Asymptotic normality of a Sobol index estimator in Gaussian process regression framework

Loic Le Gratiet

Asymptotic normality of a Sobol index estimator in Gaussian process regression framework

Loic Le Gratiet [†] [‡]

[†] Université Paris Diderot 75205 Paris Cedex 13

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

May 31, 2013

1 Abstract

Stochastic simulators such as Monte-Carlo estimators are widely used in science and engineering to study physical systems through their probabilistic representation. Global sensitivity analysis aims to identify the input parameters which have the most important impact on the output. A popular tool to perform global sensitivity analysis is the variance-based method which comes from the Hoeffding-Sobol decomposition. Nevertheless, this method requires an important number of simulations and is often unfeasible under reasonable time constraint. Therefore, an approximation of the input/output relation of the code is built with a Gaussian process regression model. This paper provides conditions which ensure the asymptotic normality of a Sobol's index estimator evaluated through this surrogate model. This result allows for building asymptotic confidence intervals for the considered Sobol index estimator. The presented method is successfully applied on an academic example on the heat equation.

Keywords: Sensitivity analysis, Gaussian process regression, asymptotic normality, stochastic simulators, Sobol index.

2 Introduction

Complex computer codes usually have a large number of input parameters. The determination of the important input parameters can be carried out by a global sensitivity analysis. We focus on the variance-based Sobol indices [1], [2], [3] and [4] coming from the Hoeffding-Sobol decomposition [5] which is valid when the input parameters are independent random variables. For an extension of the Hoeffding-Sobol decomposition in a non-independent case, the reader is referred to [6], [7], [8], [9] and [10].

Monte-Carlo methods are commonly used to estimate the Sobol indices (see [1], [11] and [12]). One of their main advantages is that they allow for quantifying the uncertainty related to the estimation errors. In particular, for non-asymptotic cases, this can be easily carried out with a bootstrap procedure as presented in [13] and [14]. Furthermore, in asymptotic cases, useful properties can be shown as the asymptotic normality [12]. The reader is referred to

[15] for an extensive presentation of asymptotic statistics. Nevertheless, Monte-Carlo methods require a large number of simulations and are often unachievable under reasonable time constraints. Therefore, in order to avoid prohibitive computational costs, we surrogate the simulator with a meta-model and we perform the estimations on it.

In this paper, we consider a special surrogate model corresponding to a Gaussian process regression. More precisely we consider an idealized regression problem for which we can deduce a posterior predictive mean and variance tractable for our purpose. In particular, we can derive the rate of convergence of the meta-model approximation error with respect to the computational budget.

Therefore, the Sobol index estimations - which are performed with a Monte-Carlo procedure by replacing the true code with the posterior predictive mean - have two sources of uncertainty: the one due to the Monte-Carlo scheme and the one due to the meta-model approximation. The error due to the Monte-Carlo procedure tends to zero when the number of particles (calls of the meta-model) tends to infinity and the error due to the meta-model tends to zero when the computational budget (calls of the complex simulator used to build the meta-model) tends to infinity. A question of interest is whether the asymptotic normality presented in [14] is maintained. The principal difficulty of the study is that the estimator lies in a product probability space which takes into account both the uncertainty of the Gaussian process and the one of the Monte-Carlo sample.

We emphasize that [14] presents such a result for noise-free Gaussian process regression using a squared exponential covariance kernel (see [16]). They give conditions on the number of simulations and the number of Monte-Carlo particles which ensure the asymptotic normality for the Sobol index estimators. A part of our developments is inspired by their work nevertheless they are different with some important respects. Indeed, the particular case of noise-free Gaussian process regression with squared exponential covariance kernel allows for not considering the probability space in which lies the Gaussian process. This significantly simplifies the mathematical developments. Unfortunately this simplification does not hold in our general framework.

In this paper, we are interested in stochastic simulators which use Monte-Carlo or Monte-Carlo Markov Chain methods to solve a system of differential equations through its probabilistic interpretation. Such simulators provide noisy observations with a noise level inversely proportional to the number of Monte-Carlo particles used by the simulator. Therefore, with a fixed computational budget, we have to make a trade-off between the number of simulations and the output accuracy. Actually, we consider the asymptotic case where the number of observations is large.

