y_i^{u_k} - \left(\frac{1}{2N}\sum (Y_i + Y_i^{u_k})\right)^2}{\frac{1}{N}\sum M^{\mathbf{u}}_i - \left(\frac{1}{N}\sum Z^{\mathbf{u}}_i\right)^2} \right). \tag{4}$$

This estimator (in the $k=1$ case) was first introduced by Monod in [9] and Janon et al. studied its asymptotic properties (CLT, efficiency) in [7]. In [11, 10] Owen introduces new estimators for Sobol indices and compares numerically their performances. The delta method can also be used on these pick-freeze estimators to derive their asymptotic properties.

Remark 2.2. *One could use all the information available in the sample by defining the following estimator:*

$$\left(\frac{\frac{1}{N}\sum Y_i Y_i^{u_1} - \left(\frac{1}{N}\sum Z_i^u\right)^2}{\frac{1}{N}\sum M_i^u - \left(\frac{1}{N}\sum Z_i^u\right)^2}, \ldots, \frac{\frac{1}{N}\sum Y_i Y_i^{u_k} - \left(\frac{1}{N}\sum Z_i^u\right)^2}{\frac{1}{N}\sum M_i^u - \left(\frac{1}{N}\sum Z_i^u\right)^2} \right).$$

However, our empirical studies show that this estimator has a larger variance than $T^{\mathbf{u}}_{N,\mathrm{Cl}}$.

3 Joint CLT for Sobol index estimates with applications to significance tests

3.1 Main results

Theorem 3.1. *Assume that* $\mathbb{E}(Y^4) < \infty$. *Then:*

1. $$\sqrt{N}\left(S^{\mathbf{u}}_{N,\mathrm{Cl}} - S^{\mathbf{u}}_{\mathrm{Cl}}\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}_k\left(0, \Gamma_{\mathbf{u},S}\right) \tag{5}$$

   where $\Gamma_{\mathbf{u},S} = \left((\Gamma_{\mathbf{u},S})_{l,j}\right)_{1\le l,j\le k}$ *with*

   $$(\Gamma_{\mathbf{u},S})_{l,j} = \frac{\mathrm{Cov}(YY^{u_l}, YY^{u_j}) - S^{u_l}_{\mathrm{Cl}}\mathrm{Cov}(YY^{u_j}, Y^2) - S^{u_j}_{\mathrm{Cl}}\mathrm{Cov}(YY^{u_l}, Y^2) + S^{u_j}_{\mathrm{Cl}} S^{u_l}_{\mathrm{Cl}}\mathrm{Var}(Y^2)}{\left(\mathrm{Var}(Y)\right)^2}$$

2. $$\sqrt{N}\left(T^{\mathbf{u}}_{N,\mathrm{Cl}} - S^{\mathbf{u}}_{\mathrm{Cl}}\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}_k\left(0, \Gamma_{\mathbf{u},T}\right) \tag{6}$$

   where $\Gamma_{\mathbf{u},T} = \left((\Gamma_{\mathbf{u},T})_{l,j}\right)_{1\le l,j\le k}$ *with*

   $$(\Gamma_{\mathbf{u},T})_{l,j} = \frac{\mathrm{Cov}(YY^{u_l}, YY^{u_j}) - S^{u_l}_{\mathrm{Cl}}\mathrm{Cov}(YY^{u_j}, M^{\mathbf{u}}) - S^{u_j}_{\mathrm{Cl}}\mathrm{Cov}(YY^{u_l}, M^{\mathbf{u}}) + S^{u_j}_{\mathrm{Cl}} S^{u_l}_{\mathrm{Cl}}\mathrm{Var}(M^{\mathbf{u}})}{\left(\mathrm{Var}(Y)\right)^2}.$$

3.2 Some particular cases

1. Assume $k=p$, $u=(\{1\},\ldots,\{p\})$ and $\mathbb{E}(Y^4)<\infty$. We denote $Y_i^{\{j\}}$ by Y_i^j. Here

   $$S^{\mathbf{u}}_{\mathrm{Cl}} = \left(\frac{\mathrm{Var}(\mathbb{E}(Y|X_1))}{\mathrm{Var}(Y)}, \ldots, \frac{\mathrm{Var}(\mathbb{E}(Y|X_p))}{\mathrm{Var}(Y)}\right)$$

   and

   $$T^{\mathbf{u}}_{N,\mathrm{Cl}} = \left(\frac{\frac{1}{N}\sum Y_i Y_i^1 - \left(\frac{1}{2N}\sum (Y_i + Y_i^1)\right)^2}{\frac{1}{N}\sum M_i^{\mathbf{u}} - \left(\frac{1}{N}\sum Z_i^{\mathbf{u}}\right)^2}, \ldots, \frac{\frac{1}{N}\sum Y_i Y_i^p - \left(\frac{1}{2N}\sum (Y_i + Y_i^p)\right)^2}{\frac{1}{N}\sum M_i^{\mathbf{u}} - \left(\frac{1}{N}\sum Z_i^{\mathbf{u}}\right)^2} \right).$$

   The CLT becomes

   $$\sqrt{N}\left(T^{\mathbf{u}}_{N,\mathrm{Cl}} - S^{\mathbf{u}}_{\mathrm{Cl}}\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}_p\left(0, \Gamma_{\mathbf{u},T}\right)$$

   where $\Gamma_{\mathbf{u},T} = \left((\Gamma_{\mathbf{u},T})_{l,j}\right)_{1\le l,j\le k}$ with

   $$\left(\mathrm{Var}(Y)\right)^2 (\Gamma_{\mathbf{u},T})_{l,j} = \mathrm{Cov}(YY^l, YY^j) - S^l_{\mathrm{Cl}}\mathrm{Cov}(YY^j, M^{\mathbf{u}}) - S^j_{\mathrm{Cl}}\mathrm{Cov}(YY^l, M^{\mathbf{u}}) + S^j_{\mathrm{Cl}} S^l_{\mathrm{Cl}}\mathrm{Var}(M^{\mathbf{u}}).$$

2. We can obviously have a CLT for any index of order 2. Indeed if we take $k=1$ and $(i,j)\in\{1,\dots,p\}^2$ with $i\neq j$ and $u=\{i,j\}$. We get $Z^{\mathbf{u}}=\frac{1}{2}\left(Y+Y^{\mathbf{u}}\right)$ and $M^{\mathbf{u}}=\frac{1}{2}\left(Y^2+(Y^{\mathbf{u}})^2\right)$; thus

$$S^{\mathbf{u}}_{\mathrm{Cl}}=\frac{\mathrm{Var}(\mathbb{E}(Y|X_i,X_j))}{\mathrm{Var}(Y)}\text{ and }T^{\mathbf{u}}_{N,\mathrm{Cl}}=\frac{\frac{1}{N}\sum Y_iY_i^{\mathbf{u}}-\left(\frac{1}{2N}\sum(Y_i+Y_i^{\mathbf{u}})\right)^2}{\frac{1}{2N}\sum\left(Y^2+(Y^{\mathbf{u}})^2\right)-\left(\frac{1}{2N}\sum(Y_i+Y_i^{\mathbf{u}})\right)^2}.$$

The CLT becomes

$$\sqrt{N}\left(T^{\mathbf{u}}_{N,\mathrm{Cl}}-S^{\mathbf{u}}_{\mathrm{Cl}}\right)\underset{N\to\infty}{\overset{\mathcal{L}}{\to}}\mathcal{N}_1\left(0,\Gamma_{\mathbf{u},T}\right)$$

with

$$\left(\mathrm{Var}(Y)\right)^2\left(\Gamma_{\mathbf{u},T}\right)=\mathrm{Var}(YY^{\mathbf{u}})-2S^{\mathbf{u}}_{\mathrm{Cl}}\mathrm{Cov}(YY^{\mathbf{u}},Y^2)+\frac{\left(S^{\mathbf{u}}_{\mathrm{Cl}}\right)^2}{2}\left(\mathrm{Var}(Y^2)+\mathrm{Cov}(Y^2,(Y^{\mathbf{u}})^2)\right).$$

3. One can also straightforwardly deduce the joint distribution of the vector of all indices of order 2. For example, if $p=3$ take $k=3$ and $\mathbf{u}=(\{1,2\},\{1,3\},\{2,3\})$ and apply Theorem 3.1.

3.3 Proof of Theorem 3.1

Since $S^{\mathbf{u}}_{N,\mathrm{Cl}}$ and $T^{\mathbf{u}}_{N,\mathrm{Cl}}$ are invariant by any centering (translation) of the Y_i's and $Y_i^{u_j}$'s for $j=1,\dots,k$, we can simplify the next calculations translating by $\mathbb{E}(Y)$. For the sake of simplicity, Y_i and $Y_i^{u_j}$ now denote the centered random variables.