The main result of this paper is a theorem giving sufficient conditions to ensure the asymptotic normality of the Sobol index estimators based on the Monte-Carlo procedure presented in [1] and using the presented Gaussian process regression model. We note that the presented theorem holds for a large class of covariance kernels. The asymptotic normality is of interest since it allows for giving asymptotic confidence intervals on the Sobol index estimators. This result is illustrated with an academic example dealing with the heat equation problem.

3 Gaussian process regression for stochastic simulators

We present in Subsection 3.1 the practical problem that we want to deal with. In order to handle the asymptotic framework of a large number of observations, we replace the true

problem by an idealized version of it in Subsection 3.2. This idealization allows us to study the asymptotic normality of the Sobol's index estimator in Section 4.

3.1 Gaussian process regression with a large number of observations

Let us suppose that we want to surrogate a function $f(x)$, $x \in Q \subset \mathbb{R}^d$, from noisy observations of it at points $(x_i)_{i=1,\dots,n}$ sampled from the probability measure μ - μ is called the design measure and Q is an nonempty open set. Furthermore, we consider that we have r replications at each point. We hence have nr experiments of the form $z_{i,j} = f(x_i) + \varepsilon_{i,j}$, $i = 1, \dots, n$, $j = 1, \dots, r$ and we consider that $(\varepsilon_{i,j})_{\substack{i=1,\dots,n \\ j=1,\dots,r}}$ are independently sampled from a Gaussian distribution with mean zero and variance σ_ε^2. A stochastic simulator provides outputs of the following form

$$z_i = \frac{1}{r}\sum_{j=1}^{r} z_{i,j} = f(x_i) + \varepsilon_i, \quad \forall i = 1, \dots, n$$

where $(\varepsilon_i)_{i=1,\dots,n}$ are the observation noises sampled from a zero-mean Gaussian distribution with variance σ_ε^2/r. Therefore, if we consider a fixed number of experiments $T = nr$, we have an observation noise variance equal to $n\sigma_\varepsilon^2/T$.

Note that an observation noise variance proportional to n is natural in the framework of stochastic simulators. Indeed, for a fixed total number of experiments $T = nr$, we can either decide to perform them in few points (i.e. n small) but with lot of replications (i.e. r large) or decide to perform them in lot of points (i.e. n large) but with few replications (i.e. r small).

In a Gaussian process regression framework, we model $f(x)$ as a Gaussian process with a known mean (that we take equal to zero without loss of generality) and a covariance kernel $k(x, \tilde{x})$. Therefore, in the remainder of this paper, the function $f(x)$ is random. The predictive Mean Squared Error (MSE) of the Best Linear Unbiased Predictor (BLUP) given by

$$\hat{z}_{T,n}(x) = \mathbf{k}'(x)\left(\mathbf{K} + \frac{n\sigma_\varepsilon^2}{T}\mathbf{I}\right)^{-1}\mathbf{z}^n \tag{1}$$

is

$$\sigma_{T,n}^2(x) = k(x,x) - \mathbf{k}'(x)\left(\mathbf{K} + \frac{n\sigma_\varepsilon^2}{T}\mathbf{I}\right)^{-1}\mathbf{k}(x) \tag{2}$$

where $\mathbf{z}^n = (z_i)_{i=1,\dots,n}$ denotes the vector of the observed values, $k(x) = [k(x, x_i)]_{1\le i\le n}$ is the n-vector containing the covariances between $f(x)$ and $f(x_i)$, $1 \le i \le n$, $\mathbf{K} = [k(x_i, x_j)]_{1\le i,j\le n}$ is the $n\times n$-matrix containing the covariances between $f(x_i)$ and $f(x_j)$, $1 \le i, j \le n$ and $\mathbf{I}$ is the $n \times n$ identity matrix.

In this paper, we consider the case $n \gg 1$. It corresponds to a massive experimental design set but with observations with a large noise variance. This case is realistic for stochastic simulators where the computational cost resulting from one Monte-Carlo particle is very low and thus can be run in lot of points $(x_i)_{i=1,\dots,n}$.

3.2 Idealized Gaussian process regression

We assume from now on that the positive kernel $k(x, \tilde{x})$ is continuous and that $\sup_{x\in Q} k(x, x) < \infty$ where Q is a nonempty open subset of $\mathbb{R}^d$. We introduce the Mercer's decomposition of

$k(x,\tilde{x})$ [17], [18]:

$$k(x,\tilde{x}) = \sum_{p\geq 0} \lambda_p \phi_p(x)\phi_p(\tilde{x}) \tag{3}$$

where $(\phi_p(x))_p$ is an orthonormal basis of $L^2_\mu(\mathbb{R}^d)$ consisting of eigenfunctions of the integral operator $(T_{\mu,k}g)(x) = \int_{\mathbb{R}^d} k(x,u)g(u)d\mu(u)$ and λ_p is the nonnegative sequence of corresponding eigenvalues sorted in decreasing order.

Let us consider the following predictor:

$$\hat{z}_T(x) = \sum_{p\geq 0} \frac{\lambda_p}{\lambda_p + \sigma_\varepsilon^2/T} z_p \phi_p(x) \tag{4}$$

where $z_p = f_p + \varepsilon_p^*$, $f_p = \int f(x)\phi_p(x)\, d\mu(x)$, $\varepsilon_p^* \sim \mathcal{N}\left(0, \sigma_\varepsilon^2/T\right)$, ε_p^* independent of ε_q^* for $p \neq q$ and $(\varepsilon_p^*)_{p\geq 0}$ independent of $(f_p)_{p\geq 0}$. Note that we have $f_p \sim \mathcal{N}(0,\lambda_p)$, f_p independent of f_q for $p \neq q$ and $f(x) = \sum_{p\geq 0} f_p \phi_p(x)$.

Let us introduce the probability space $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z) = (\Omega_f \times \Omega_\varepsilon, \sigma(\mathcal{F}_f \times \mathcal{F}_\varepsilon), \mathbb{P}_f \times \mathbb{P}_\varepsilon)$ where $(\Omega_f, \mathcal{F}_f, \mathbb{P}_f)$ corresponds to the probability space where $f(x)$ and the sequence $(f_p)_{p\geq 0}$ are defined and $(\Omega_\varepsilon, \mathcal{F}_\varepsilon, \mathbb{P}_\varepsilon)$ is the probability space where the observation noises $(\varepsilon_i)_{i\in\mathbb{N}}$ and the sequence $(\varepsilon_p^*)_{p\geq 0}$ are defined. Further, let us consider the sequence of independent random variables $(X_i)_{i\in\mathbb{N}}$ with probability measure μ on $Q \subset \mathbb{R}^d$ and defined on the probability space $(\Omega_D, \mathcal{F}_D, \mathbb{P}_D)$. The sequence $(X_i)_{i=1,\ldots,n}$ represents the experimental design set considered as a random variable. Therefore, the predictors $\hat{z}_{T,n}(x)$ in (1) and $\hat{z}_T(x)$ in (4) are associated to the random experimental design set $(X_i)_{i\in\mathbb{N}}$. We have the following convergence in probability when $n \to \infty$ [19]:

$$\sigma^2_{T,n}(x) \xrightarrow[n\to\infty]{\mathbb{P}_D} \sigma^2_T(x) \tag{5}$$

where $\sigma^2_{T,n}(x) = \mathbb{E}_Z\left[(\hat{z}_{T,n}(x) - f(x))^2\right]$ (2) and $\sigma^2_T(x) = \mathbb{E}_Z\left[(\hat{z}_T(x) - f(x))^2\right]$. Therefore $\hat{z}_T(x)$ in (4) is a relevant candidate for an idealized version of $\hat{z}_{T,n}(x)$ in (1) for the considered asymptotics $n \to \infty$. The following proposition allows for completing the justification of the relevance of $\hat{z}_{T,n}(x)$.

Proposition 1. *Let us consider $f(x)$ a Gaussian process of zero mean and covariance kernel $k(x,\tilde{x})$, $\hat{z}_{T,n}(x)$ in (1) and $\hat{z}_T(x)$ in (4) both associated to the random experimental design set $(X_i)_{i\in\mathbb{N}}$. Consequently $f(x) = \sum_{p\geq 0} f_p\phi_p(x)$ where $f_p \sim \mathcal{N}(0,\lambda_p)$, $(f_p)_{p\geq 0}$ independent and $(\phi_p(x))_{p\geq 0}$ defined in (3). The following convergence holds $\forall \delta > 0$ and for any Borel set $A \subset \mathbb{R}^2$ such that the Lebesgue measure of its boundary is zero:*

$$\mathbb{P}_D\left(\left|\mathbb{P}_Z\left((\hat{z}_{T,n}(x), f(x)) \in A\right) - \mathbb{P}_Z\left((\hat{z}_T(x), f(x)) \in A\right)\right| > \delta\right) \overset{n\to\infty}{\longrightarrow} 0 \tag{6}$$