Proof of (5) :
Recall that

$$S^{\mathbf{u}}_{N,\mathrm{Cl}}-S^{\mathbf{u}}_{\mathrm{Cl}}=\left(\frac{\frac{1}{N}\sum Y_iY_i^{u_1}-\left(\frac{1}{N}\sum Y_i\right)\left(\frac{1}{N}\sum Y_i^{u_1}\right)}{\frac{1}{N}\sum Y_i^2-\left(\frac{1}{N}\sum Y_i\right)^2}-S^{u_1}_{\mathrm{Cl}},\dots,\frac{\frac{1}{N}\sum Y_iY_i^{u_k}-\left(\frac{1}{N}\sum Y_i\right)\left(\frac{1}{N}\sum Y_i^{u_k}\right)}{\frac{1}{N}\sum Y_i^2-\left(\frac{1}{N}\sum Y_i\right)^2}-S^{u_k}_{\mathrm{Cl}}\right).$$

Let $W_i=(Y_iY_i^{u_j},j=1,\dots,k,Y_i,Y_i^{u_j},j=1\dots,k,Y_i^2)^t$ $(i=1,\dots)$ and g the mapping from $\mathbb{R}^{2k+2}$ to $\mathbb{R}^k$ defined by

$$g(x_1,\dots,x_k,y,y_1,\dots,y_k,z)=\left(\frac{x_1-yy_1}{z-y^2},\dots,\frac{x_k-yy_k}{z-y^2}\right).$$

Let Σ denote the covariance matrix of W_i. The vectorial central limit theorem implies that

$$\sqrt{N}\left(\frac{1}{N}\sum W_i-\mathbb{E}(W)\right)\underset{N\to\infty}{\overset{\mathcal{L}}{\to}}\mathcal{N}_{2k+2}\left(0,\Sigma\right)$$

We then apply the so-called Delta method [17] to W and g so that

$$\sqrt{N}\left(g(\overline{W}_N)-g(\mathbb{E}(W))\right)\underset{N\to\infty}{\overset{\mathcal{L}}{\to}}\mathcal{N}\left(0,J_g(\mathbb{E}(W))\Sigma J_g(\mathbb{E}(W))^t\right)$$

with $J_g(\mathbb{E}(W))$ the Jacobian of g at point $\mathbb{E}(W)$.
Define $(g_1,\dots,g_k):=\varphi$. For $i=1,\dots,k$, $j=1,\dots,k$,

$$\begin{cases}\frac{\partial g_j}{\partial x_i}(\mathbb{E}(W))=\frac{1}{V}\delta_{i,j}\\ \frac{\partial g_j}{\partial y}(\mathbb{E}(W))=0\\ \frac{\partial g_j}{\partial y_i}(\mathbb{E}(W))=0\\ \frac{\partial g_j}{\partial z}(\mathbb{E}(W))=-\frac{S^{u_j}_{\mathrm{Cl}}}{V}\end{cases}$$

with $\delta_{i,i}=1$ and $\delta_{i,j}=0$ if $i\neq j$. Thus $\Gamma_{\mathbf{u},S}=J_g(\mathbb{E}(W))\Sigma J_g(\mathbb{E}(W))^t$ is as stated in Theorem 3.1.

Proof of (6) :
The proof is similar to the one of (5). We now define $W_i=(Y_iY_i^{u_j},j=1,\dots,k,Y_i,Y_i^{u_j},j=1\dots,k,\overline{(Y_i^{\mathbf{u}})^2})^t$. We apply the delta method to g from $\mathbb{R}^{2k+2}$ into $\mathbb{R}^k$ defined by

$$g(x_1,\dots,x_k,y,y_1,\dots,y_k,z)=\left(\frac{x_1-\left(\frac{y+y_1}{2}\right)^2}{z-\left(\frac{y+y_1+\dots+y_k}{k+1}\right)^2},\dots,\frac{x_k-\left(\frac{y+y_k}{2}\right)^2}{z-\left(\frac{y+y_1+\dots+y_k}{k+1}\right)^2}\right).$$

For $i = 1, \dots, k$, $j = 1, \dots, k$,

$$\begin{cases} \frac{\partial g_j}{\partial x_i} u(\mathbb{E}(W)) = \frac{1}{V}\delta_{i,j} \\ \frac{\partial g_j}{\partial y}(\mathbb{E}(W)) = 0 \\ \frac{\partial g_j}{\partial y_i}(\mathbb{E}(W)) = 0 \\ \frac{\partial g_j}{\partial z}(\mathbb{E}(W)) = -\frac{S_{\mathrm{Cl}}^{u_j}}{V}. \end{cases}$$

3.4 Significance tests

In order to simplify the notation we will write the vectors $S_{\mathrm{Cl}}^{\mathbf{u}}$ as column vectors. In this section, we give a general procedure to build significance tests of level α and then illustrate this procedure on two examples.

Let $\mathbf{u} := (u_1, \dots, u_k)$ so that for any $i = 1, \dots, k$, u_i is a subset of $I_p := \{1, \dots, p\}$. Similarly, let $\mathbf{v} := (v_1, \dots, v_l)$ and $\mathbf{w} := (w_1, \dots, w_l)$ be l be so that for any $i = 1, \dots, l$, $v_i \subseteq I_p$ and $w_i \subseteq I_p$.

Consider the following general testing problem

$$H_0 : S_{\mathrm{Cl}}^{\mathbf{u}} = 0 \text{ and } S_{\mathrm{Cl}}^{\mathbf{v}} = S_{\mathrm{Cl}}^{\mathbf{w}} \quad \text{against} \quad H_1 : H_0 \text{ is not true.}$$

Remark 3.2. *Note that one can also test*

$$H_0 : S_{\mathrm{Cl}}^{\mathbf{u}} \leq s \quad \textit{against} \quad H_1 : S_{\mathrm{Cl}}^{\mathbf{u}} > s,$$

or

$$H_0 : S_{\mathrm{Cl}}^{\mathbf{u}} \leq S_{\mathrm{Cl}}^{\mathbf{v}} \quad \textit{against} \quad H_1 : S_{\mathrm{Cl}}^{\mathbf{u}} > S_{\mathrm{Cl}}^{\mathbf{v}}.$$

Appling Theorem 3.1 we have

$$G_N := \sqrt{N}\left(\begin{pmatrix} S_{N,\mathrm{Cl}}^{\mathbf{u}} \\ S_{N,\mathrm{Cl}}^{\mathbf{v}} - S_{N,\mathrm{Cl}}^{\mathbf{w}} \end{pmatrix} - \begin{pmatrix} S_{\mathrm{Cl}}^{\mathbf{u}} \\ S_{\mathrm{Cl}}^{\mathbf{v}} - S_{\mathrm{Cl}}^{\mathbf{w}} \end{pmatrix}\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}_{k+l}(0, \Gamma). \tag{7}$$

Since we have an explicit expression of Γ we may build an estimator Γ_N of Γ thanks to empirical means. Note that $(\Gamma_N)_N$ converges a.s. to Γ. Define

$$\widetilde{G}_N := \sqrt{N}\begin{pmatrix} S_{N,\mathrm{Cl}}^{\mathbf{u}} \\ S_{N,\mathrm{Cl}}^{\mathbf{v}} - S_{N,\mathrm{Cl}}^{\mathbf{w}} \end{pmatrix}.$$

Then:

$$G_N = \widetilde{G}_N - \begin{pmatrix} S_{\mathrm{Cl}}^{\mathbf{u}} \\ S_{\mathrm{Cl}}^{\mathbf{v}} - S_{\mathrm{Cl}}^{\mathbf{w}} \end{pmatrix}.$$

Corollary 3.3. *Under H_0, $\widetilde{G}_N \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}_{k+l}(0, \Gamma)$.*
Under H_1, $|\widetilde{G}_N(1)| + |\widetilde{G}_N(2)| \underset{N\to\infty}{\overset{a.s.}{\to}} \infty$.

This corollary allows us to construct several tests. It is a well-known fact that in the case of a vectorial null hypothesis "there exists no uniformly most powerful test, not even among the unbiased tests" (see Chapter 15 in [17]). In practice, we return to the dimension 1 introducing a function $F : \mathbb{R}^{k+l} \to \mathbb{R}$ and testing $H_0(F) : F(h) = 0$ (respectively $H_1(F) : F(h) \neq 0$) instead of $H_0 : h = 0$ (resp. $H_1 : h \neq 0$). The choice of a reasonable test "depends on the alternatives at which we wish a high power".

Remark 3.4. *If we take as test statistic $T_N = A\widetilde{G}_N$ where A is a linear form defined on $\mathbb{R}^{l+k}$, under H_0, $T_N \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, A\Gamma A')$. Replacing Γ by Γ_N and using Slutsky's lemma we get*

$$(A\Gamma_N A')^{-1/2} T_N \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, 1).$$

Thus we reject H_0 if $(A\Gamma_N A')^{-1/2} T_N \geq z_\alpha$ where z_α is the $1-\alpha$ quantile of a standard Gaussian random variable.
One can have a similar result when A is not anymore linear but only C^1 by applying the so-called Delta method.

3.4.1 Numerical applications: toy examples

Example 1 In this first toy example, we compare 5 different test statistics through their power function. Let $X = (X_1, X_2) \sim \mathcal{N}(0, I_2)$, and

$$Y = f(X) = \lambda_1 X_1 + \lambda_1 X_2 + \lambda_2 X_1 X_2,$$

with $2\lambda_1^2 + \lambda_2^2 = 1$. We consider here the following testing problem

$$H_0 : S^{\mathbf{1}}_{\mathrm{Cl}} = S^{\mathbf{2}}_{\mathrm{Cl}} = \lambda_1^2 = 0 \quad \text{against} \quad H_1 : \ \lambda_1 \neq 0.$$

Then, computations lead to

$$\Gamma(1,1) = \Gamma(2,2) = 3 - 2\lambda_1^2 - 11\lambda_1^4 + 24\lambda_1^6 - 24\lambda_1^8$$
$$\Gamma(2,1) = \Gamma(1,2) = -7\lambda_1^4 + 24\lambda_1^6 - 24\lambda_1^8.$$

The Gaussian limit in Theorem 3.1 is $\mathcal{N}_2(0, 3Id_2)$ under H_0 while it is asymptotically distributed as $\mathcal{N}_2(0, \Gamma)$ under H_1.

Test 1: we take as test statistic $T_{N,1} = \widetilde{G}_N(1) + \widetilde{G}_N(2)$.
Under H_0, $T_{N,1} \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, 6)$ so we reject H_0 if $T_{N,1} > z_\alpha$ where $z_\alpha/\sqrt{6}$ is the $(1-\alpha)$ quantile of a standard Gaussian random variable.While under H_1, following the procedure of Remark 3.4 with $A = (1\ 1)$.

$$\left(T_{N,1} - 2\sqrt{N}\lambda_1^2\right) / (2[\Gamma(1,1) + \Gamma(1,2)])^{1/2} \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0,1).$$

It is then easy to compute the theoretical power function. In Figure 1 we plot this theoretical function called **true power fct t1** and the empirical power function called **estimated power fct t1**. To compute the empirical power function we didn't assume the knowledge of the matrix Γ nor the one of the function f.

Test 2: since the Sobol indices are non negative, the testing problem is naturally unilateral. However in view of more general contexts we introduce the test statistic $T_{N,2} = |\widetilde{G}_N(1)| + |\widetilde{G}_N(2)|$. We reject H_0 if $T_{N,2} > z_\alpha$ where $z_\alpha/\sqrt{3}$ is the $(1-\alpha)$ quantile of the random variable having

$$\frac{2}{\sqrt{\pi}} e^{-u^2/4} \Phi(u/\sqrt{2}) \mathbb{1}_{\mathbb{R}_+}(u)$$

as density (Φ being the distribution function of a standard Gaussian random variable). Under H_1, the power function of $T_{N,2}$ and the limit variance are estimated using Monte Carlo technics. In Figure 1 we plot this empirical power function called **estimated power fct t2**.

Test 3: in the same spirit, we introduce the test statistic $T_{N,3} = |\widetilde{G}_N(1) + \widetilde{G}_N(2)|$. We reject H_0 if $T_{N,3} > z_\alpha$ where $z_\alpha/\sqrt{6}$ is the $(1-\alpha/2)$ quantile of a standard Gaussian random variable.Under H_1, the power function of $T_{N,3}$ and the limit variance are estimated using Monte Carlo technics. In Figure 1 we plot this empirical power function called **estimated power fct t3**.

Test 4: we use the L^2 norm and consider $T_{N,4} = (G_N(1))^2 + (G_N(2))^2$. Under H_0, $T_{N,4}/3 \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \chi_2(2)$ so we reject H_0 if $T_{N,4} > z_\alpha$ where $z_\alpha/3$ is the $(1-\alpha)$ quantile of a χ_2 random variablewith 2 degrees of freedom. Under H_1, the power function of $T_{N,4}$ and the limit variance are estimated using Monte Carlo technics. We plot this empirical power function in Figure 1 called **estimated power fct t4**.

Test 5: we use the infinity norm and consider $T_{N,5} = max(|G_N(1)|; |G_N(2)|)$. We reject H_0 if $T_{N,5} > z_\alpha$ where $z_\alpha/\sqrt{3}$ is the $[1+\sqrt{1-\alpha}]/2$ quantile of a standard Gaussian random variable.Under H_1, the power function of $T_{N,5}$ and the limit variance are estimated using Monte Carlo technics. In Figure 1 we plot this theoretical function called **true power fct t5** and the empirical power function called **estimated power fct t5**.

In Figure 1 we thus present the plot of the different power functions for $N = 100,\ 500$ and 1000.

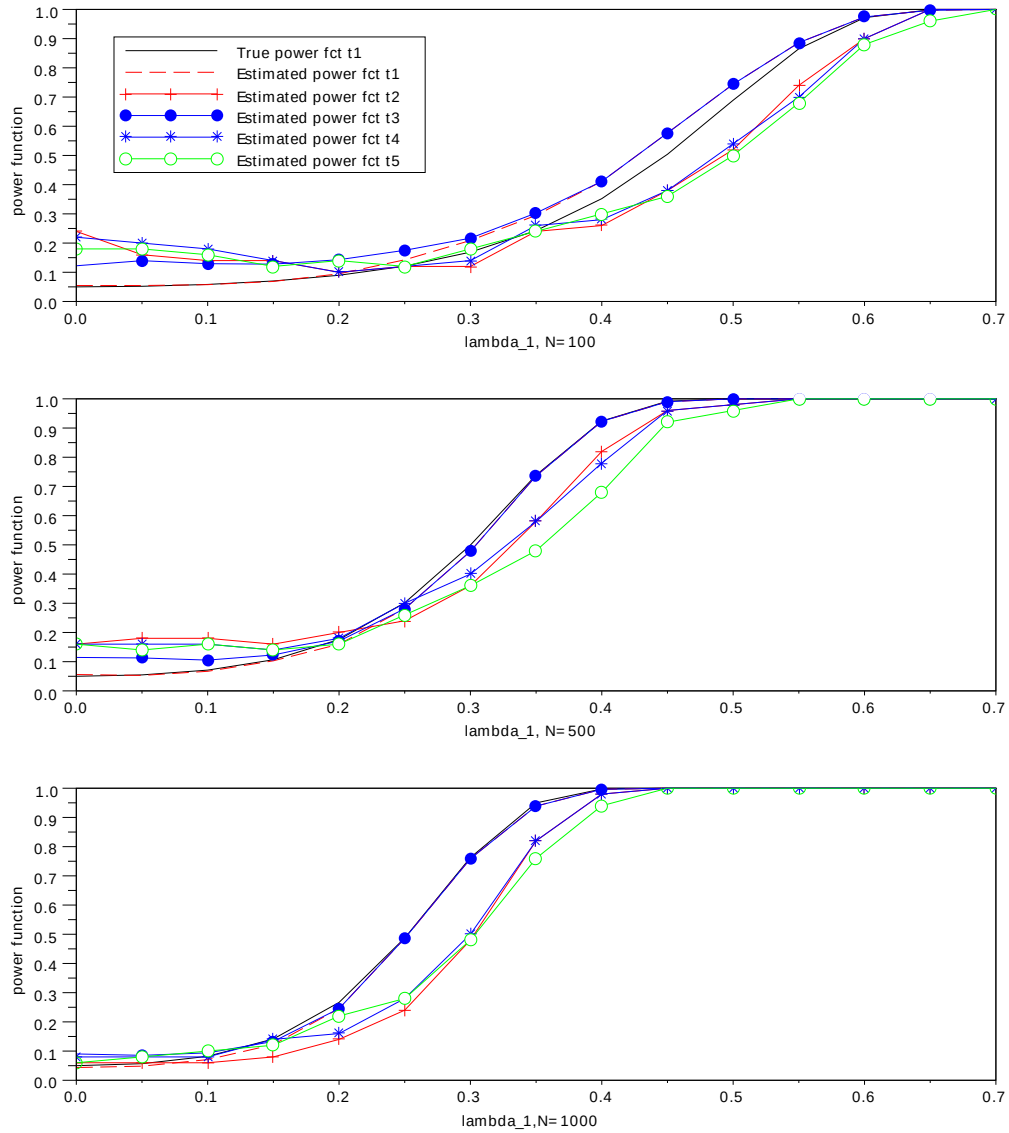

Figure 1: Power functions

Example 2 Let $X = (X_1, X_2, X_3) \sim \mathcal{N}(0, I_3)$, $2\lambda_1^2 + \lambda_2^2 = 1$ and