Proof of Proposition 1. First of all, we note that for a fixed $\omega_D \in \Omega_D$ the random variables $(\hat{z}_{T,n}(x), f(x))$ and $(\hat{z}_T(x), f(x))$ are Gaussian since they are linear transformations of $((\varepsilon_i)_{i\in\mathbb{N}}, (f_p)_{p\geq 0})$ and $((\varepsilon_p^*)_{p\geq 0}, (f_p)_{p\geq 0})$ which are both independently distributed from Gaussian distributions.

Thanks to the equality $\mathbb{E}_Z\left[(\hat{z}_{T,n}(x))^2\right] = k(x,x) - \sigma^2_{T,n}(x)$ with $k(x,x) = \sum_{p\geq 0}\lambda_p\phi_p(x)^2$, to the definition of $\hat{z}_T(x)$ in (4) and to the convergence (5), the following convergence holds in probability when $n \to \infty$:

$$\mathbb{E}_Z\left[(\hat{z}_{T,n}(x))^2\right] \xrightarrow[n\to\infty]{\mathbb{P}_D} \mathbb{E}_Z\left[(\hat{z}_T(x))^2\right] \tag{7}$$

Furthermore, we also have the equality $\mathbb{E}_Z\left[\hat{z}_{T,n}(x)f(x)\right] = k(x,x) - \sigma^2_{T,n}(x)$ that leads the convergence for $n \to \infty$:

$$\mathbb{E}_Z\left[\hat{z}_{T,n}(x)f(x)\right] \xrightarrow{\mathbb{P}_D} \mathbb{E}_Z\left[\hat{z}_T(x)f(x)\right] \tag{8}$$

We can deduce the following convergence of the covariance of the two-dimensional Gaussian vector $(\hat{z}_{T,n}(x), f(x))$ to the one of the two-dimensional Gaussian vector $(\hat{z}_T(x), f(x))$ when $n \to \infty$:

$$\mathrm{cov}_Z\left((\hat{z}_{T,n}(x), f(x))\right) \xrightarrow{\mathbb{P}_D} \mathrm{cov}_Z\left((\hat{z}_T(x), f(x))\right) \tag{9}$$

Furthermore, the following equality holds:

$$\mathbb{E}_Z\left[(\hat{z}_{T,n}(x), f(x))\right] = \mathbb{E}_Z\left[(\hat{z}_T(x), f(x))\right] = (0,0) \tag{10}$$

Let us denote by $C_n = \mathrm{cov}_Z\left((\hat{z}_{T,n}(x), f(x))\right)$, for all Borel sets $A \subset \mathbb{R}^2$ such that $\nu(\partial A) = 0$ (ν denotes the Lebesgue measure and ∂A the boundary of A), we have the following equality almost surely with respect to $(\Omega_D, \mathcal{F}_D, \mathbb{P}_D)$:

$$\mathbb{P}_Z\left((\hat{z}_{T,n}(x), f(x)) \in A\right) = \phi_2\left(C_n^{-1/2}A\right)$$

where ϕ_2 stands for the bivariate normal distribution $\mathcal{N}(0, \mathbf{I}_2)$. We note that C_n is a random variable defined on the probability space $(\Omega_D, \mathcal{F}_D, \mathbb{P}_D)$. Let us denote by $C = \mathrm{cov}_Z\left((\hat{z}_T(x), f(x))\right)$. The matrix C being nonsingular, the convergence (9) implies the following one when $n \to \infty$:

$$C_n^{-1/2} \xrightarrow[n\to\infty]{\mathbb{P}_D} C^{-1/2}$$

Therefore, for all Borel sets $A \subset \mathbb{R}^2$ such that $\nu(\partial A) = 0$, we have when $n \to \infty$:

$$\phi_2(C_n^{-1/2}A) \xrightarrow[n\to\infty]{\mathbb{P}_D} \phi_2(C^{-1/2}A)$$

Finally, we can deduce that $\forall \delta > 0$ and for all Borel sets $A \subset \mathbb{R}^2$ such that $\nu(\partial(A) = 0$, the convergence in (6) holds. □

The function $\hat{z}_T(x)$ is the surrogate model that we consider in this paper. We note that $\hat{z}_T(x)$ is not equal to the objective function $f(x)$ since $\sigma^2_\varepsilon/T \neq 0$. In practical applications, we expect that the idealized model (4) is close enough to the actual surrogate model (1) so that it provides relevant confidence intervals.