$$Y = f(X) = \lambda_1(X_2 + X_3) + \lambda_2 X_1 X_2.$$

Let us test if X_1 has any influence ie $H_0 : S_{\mathrm{Cl}}^{\{\mathbf{1}\}} = 0$, $S_{\mathrm{Cl}}^{\{\mathbf{1},\mathbf{2}\}} = S_{\mathrm{Cl}}^{\{\mathbf{2}\}}$ and $S_{\mathrm{Cl}}^{\{\mathbf{1},\mathbf{3}\}} = S_{\mathrm{Cl}}^{\{\mathbf{3}\}}$. Applying Theorem 3.1 we easily get

$$G_N := \sqrt{N}\left(\begin{pmatrix} S_{N,\mathrm{Cl}}^1 \\ S_{N,\mathrm{Cl}}^{1,2} - S_{N,\mathrm{Cl}}^2 \\ S_{N,\mathrm{Cl}}^{1,3} - S_{N,\mathrm{Cl}}^3 \end{pmatrix} - \begin{pmatrix} S_{\mathrm{Cl}}^1 \\ S_{\mathrm{Cl}}^{1,2} - S_{\mathrm{Cl}}^2 \\ S_{\mathrm{Cl}}^{1,3} - S_{\mathrm{Cl}}^3 \end{pmatrix}\right) \underset{N\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}_3(0, \Gamma).$$

Here under H_0 the covariance limit Γ in Theorem 3.1 is the identity matrix. Under H_1 we use its explicit expression given in Theorem 3.1 to compute an empirical estimator Γ_N. We compare Test 1, Test 3, Test 4 and Test 5 defined in the previous example. We present in Figure 2 the plot of the different estimated power functions for $N = 100,\ 500$ and 1000.

Figures 1 and 2 show, as expected, that increasing N leads to a steeper power function (hence, a better discrimination between the hypothesis), and that the estimated power function gets closer to

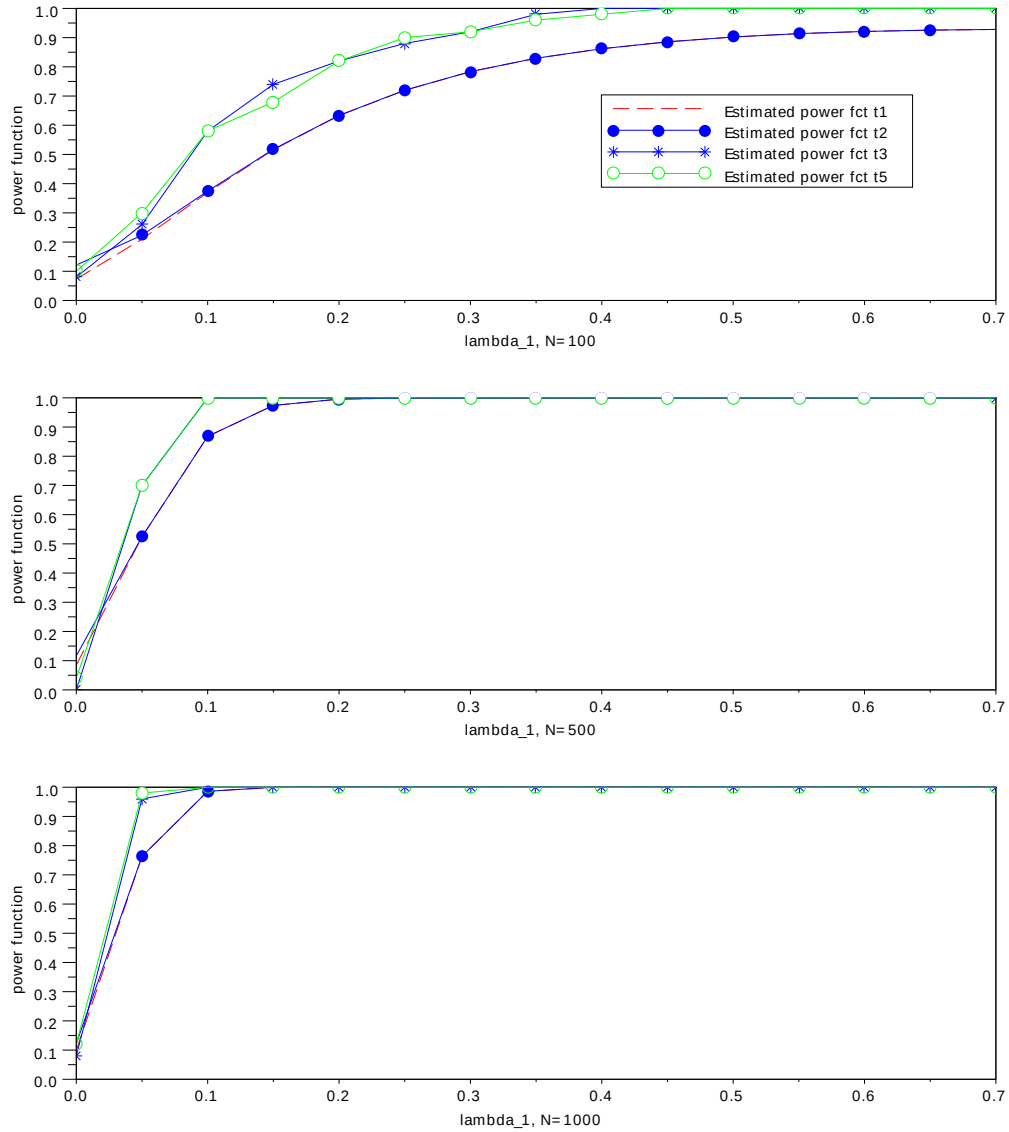

Figure 2: Estimated power functions for different values of N.

the true one. We also see that no test is the most powerful, uniformly in λ_1, in accordance with the theory quoted above.

Ishigami function The Ishigami model [6] is given by:

$$Y = f(X_1, X_2, X_3) = \sin X_1 + 7 \sin^2 X_2 + 0.1 X_3^4 \sin X_1 \tag{8}$$

for $(X_j)_{j=1,2,3}$ are i.i.d. uniform random variables in $[-\pi; \pi]$. Exact values of these indices are analytically known:

$$S_{\mathrm{Cl}}^{\{1\}} = 0.3139, \quad S_{\mathrm{Cl}}^{\{2\}} = 0.4424, \quad S_{\mathrm{Cl}}^{\{3\}} = 0.$$

We perform simulations in order to show that our test procedure allows us to recover the fact that $S_{\mathrm{Cl}}^{\{3\}} = 0$, even for relatively small values of N. In Table 1, we present the simulated confidence levels obtained for $N \in \{10, 50, 100, 500, 1000\}$ by the following procedure. For each value of N, we use a 1000 sample to estimate the confidence level and we repeat this scheme 20 times. We give in Table 1 the minimum, the mean and the maximum of these 20 distinct simulated values of the confidence levels.

N	Min	Mean	Max
10	0.041	0.0463	0.048
50	0.042	0.0482	0.050
100	0.044	0.0489	0.051
500	0.047	0.0510	0.053
1000	0.049	0.0510	0.055

Table 1: Results for the Ishigami function

3.4.2 Numerical applications: a real test case

It is customary in aeronautics to model the fuel mass needed to link two fixed countries with a commercial aircraft by the Bréguet formula:

$$M_{fuel} = (M_{empty} + M_{pload}) \left(e^{\frac{SFC \cdot g \cdot Ra}{V \cdot F} 10^{-3}} - 1 \right) . \tag{9}$$

See [13] for the description of the model with more details.
The fixed variables are

- M_{empty} : *Empty weight* = basic weight of the aircraft (excluding fuel and passengers)
- M_{pload} : *Payload* = maximal carrying capacity of the aircraft
- g : Gravitational constant
- Ra : *Range* = distance traveled by the aircraft

The uncertain variables are

- V : *Cruise speed* = aircraft speed between ascent and descent phase
- F : *Lift-to-drag ratio* = aerodynamic coefficient
- SFC : *Specific Fuel Consumption* = characteristic value of engines

We follow [13] and model the uncertainties as presented in Table 2.

variable	density	parameter
V	*Uniform*	(V_{min}, V_{max})
F	*Beta*	$(7, 2, F_{min}, F_{max})$
SFC	$\theta_2 \, e^{-\theta_2(u-\theta_1)} \, \mathbb{1}_{[\theta_1, +\infty[}$	$\theta_1 = 17.23, \theta_2 = 3.45$

Table 2: Uncertainty modeling

The probability density function of a beta distribution on $[a, b]$ with shape parameters (α, β) is

$$g_{(\alpha,\beta,a,b)}(x) = \frac{(x-a)^{(\alpha-1)}(b-x)^{\beta-1}}{(b-a)^{\beta-1} B(\alpha, \beta)} \mathbb{1}_{[a,b](x)} ,$$

where $B(\cdot, \cdot)$ is the beta function. Still following [13], we take the nominal and extremal values of V and F as in Table 3.

variable	nominal value	min	max
V	**231**	226	234
F	**19**	18.7	19.05

Table 3: Minimal and maximal values of uncertain variables

The uncertainty on the cruise speed V represents a relative difference of arrival time of 8 minutes.

The airplane manufacturer may wonder whether he has to improve the quality of the engine (SFC) or the aerodynamical property of the plane (F). Thus we study the sensitivity of M_{fuel} with respect to F and SFC and we want to know if $H_0 : \ S^{SFC} > S^F$ or $H_1 : \ S^{SFC} \leq S^F$. Applying the test procedure described previously we can not reject H_0.

4 Concentration inequalities

In this section we give concentration inequalities satisfied by the Sobol indices in one dimension (i.e. $k = 1$).

We define the h function by $h(x) = (1+x)\ln(1+x) - x$ for all $x > -1$.

4.1 Concentration inequalities for $S^{\mathbf{u}}_{N,\mathrm{Cl}}$

We introduce the random variables

$$U_i^{\pm} = Y_i Y_i^{\mathbf{u}} - (S^{\mathbf{u}}_{\mathrm{Cl}} \pm y)(Y_i)^2 \text{ and } J_i^{\pm} = (S^{\mathbf{u}}_{\mathrm{Cl}} \pm y) Y_i - Y_i^{\mathbf{u}}$$

and denote V_U^+ (respectively V_U^-, V_J^+ and V_J^-) the second moment of the i.i.d. random variable U_i^+ (resp. U_i^-, J_i^+ and J_i^-).

Theorem 1. *Let $b > 0$ and $y > 0$. Assume that all the random variables Y_i and $Y_i^{\mathbf{u}}$ belong to $[-b, b]$. Then*

$$\mathbb{P}\left(S^{\mathbf{u}}_{N,\mathrm{Cl}} \geq S^{\mathbf{u}}_{\mathrm{Cl}} + y\right) \leq M_1 + 2M_2 + 2M_3, \tag{10}$$

$$\mathbb{P}\left(S^{\mathbf{u}}_{N,\mathrm{Cl}} \leq S^{\mathbf{u}}_{\mathrm{Cl}} - y\right) \leq M_4 + 2M_2 + 2M_5, \tag{11}$$

where

$$M_1 = \exp\left\{-\frac{NV_U^+}{b_U^2} h\left(\frac{b_U}{V_U^+}\frac{yV}{2}\right)\right\} \qquad M_3 = \exp\left\{-\frac{NV_J^+ b^2}{b_U^2} h\left(\frac{b_U}{bV_J^+}\sqrt{\frac{yV}{2}}\right)\right\}$$

$$M_2 = \exp\left\{-\frac{NV}{b^2} h\left(\frac{b}{V}\sqrt{\frac{yV}{2}}\right)\right\} \qquad M_4 = \exp\left\{-\frac{NV_U^-}{b_U^2} h\left(\frac{b_U}{V_U^-}\frac{yV}{2}\right)\right\}$$

$$M_5 = \exp\left\{-\frac{NV_J^- b^2}{b_U^2} h\left(\frac{b_U}{bV_J^-}\sqrt{\frac{yV}{2}}\right)\right\}$$

and $b_U = b^2(1 + S^{\mathbf{u}}_{\mathrm{Cl}} + y)$.

Remark 4.1. *One must be cautious since the variables $Y_i - \overline{Y}_N$ are dependent.*

Proof. Since $S^{\mathbf{u}}_{\mathrm{Cl}}$ and $S^{\mathbf{u}}_{N,\mathrm{Cl}}$ are invariant by translation on Y and $Y^{\mathbf{u}}$, one may assume without loss of generality that Y is centered.

1. Obviously U_i^+ and U_i^- are upper-bounded by b_U, J_i^+ and J_i^- by b_U/b,

$$\mathbb{E}(U_i^+) = -yV \qquad \mathbb{E}(J_i^+) = 0$$
$$\mathbb{E}(U_i^-) = yV \qquad \mathbb{E}(J_i^-) = 0$$

and

$$V_U^{\pm} = \mathrm{Var}(YY^{\mathbf{u}}) + (S^{\mathbf{u}}_{\mathrm{Cl}} + y)^2 \mathrm{Var}(Y^2) - 2(S^{\mathbf{u}}_{\mathrm{Cl}} \pm y)\mathrm{Cov}(YY^{\mathbf{u}}, Y^2) + y^2V^2$$

$$V_J^{\pm} = ((S^{\mathbf{u}}_{\mathrm{Cl}} \pm y)^2 + 1)V - 2(S^{\mathbf{u}}_{\mathrm{Cl}} \pm y)C_u.$$

2. Proof of (10). Using

$$\{a + b \geq c\} \subset \{a \geq c/2\} \cup \{b \geq c/2\} \quad \text{and} \quad \{ab \geq c\} \subset \{|a| \geq \sqrt{c}\} \cup \{|b| \geq \sqrt{c}\}$$

one gets

$$\begin{aligned}
\mathbb{P}\left(S^{\mathbf{u}}_{N,\mathrm{Cl}} \geq S^{\mathbf{u}}_{\mathrm{Cl}} + y\right) &= \mathbb{P}\left(\frac{\frac{1}{N}\sum_{i=1}^N Y_i Y_i^{\mathbf{u}} - \overline{Y}_N \overline{Y}^{\mathbf{u}}_N}{\frac{1}{N}\sum_{i=1}^N (Y_i)^2 - \left(\overline{Y}_N\right)^2} \geq S^{\mathbf{u}}_{\mathrm{Cl}} + y\right) \\
&= \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^N \left(U_i^+ - \mathbb{E}(U^+)\right) + \overline{Y}_N \overline{J}^+_N \geq yV\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^N \left(U_i^+ - \mathbb{E}(U^+)\right)] \geq N\frac{yV}{2}\right) + \mathbb{P}\left(\sum_{i=1}^N Y_i \geq N\sqrt{\frac{yV}{2}}\right) \\
&\quad + \mathbb{P}\left(\sum_{i=1}^N (-Y_i) \geq N\sqrt{\frac{yV}{2}}\right) + \mathbb{P}\left(\sum_{i=1}^N J_i^+ \geq N\sqrt{\frac{yV}{2}}\right) \\
&\quad + \mathbb{P}\left(\sum_{i=1}^N (-J_i^+) \geq N\sqrt{\frac{yV}{2}}\right).
\end{aligned}$$

Inequality (10) comes directly by applying five times Bennett inequality (see [1] and references therein).

3. Proof of (11). In the same way, one gets

$$\begin{aligned}
\mathbb{P}\left(S^{\mathbf{u}}_{N,\mathrm{Cl}} \leq S^{\mathbf{u}}_{\mathrm{Cl}} - y\right) &= \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^N \left(-U_i^- + \mathbb{E}(U^-)\right) + (-\overline{Y}_N)\overline{J}^-_N \geq yV\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^N \left(-U_i^- + \mathbb{E}(U^-)\right) \geq N\frac{yV}{2}\right) + \mathbb{P}\left(\sum_{i=1}^N Y_i \geq N\sqrt{\frac{yV}{2}}\right) \\
&\quad + \mathbb{P}\left(\sum_{i=1}^N (-Y_i) \geq N\sqrt{\frac{yV}{2}}\right) + \mathbb{P}\left(\sum_{i=1}^N J_i^- \geq N\sqrt{\frac{yV}{2}}\right) \\
&\quad + \mathbb{P}\left(\sum_{i=1}^N (-J_i^-) \geq N\sqrt{\frac{yV}{2}}\right).
\end{aligned}$$

Inequality (11) comes directly by applying five times Bennett inequality. □

4.2 Concentration inequalities for $T^{\mathbf{u}}_{N,\mathrm{Cl}}$

Now remind $Z_i = \frac{Y_i + Y_i^{\mathbf{u}}}{2}$ and introduce the random variables $K_i^{\pm} = Y_i Y_i^{\mathbf{u}} - (S^{\mathbf{u}}_{\mathrm{Cl}} \pm y)\frac{(Y_i)^2 + (Y_i^{\mathbf{u}})^2}{2}$. Denote V_K^+ (resp. V_K^-) the second moment of the i.i.d. random variable K_i^+ (resp. K_i^-).