Note that with this formalism $f(x)$ is a random process defined on the probability space $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$. The random series $(z_p)_{p\geq 0}$ is defined on $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$ as well. In order to study the convergence of $\hat{z}_T(x)$ to the real function $f(x)$, let us consider the following equality:

$$\sigma^2_T(x) = \sum_{p\geq 0} \frac{\sigma^2_\varepsilon \lambda_p / T}{\sigma^2_\varepsilon/T + \lambda_p} \phi_p(x)^2 \tag{11}$$

Then, let us define the Integrated Mean Squared Error (IMSE):

$$\mathrm{IMSE}_T = \int_{\mathbb{R}^d} \sigma^2_T(x)\, d\mu(x) = \mathbb{E}_Z\left[||\hat{z}_T(x) - f(x)||^2_{L^2_\mu}\right] \tag{12}$$

The following equality holds:

$$\mathrm{IMSE}_T = \sum_{p\geq 0} \frac{\sigma_\varepsilon^2 \lambda_p / T}{\sigma_\varepsilon^2/T + \lambda_p} \tag{13}$$

We can link the asymptotic rate of convergence of the IMSE (13) with the asymptotic decay of the eigenvalues $(\lambda_p)_{p\geq 0}$ thanks to the following inequalities [19]:

$$B_T^2/2 \leq \mathrm{IMSE}_T \leq B_T^2 \tag{14}$$

with:

$$B_T^2 = \sum_{p \,\mathrm{s.t.}\, \lambda_p \leq \sigma_\varepsilon^2/T} \lambda_p + \frac{\sigma_\varepsilon^2}{T} \#\{p \,\mathrm{s.t.}\, \lambda_p > \sigma_\varepsilon^2/T\} \tag{15}$$

4 Asymptotic normality of a Sobol index estimator

We present in this section the main theorem of this paper about the asymptotic normality of a Sobol index estimator using Monte-Carlo integrations and the meta-model $\hat{z}_T(x)$ presented in Subsection 3.2. In the forthcoming development, we suppose that T is an increasing sequence indexed by the number of Monte-Carlo m particles used to estimate the variance and covariance terms involved in the Sobol index. We use the notation T_m to emphasize that T depends on m. First of all, let us define in Subsection 4.1 the Sobol indices and the considered Monte-Carlo estimator.

4.1 The Sobol indices

Let us suppose that the input parameter is a random vector X with probability measure $\mu = \mu_1 \otimes \mu_2$ on $(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}, \mathcal{B}(\mathbb{R}^{d_1} \times \mathbb{R}^{d_2}))$ with $d = d_1 + d_2$. We consider the random vector $(X, \tilde{X})$ defined on the probability space $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ with $X = (X^1, X^2)$ and $\tilde{X} = (X^1, \tilde{X}^2)$ where X^1 is a random vector with values in $\mathbb{R}^{d_1}$ and with distribution μ_1, X^2 and $\tilde{X}^2$ are random vectors with values in $\mathbb{R}^{d_2}$ with distribution μ_2, and X^1, X^2 and $\tilde{X}^2$ are independent.

We are interested in the following closed Sobol index of parameter X^1 (see [1], [2]):

$$S^{X^1} = \frac{V^{X^1}}{V} = \frac{\mathrm{var}_X\left(\mathbb{E}_X\left[f(X)|X^1\right]\right)}{\mathrm{var}_X\left(f(X)\right)} = \frac{\mathrm{cov}_X\left(f(X), f(\tilde{X})\right)}{\mathrm{var}_X\left(f(X)\right)} \tag{16}$$

where the random variables $f(X)$ and $f(\tilde{X})$ are defined on the product probability space $(\Omega_Z \times \Omega_X, \sigma\left(\mathcal{F}_Z \times \mathcal{F}_X\right), \mathbb{P}_Z \times \mathbb{P}_X)$ and S^{X^1}, V^{X^1} and V are defined on the probability space $(\Omega_Z, \mathcal{F}_Z, \mathcal{F}_Z)$. The Sobol index S^{X^1} can be simply interpreted as a measure of the part of variance of $f(x)$ explained by the factor X^1. We note that $\mathrm{var}_X\left(\right)$, $\mathbb{E}_X\left[\right]$, $\mathrm{cov}_X\left(,\right)$ stand for the variance, the expectation and the covariance in the probability space $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$.