Theorem 2. *Let $b > 0$ and $y > 0$. Assume that $Y \in [-b, b]$. Then*

$$\mathbb{P}\left(T^{\mathbf{u}}_{N,\mathrm{Cl}} \geq S^{\mathbf{u}}_{\mathrm{Cl}} + y\right) \leq m_1 + 2m_2 \mathbb{1}_{\{S^{\mathbf{u}}_{\mathrm{Cl}} + y - 1 \geq 0\}}, \tag{12}$$

$$\mathbb{P}\left(T^{\mathbf{u}}_{N,\mathrm{Cl}} \leq S^{\mathbf{u}}_{\mathrm{Cl}} - y\right) \leq m_3 + 2m_4 \mathbb{1}_{\{S^{\mathbf{u}}_{\mathrm{Cl}} + y - 1 \geq 0\}}, \tag{13}$$

where

$$m_1 = \exp\left\{-\frac{NV_K^+}{b_U^2} h\left(\frac{b_U}{V_K^+}\frac{yV}{2}\right)\right\}, \quad m_2 = \exp\left\{-\frac{N(V+C)}{2b^2} h\left(\frac{b}{V+C}\sqrt{\frac{2yV}{S^{\mathbf{u}}_{\mathrm{Cl}} + y - 1}}\right)\right\},$$

$$m_3 = \exp\left\{-\frac{NV_K^-}{b_U^2} h\left(\frac{b_U}{V_K^-}\frac{yV}{2}\right)\right\}, \quad m_4 = \exp\left\{-\frac{N(V+C)}{2b^2} h\left(\frac{b}{V+C}\sqrt{\frac{2yV}{y + 1 - S^{\mathbf{u}}_{\mathrm{Cl}}}}\right)\right\}.$$

Proof. Since $T^{\mathbf{u}}_{N,\mathrm{Cl}}$ is invariant by translation on Y and $Y^{\mathbf{u}}$, one may assume without loss of generality that Y is centered.

1. Obvisouly K_i^+ and K_i^- are upper-bounded by b_U, $\mathbb{E}(K_i^+) = -yV$, $\mathbb{E}(K_i^-) = yV$ and

$$V_K^{\pm} = V_U^{\pm} + (S^{\mathbf{u}}_{\mathrm{Cl}} \pm y)^2 \frac{\mathrm{Cov}(Y^2, (Y^{\mathbf{u}})^2) - \mathrm{Var}(Y^2)}{2}.$$

   We also have Z_i is upper-bounded by b, $\mathbb{E}(Z_i) = 0$ and $\mathbb{E}(Z_i^2) = \frac{V+C_u}{2}$.

2. Proof of (12). One gets if $S^{\mathbf{u}}_{\mathrm{Cl}} + y - 1 \geq 0$

$$\begin{aligned}
\mathbb{P}\left(T^{\mathbf{u}}_{N,\mathrm{Cl}} \geq S^{\mathbf{u}}_{\mathrm{Cl}} + y\right) &= \mathbb{P}\left(\frac{\frac{1}{N}\sum_{i=1}^N Y_i Y_i^{\mathbf{u}} - \left(\overline{Z}_N\right)^2}{\frac{1}{N}\sum_{i=1}^N \frac{Y_i^2 + (Y_i^{\mathbf{u}})^2}{2} - \left(\overline{Z}_N\right)^2} \geq S^{\mathbf{u}}_{\mathrm{Cl}} + y\right) \\
&= \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^N \left(K_i^+ - \mathbb{E}(K^+)\right) + \left(\overline{Z}_N\right)^2 (S^{\mathbf{u}}_{\mathrm{Cl}} + y - 1) \geq yV\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^N \left(K_i^+ - \mathbb{E}(K^+)\right) \geq N\frac{yV}{2}\right) + \mathbb{P}\left(\sum_{i=1}^N Z_i \geq N\sqrt{\frac{yV}{2(S^{\mathbf{u}}_{\mathrm{Cl}} + y - 1)}}\right) \\
&\quad + \mathbb{P}\left(\sum_{i=1}^N (-Z_i) \geq N\sqrt{\frac{yV}{2(S^{\mathbf{u}}_{\mathrm{Cl}} + y - 1)}}\right).
\end{aligned}$$

   Inequality (12) comes directly by applying Bennett inequality to the random variables K_i^+, Z_i and $-Z_i$.

3. Proof of (13). One gets since $y + 1 - S^{\mathbf{u}}_{\mathrm{Cl}} > 0$

$$\begin{aligned}
\mathbb{P}\left(T^{\mathbf{u}}_{N,\mathrm{Cl}} \leq S^{\mathbf{u}}_{\mathrm{Cl}} - y\right) &= \mathbb{P}\left(\frac{1}{N}\sum_{i=1}^N \left(-K_i^- + \mathbb{E}(K^-)\right) + \left(\overline{Z}_N\right)^2 (y + 1 - S^{\mathbf{u}}_{\mathrm{Cl}}) \geq yV\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^N \left(-K_i^- + \mathbb{E}(K^-)\right) \geq N\frac{yV}{2}\right) + \mathbb{P}\left(\sum_{i=1}^N Z_i \geq N\sqrt{\frac{yV}{2(y + 1 - S^{\mathbf{u}}_{\mathrm{Cl}})}}\right) \\
&\quad + \mathbb{P}\left(\sum_{i=1}^N (-Z_i) \geq N\sqrt{\frac{yV}{2(y + 1 - S^{\mathbf{u}}_{\mathrm{Cl}})}}\right).
\end{aligned}$$

   Inequality (13) comes from Bennett inequality to the random variables K_i^-, Z_i and $-Z_i$. □

4.3 Numerical applications

In this section, we provide numerical illustrations of the concentration inequalities stated in Sections 4.1 and 4.2.

The upper bounds appearing in Theorem 1 involve the (a priori) unknown quantities:

$$Q = \left(V, V_U^+, V_U^-, V_J^+, V_J^-, S^{\mathbf{u}}_{\mathrm{Cl}}\right).$$

We denote by $pAbove(y, N)$ and $pBelow(y, N)$ the estimators of the right-hand sides of (10) and (11), respectively, obtained by replacing the Q vector by its empirical estimate.

Similarly, we denote by $pAbove'(y, N)$ and $pBelow'(y, N)$ the estimators of the right-hand sides of (12) and (13) when:

$$Q' = \left(V, C, V_K^+, V_K^-, S^{\mathbf{u}}_{\mathrm{Cl}}\right)$$

is replaced by its empirical estimate.

One should note at this point that, on the one hand, the bounds of Theorems 1 and 2 are fully rigorous for any N. From a practical point of view, these bounds are not computable, unless the Q (resp. Q')

vector is known. On the other hand, $pAbove$ and $pBelow$ (resp. $pAbove'$ and $pBelow'$) are computable but are not fully justified for finite N, as they rely on the estimation of Q (resp. Q'). However, as pointed out in [4], these bounds are *conservative*, hence they are less sensitive to a bad estimation than the asymptotic confidence interval given by the CLT

We again take for f the Ishigami function considered in 3.4.1.

In this case, it is easy to check that $Y \in [-b, b]$, where:

$$b = 8 + 0.1 \times \pi^4.$$

When such a majoration of Y is not possible, b can be put into the Q (or Q') vector and estimator of it can be plugged in to obtain $pAbove$ and $pBelow$ (or $pAbove'$ and $pBelow'$).

We also choose $\mathbf{u} = \{1\}$.

Figure 3 show, for different values of N, the plot of $pAbove(y, N)$ and $pBelow(y, N)$ (respectively, $pAbove'(y, N)$ and $pBelow'(y, N)$) as functions of y.