Furthermore, let us consider the sequence $(X_i, \tilde{X}_i)_{i=1}^{\infty}$ of random variables defined on $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ independent and identically distributed such that $(X_i, \tilde{X}_i) \overset{\mathcal{L}}{=} (X, \tilde{X})$ for all $i \in \mathbb{N}^*$ ($\overset{\mathcal{L}}{=}$ stands for the equality in distribution). We use the following classical Monte-Carlo estimator for (16) (see [1]):

$$S_m^{X^1} = \frac{V_m^{X^1}}{V_m} = \frac{m^{-1}\sum_{i=1}^m f(X_i) f(\tilde{X}_i) - m^{-2}\sum_{i,j=1}^m f(X_i) f(\tilde{X}_j)}{m^{-1}\sum_{i=1}^m f^2(X_i) - m^{-2}(\sum_{i=1}^m f(X_i))^2} \tag{17}$$

where the random variables $S_m^{X^1}$, $V_m^{X^1}$ and V_m are defined on the probability space $(\Omega_Z \times \Omega_X, \sigma(\mathcal{F}_Z \times \mathcal{F}_X), \mathbb{P}_Z \times \mathbb{P}_X)$.

Furthermore, after substituting $f(x)$ with the meta-model $\hat{z}_{T_m}(x)$, we obtain the following estimator:

$$S_{T_m,m}^{X^1} = \frac{V_{T_m,m}^{X^1}}{V_{T_m,m}} = \frac{m^{-1}\sum_{i=1}^m \hat{z}_{T_m}(X_i)\hat{z}_{T_m}(\tilde{X}_i) - m^{-2}\sum_{i,j=1}^m \hat{z}_{T_m}(X_i)\hat{z}_{T_m}(\tilde{X}_j)}{m^{-1}\sum_{i=1}^m \hat{z}_{T_m}^2(X_i) - m^{-2}(\sum_{i=1}^m \hat{z}_{T_m}(X_i))^2} \tag{18}$$

where the random variables $S_{T_m,m}^{X^1}$, $V_{T_m,m}^{X^1}$, $V_{T_m,m}$, $\hat{z}_{T_m}(X_i)$ and $\hat{z}_{T_m}(\tilde{X}_j)$ are defined on the product probability space $(\Omega_Z \times \Omega_X, \sigma(\mathcal{F}_Z \times \mathcal{F}_X), \mathbb{P}_Z \times \mathbb{P}_X)$.

4.2 Theorem on the asymptotic normality of the Sobol index estimator

The theorem below gives the relation between T_m and m which ensures the asymptotic normality of the estimator $S_{T_m,m}^{X^1}$ when $m \to \infty$. We note that $S_{T_m,m}^{X^1}$ is the estimator of the Sobol index $S^{X^1} = \mathrm{cov}_X\left(f(X), f(\tilde{X})\right)/\mathrm{var}_X(f(X))$ when we replace the true function by the surrogate model (4) and when we use the Monte-Carlo estimator (17) for the variance and covariance involved in the Sobol index.

Theorem 1 (Asymptotic normality of $S_{T_m,m}^{X^1}$)**.** *Let us consider the estimator $S_{T_m,m}^{X^1}$ (18) of S^{X^1} (16) with T_m an increasing function of $m \in \mathbb{N}^*$. We have the following convergences:*

If $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} 0$, then for all interval $I \in \mathbb{R}$ and $\forall \delta > 0$, we have the convergence:

$$\mathbb{P}_Z\left(\left|\mathbb{P}_X\left(\sqrt{m}\left(S_{T_m,m}^{X^1} - S^{X^1}\right) \in I\right) - \int_I g(x)dx\right| > \delta\right) \overset{m\to\infty}{\longrightarrow} 0 \tag{19}$$

where $g(x)$ is the probability density function of a zero-mean Gaussian random variable with variance:

$$\frac{\mathrm{var}_X\left((f(X) - \mathbb{E}_X[f(X)])\left(f(\tilde{X}) - \mathbb{E}_X[f(X)] - S^{X^1}f(X) + S^{X^1}\mathbb{E}_X[f(X)]\right)\right)}{\left(\mathrm{var}_X(f(X))\right)^2} \tag{20}$$

with $B_{T_m}^2$ given by (15).