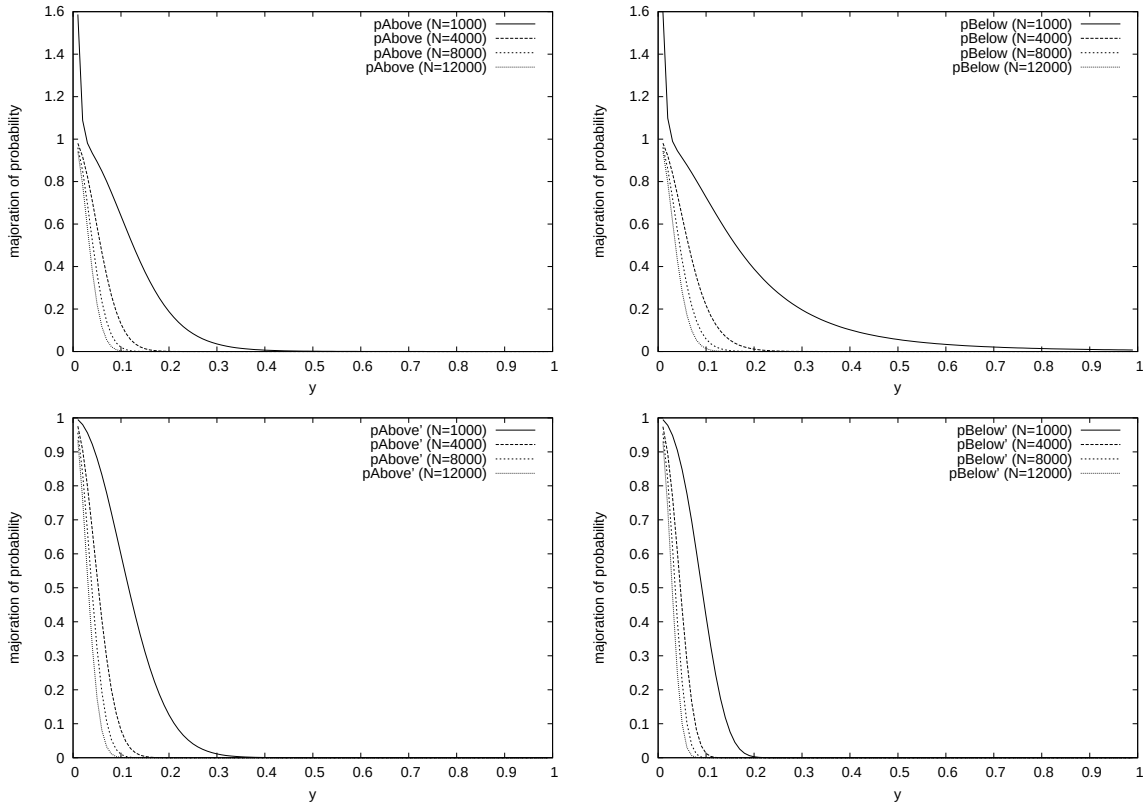

Figure 3: Plots, for $N \in \{1000, 4000, 8000, 12000\}$, of $pAbove(y, N)$ (left-top) and $pBelow(y, N)$ (right-top), $pAbove'(y, N)$ (left-bottom) and $pBelow'(y, N)$ (right-bottom) for the Ishigami model and for $\mathbf{u} = \{1\}$.

As expected, the concentration inequalities are more conservative than the asymptotic confidence interval. These plots confirm that the $T^{\mathbf{u}}_{N,\mathrm{Cl}}$ concentrates faster than $S^{\mathbf{u}}_{N,\mathrm{Cl}}$, and the inequality, while conservative, is sharp enough for this desirable property of $T^{\mathbf{u}}_{N,\mathrm{Cl}}$ to be reported. We also notice that there is a dissimetry in the bounds for above and below deviations, as this is often the case for concentration inequalities. Finally, the expected convergence for $N \to +\infty$ is observed.

5 Berry-Esseen Theorems

In this section we will give a general Berry-Esseen type Theorem for the estimator $S^{\mathbf{u}}_{N,\mathrm{Cl}}$ in one dimension (i.e. $k = 1$). Let Φ be the cumulative distribution function of the standard Gaussian

distribution.

5.1 Pinelis' Theorem

We first recall a general Berry-Esseen type theorem proved in [12]. Let $(V_i)_{i\geq 1}$ a sequence of i.i.d. centered random variables in $\mathbb{R}^d$, for some $d \in \mathbb{N}^*$. Let f some measurable function: $\mathbb{R}^d \to \mathbb{R}$ with $f(0) = 0$ and such that:

$$\exists\, \varepsilon > 0,\, \exists\, M_\varepsilon > 0 \text{ s. t. } |f(x) - L(x)| \leq \frac{M_\varepsilon}{2}||x||^2 \tag{14}$$

where $L := Df(0)$ is the Fréchet derivative of f at point 0.

Remark 5.1. *Remark that condition (14) is satisfied as soon as f is twice continuously differentiable in a neighborhood of 0.*

Theorem 5.2 (Corollary 3.7 in [12])**.** *Take any $p \in (2, 3]$. Assume (14) holds,*

$$\sigma := \sqrt{\mathbb{E}\left(L(V)^2\right)} > 0\,,$$

and $\left(\mathbb{E}\left(||V||^p\right)\right)^{1/p} < \infty$ where $\|\cdot\|$ denotes the euclidean norm on $\mathbb{R}^d$. Then for all $z \in \mathbb{R}$

$$\left|\mathbb{P}\left(\frac{f(\overline{V}_n)}{\sigma/\sqrt{n}} \leq z\right) - \Phi(z)\right| \leq \frac{\kappa}{n^{p/2-1}}\,, \tag{15}$$

where κ above is a generic constant that depends only upon p.

5.2 Theoretical result for the general case

For any random variable Z, denote by Z^c its centered version $Z - \mathbb{E}(Z)$.

Theorem 5.3. *Assume that the random variable Y has finite moments up to order 6. Then, for all $z \in \mathbb{R}$,*

$$\left|\mathbb{P}\left(\frac{\sqrt{N}}{\sigma}\left[S^{\mathbf{u}}_{N,\mathrm{Cl}} - S^{\mathbf{u}}_{\mathrm{Cl}}\right] \leq z\right) - \Phi(z)\right| \leq \frac{\kappa}{\sqrt{N}}\,. \tag{16}$$

Here

$$\sigma^2 := \mathrm{Var}\left(\frac{1}{V}\left(Y^c(Y^{\mathbf{u}})^c - S^{\mathbf{u}}_{\mathrm{Cl}}(Y^c)^2\right)\right)$$

is the asymptotic variance of $\sqrt{N}S^{\mathbf{u}}_{N,\mathrm{Cl}}$.

Proof. We define $V_i = \left(Y_i^c(Y_i^{\mathbf{u}})^c - C_u, Y_i^c, (Y_i^{\mathbf{u}})^c, (Y_i^c)^2 - V\right)^t$ and $f : \mathbb{R}^4 \to \mathbb{R}$ as $f(x, y, z, t) = \frac{x - yz + C_u}{t - y^2 + V} - S^{\mathbf{u}}_{\mathrm{Cl}}$. Note that $f(0, 0, 0, 0) = 0$, $f(\overline{V}_N) = S^{\mathbf{u}}_{N,\mathrm{Cl}} - S^{\mathbf{u}}_{\mathrm{Cl}}$ and by Remark 5.1, (14) holds. The result is then a direct application of Theorem 5.2 once $\sigma^2 = \mathbb{E}\left(L(V)^2\right)$ will be computed. We have

$$\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}, \frac{\partial f}{\partial t}\right)(0, 0, 0, 0) = \left(\frac{1}{V}, 0, 0, \frac{-C_u}{V^2}\right)\,.$$

Using notation in Section 5.1 one gets

$$L(x, y, z, t) = \frac{1}{V}\left(x - S^{\mathbf{u}}_{\mathrm{Cl}}t\right) \quad \text{and} \quad L(V) = \frac{1}{V}\left(Y^c(Y^{\mathbf{u}})^c - S^{\mathbf{u}}_{\mathrm{Cl}}(Y^c)^2\right)\,.$$

Straightforward computations lead to the required result. □

Then we have a Berry-Essen theorem for Sobol index estimator in a general case (whatever the first moment of Y). However, the constant of the bound is hard or even too complex to express explicitely. In the next section we present a Berry-Essen theorem with explicit bounds in the centered case but with an estimator of $S^{\mathbf{u}}_{\mathrm{Cl}}$ slightly different from $S^{\mathbf{u}}_{N,\mathrm{Cl}}$.

5.3 Practical result in the centered case

In this section we give a Berry-Esseen theorem for the estimator

$$\widetilde{S}^{\mathbf{u}}_{N,\mathrm{Cl}} := \frac{\frac{1}{N}\sum Y_i Y_i^{\mathbf{u}}}{\frac{1}{N}\sum Y_i^2}$$

in the centered case and $k = 1$. Further, let $\kappa \approx 0.42$ be the last best constant known in the classical Berry-Esseen theorem ([8]). We then have

Theorem 5.4. *Assume that the random variable Y has finite moment up to order 6. Then, for all $t \in \mathbb{R}$,*

$$\left|\mathbb{P}\left(\frac{\sqrt{N}}{\sigma}(\widetilde{S}^{\mathbf{u}}_{N,\mathrm{Cl}} - S^{\mathbf{u}}_{\mathrm{Cl}}) \le t\right) - \Phi(t)\right| \le \frac{\kappa\mu_{3,N}}{\sqrt{N}} + \left|\Phi(t) - \Phi\left(\frac{t}{\sqrt{1+\frac{t\nu_N}{\sigma\sqrt{N}V^2}}}\right)\right|. \tag{17}$$

Here

$$\sigma^2 := \mathrm{Var}\left(\frac{1}{V}\left(YY^{\mathbf{u}} - S^{\mathbf{u}}_{\mathrm{Cl}}Y^2\right)\right) \tag{18}$$

is the asymptotic variance of $\sqrt{N}\widetilde{S}^{\mathbf{u}}_{N,\mathrm{Cl}}$ and

$$\begin{aligned}
\mu_{3,N} &:= \mathbb{E}\left[\left|\frac{\Delta_n - \mathbb{E}(\Delta_n)}{\sqrt{\mathrm{Var}\Delta_n}}\right|^3\right],\\
\Delta_N &:= \sigma^{-1}V\left[YY^{\mathbf{u}} - \left(S^{\mathbf{u}}_{\mathrm{Cl}} + \frac{t\sigma}{\sqrt{N}}\right)Y^2\right],\\
\nu_N &:= \left(\frac{t\sigma}{\sqrt{N}} + 2S^{\mathbf{u}}_{\mathrm{Cl}}\right)\mathrm{Var}(Y^2) - 2\mathrm{Cov}(YY^{\mathbf{u}}, Y^2).
\end{aligned}$$