If $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} \infty$, then $\forall \delta > 0$, $\exists C > 0$ such that :

$$\mathbb{P}_Z\left(\left|\mathbb{P}_X\left(B_{T_m}^{-1}\left(S_{T_m,m}^{X^1} - S^{X^1}\right) \geq C\right) - 1\right| > \delta\right) \overset{m\to\infty}{\longrightarrow} 0 \tag{21}$$

Theorem 1 is of interest since it gives how fast T_m has to increase with respect to m so that the error of the surrogate modelling and the one of the Monte-Carlo sampling have the same order of magnitude. Indeed, for a given size m of the Monte-Carlo sample, it is not necessary to use a too large T_m otherwise the Monte-Carlo estimation error will dominate (it corresponds to the case $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} 0$). On the other hand, if T_m is taken too large (it corresponds to the case $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} \infty$), the estimation error is dominated by the meta-model approximation.

Furthermore, we see that when $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} 0$, the asymptotic normality is assessed for the estimator $S_{T_m,m}^{X^1}$ with an explicit variance given in equation (20). By studying in (20) the

cases $S^{X^1} = 0$ and $S^{X^1} = 1$ we see that the given estimator is more precise for large values of Sobol indices than for small ones. A more efficient estimator for small index values is given in [11].

We show in Section 5 that the product $mB^2_{T_m}$ can easily be handled when we have an explicit formula for the asymptotic decay of the eigenvalues of the Mercer's decomposition of $k(x, \tilde{x})$. The proof of Theorem 1 is given in Appendix A. It is based on the Skorokhod's representation theorem [20], the Lindeberg-Feller central limit theorem, and the Delta method [15].

5 Examples of asymptotic normality for Sobol's index

According to the previous developments, the desired asymptotic normality is assessed under the assumption $mB^2_{T_m} \overset{m\to\infty}{\longrightarrow} 0$. In the remainder of this section, we present relations between T_m and m which lead the convergence $mB^2_{T_m} \overset{m\to\infty}{\longrightarrow} 0$ for some usual kernels.

5.1 Asymptotic normality with d-tensorised Matérn-ν kernels

We focus here on the d-tensorised Matérn-ν kernel with regularity parameter $\nu > 1/2$ [21], [16]:

$$k(x, \tilde{x}) = \prod_{i=1}^{d} \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|x^i - \tilde{x}^i|}{\theta_i} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|x^i - \tilde{x}^i|}{\theta_i} \right)$$

where K_ν is the modified Bessel function [22]. The eigenvalues of this kernel satisfy the following asymptotic behavior [23]:

$$\lambda_p = \phi(p), \quad p \gg 1$$

where $\phi(p) = \left(\log(1+p)^{2(d-1)(\nu+1/2)}\right) p^{-2(\nu+1/2)} \left(1 + \mathrm{O}(1/p)\right)$. Therefore, for $T_m \gg 1$:

$$B^2_{T_m} \approx \log(T_m/\sigma^2_\varepsilon)^{d-1} \left(\frac{\sigma^2_\varepsilon}{T_m} \right)^{1-1/2(\nu+1/2)}$$

Section 4 suggests that the asymptotic normality of the Sobol's index estimator is assessed when:

$$mB^2_{T_m} \overset{m}{\longrightarrow} 0$$

Let us consider that T_m is such that:

$$\log(T_m/\sigma^2_\varepsilon)^{d-1} \left(\frac{\sigma^2_\varepsilon}{T_m} \right)^{1-1/2(\nu+1/2)} = 1/m \tag{22}$$

It corresponds to the critical point $mB^2_{T_m} \approx 1$. In this case, the error originates both from the meta-model approximation error and the Monte-Carlo estimation error. Equation (22) leads to the following critical budget:

$$\frac{T_m}{\sigma^2_\varepsilon} = \sigma^2_\varepsilon m^{1/(1-1/2(\nu+1/2))} \log(m)^{(d-1)}, \tag{23}$$

and, the asymptotic normality is assessed for:

$$\frac{T_m}{\sigma_\varepsilon^2} = \sigma_\varepsilon^2 m^{1/(1-1/2(\nu+1/2))+\alpha} \log(m)^{(d-1)}, \forall \alpha > 0 \tag{24}$$

In practice, we want to minimize the budget allocated to the simulator and thus consider the case α tends to zero. As a consequence, for applications we will consider the allocation of the critical point (23).