Proof. To begin with, we compute the asymptotic variance σ^2 of $\sqrt{N}\widetilde{S}^{\mathbf{u}}_{N,\mathrm{Cl}}$: we apply the so-called Delta method [17] to $W_i = (Y_iY_i^{\mathbf{u}}, Y_i^2)$ and $\Psi(u,v) := uv^{-1}$, $(u \in \mathbb{R}, v > 0)$. Then $\sigma^2 = J_\Psi(\mathbb{E}(W))\Sigma J_\Psi(\mathbb{E}(W))^t$ (J_Ψ the Jacobian of Ψ) and the expression given in (18) follows obviously.
Now, for $t \in \mathbb{R}$, set

$$A_t := \left\{\frac{\sqrt{N}}{\sigma}\left(\widetilde{S}^{\mathbf{u}}_{N,\mathrm{Cl}} - S^{\mathbf{u}}_{\mathrm{Cl}}\right) \le t\right\}.$$

Obvious algrebraic manipulations lead to

$$A_t = \left\{\sqrt{N^{-1}}\sum_{j=1}^{N}\Delta_{N,j} \le 0\right\}$$

where, for $j = 1, \ldots, N$,

$$\Delta_{N,j} := \sigma^{-1}\left[Y_jY_j^{\mathbf{u}}V - \left(C_u + \frac{t\sigma}{\sqrt{N}}V\right)Y_j^2\right].$$

Now, we have

$$\mathbb{E}(\Delta_N) = \frac{-tV^2}{\sqrt{N}} \quad \text{and} \quad \mathrm{Var}(\Delta_N) = V^4\left[1 + \frac{t\nu_N}{\sigma\sqrt{N}V^2}\right].$$

So that denoting by $\Delta_{N,\cdot}$ the empirical mean of $(\Delta_{N,j})_{j=1,\ldots,N}$, we obtain

$$A_t = \left\{\sqrt{N}\Delta_{N,\cdot} \le 0\right\} = \left\{\sqrt{N}\left(\frac{\Delta_{N,\cdot} - \mathbb{E}(\Delta_N)}{\sqrt{\mathrm{Var}\Delta_N}}\right) \le \frac{t}{\sqrt{1+\frac{t\nu_N}{\sigma\sqrt{N}V^2}}}\right\}.$$

Now, to conclude we apply Berry-Esseen theorem (see [8]) and the triangular inequality to obtain (16). □

5.4 Numerical applications for the centered case

We denote by $B(t)$ the right hand side of the Berry-Esseen inequality (17). It is clear that, for any $y > 0$, we have:

$$\mathbb{P}(-y \leq \widetilde{S}^{\mathbf{u}}_{N,\mathrm{Cl}} - S^{\mathbf{u}}_{\mathrm{Cl}} \leq y) \geq \left[\Phi\left(\frac{\sqrt{N}}{\sigma}y\right) - \Phi\left(-\frac{\sqrt{N}}{\sigma}y\right)\right] - \left[B\left(\frac{\sqrt{N}}{\sigma}y\right) + B\left(-\frac{\sqrt{N}}{\sigma}y\right)\right] \quad (19)$$

and:

$$\mathbb{P}(-y \leq \widetilde{S}^{\mathbf{u}}_{N,\mathrm{Cl}} - S^{\mathbf{u}}_{\mathrm{Cl}} \leq y) \leq \left[\Phi\left(\frac{\sqrt{N}}{\sigma}y\right) - \Phi\left(-\frac{\sqrt{N}}{\sigma}y\right)\right] + \left[B\left(\frac{\sqrt{N}}{\sigma}y\right) + B\left(-\frac{\sqrt{N}}{\sigma}y\right)\right] \quad (20)$$

Hence, the actual confidence level of the asymptotic confidence interval for $S^{\mathbf{u}}$ using $\widetilde{S}^{\mathbf{u}}_{N,\mathrm{Cl}}$ is greater than the theoretical level (first term of the sum above), minus a correction term given by the Berry-Esseen theorem (second term). The upper bound given by (20) may also be of practical interest: an overly conservative (overconfident) interval is not always desirable, as a more precise interval with accurate level may exist.

As in the previous applicational section 4.3, the lower bound of the asymptotic confidence interval level involve unkown quantities (moments of Δ_N, Y, YY_u) that have to be estimated. We designate by $L(y, N)$ (resp. $U(y, N)$) the estimator of the right hand side of (19) (resp. (20)) when all unkown quantities are empirically estimated.

We take as output model the Ishigami function defined at Section 3.4.1, recentered by its true mean 7/2 :

$$Y = f(X_1, X_2, X_3) = \sin X_1 + 7\sin^2 X_2 + 0.1X_3^4 \sin X_1 - \frac{7}{2}.$$

Note that the true mean could also be replaced by an estimate of the mean. For y, we choose $y = 1.96\frac{\widehat{\sigma^2}}{\sqrt{N}}$, where $\widehat{\sigma^2}$ is an empirical estimate of σ^2, so as to compute (estimators of) upper and lower bounds of the actual level of the 95%-level confidence interval.

We present the numerical results, as functions of N, and for $\mathbf{u} = \{1\}$ in Figure 4; for $\mathbf{u} = \{2\}$ or $\mathbf{u} = \{3\}$, the results were very similar.

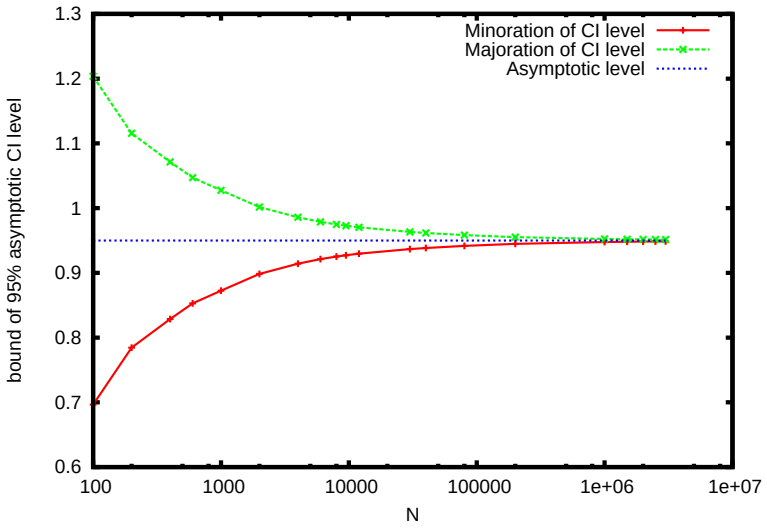

Figure 4: Plots of $L(N)$ (minoration of CI level)) and $U(N)$ (majoration of CI level) for $\mathbf{u} = \{1\}$ and different values of N.

As expected, the actual confidence level is estimated under the "target" level of the confidence interval (0.95). As $N \to +\infty$, our bound converges (quite slowly) to 0.95. Nevertheless, the Berry-Esseen

bound we have presented quickly attains confidence levels which are very close to the asymptotic level, and it can be used so as to provide a certification, at finite sample size, of the level of the asymptotic confidence interval.

Acknowledgements. 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] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

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

[3] 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.

[4] Fred J Hickernell, Lan Jiang, Yuewei Liu, and Art Owen. Guaranteed conservative fixed width confidence intervals via monte carlo sampling. *arXiv preprint arXiv:1208.4318*, 2012.

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

[6] 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.

[7] Alexandre Janon, Thierry Klein, Agnès Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two Sobol index estimators.

[8] V. Yu. Korolev and I. G. Shevtsova. An upper bound for the absolute constant in the Berry-Esseen inequality. *Teor. Veroyatn. Primen.*, 54(4):671–695, 2009.

[9] 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.

[10] Art B Owen. Better estimation of small sobol'sensitivity indices. *arXiv preprint arXiv:1204.4763*, 2012.

[11] Art B Owen. Variance components and generalized sobol' indices. Preprint available at `http://arxiv.org/abs/1205.1774`, 2012.

[12] I. Pinelis and R. Molzon. Berry-esseen bounds for general nonlinear statistics, with applications to pearson's and non-central student's and hotelling's. *Arxiv preprint arXiv:0906.0177v3*, 2012.

[13] Nabil Rachdi, Jean-Claude Fort, and Thierry Klein. Stochastic inverse problem with noisy simulator-application to aeronautical model. *Annales de la Faculté des Sciences de Toulouse*, 6, 21:593–622, 2012.

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

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

[16] 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.

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