5.2 Asymptotic normality for d-dimensional Gaussian kernels

Let us consider the d-dimensional Gaussian kernel:

$$k(x, \tilde{x}) = \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{(x^i - \tilde{x}^i)^2}{\theta_i^2}\right) \tag{25}$$

Thanks to [24], we have the following upper bound for the eigenvalues:

$$\lambda_p \leq c' \exp\left(-cp^{1/d}\right) \tag{26}$$

with c and c' constants. From this inequality, we can deduce that $\exists C > 0$ such that:

$$B_{T_m}^2 \approx C \log(T_m/\sigma_\varepsilon^2)^d \left(\frac{\sigma_\varepsilon^2}{T_m}\right)$$

Therefore, the critical budget corresponding to the critical point $mB_{T_m}^2 \approx 1$ is given by

$$T_m/\sigma_\varepsilon^2 = m \log(m)^d \tag{27}$$

and the asymptotic normality for the Sobol index estimator is assessed with:

$$T_m/\sigma_\varepsilon^2 = m^{1+\alpha} \log(m)^d, \forall \alpha > 0 \tag{28}$$

We note that the condition is only sufficient since we have an inequality in (26).

5.3 Asymptotic normality for d-dimensional Gaussian kernels with a Gaussian measure $\mu(x)$

Let us consider a Gaussian measure $\mu \sim \mathcal{N}(0, \sigma_\mu^2 \mathbf{I})$ in dimension d and the Gaussian kernel (25). As presented in [25], we have analytical expressions for the eigenvalues and eigenfunctions of $k(x, \tilde{x})$:

$$\lambda_p = \prod_{i=1}^{d} \sqrt{\frac{2a}{A_i}} B_i^p$$

$$\phi_p(x) = \exp\left(-\sum_{i=1}^{d}(c_i - a)(x^i)^2\right) \prod_{i=1}^{d} H_p(\sqrt{2c_i}x^i)$$

where $H_p(x) = (-1)^p \exp(x^2)\frac{d^p}{dx^p}\exp(-x^2)$ is the p^{th} order Hermite polynomial (see [26]), $a = 1/(2\sigma_\mu)^2$, $b_i = 1/(2\theta_i^2)$ and

$$c_i = \sqrt{a^2 + 2ab_i}, \quad A_i = a + b_i + c_i, \quad B_i = b_i/A_i.$$

Therefore, the eigenvalues satisfy the following asymptotic behavior

$$\lambda_p \propto \exp\left(-p\xi_d\right) \tag{29}$$

where $\xi_d = \sum_{i=1}^{d} \log\left(1/B_i\right)$. For $T_m \gg 1$, we have:

$$B_{T_m}^2 \approx \left(\sigma_\varepsilon^2/T_m\right) \log\left(T_m/\sigma_\varepsilon^2\right)/\xi_d \tag{30}$$

Let us consider the critical point $B_{T_m}^2 = 1/m$. Then, the critical budget is given by

$$\frac{T_m}{\sigma_\varepsilon^2} = \xi_d m \log(m)$$

and the asymptotic normality is assessed for:

$$\frac{T_m}{\sigma_\varepsilon^2} = \xi_d m^{1+\alpha} \log(m),\ \forall \alpha > 0 \tag{31}$$

6 Numerical illustration

The purpose of this section is to perform a global sensitivity analysis of a stochastic code solving the following heat equation:

$$\frac{\partial u}{\partial t}(x,t) - \frac{1}{2}\Delta u(x,t) = 0 \tag{32}$$

with $x \in \mathbb{R}^d$ and $u(x,0) = g(x) = \exp(-\sum_{i=1}^{d} x_i^2/(2\sigma_{g,i}^2))$. The function $u(x,t)$ has the following probabilistic representation:

$$u(x,t) = \mathbb{E}_{W_t}[g(x+W_t)] \tag{33}$$

where W_t is the 1-dimensional Brownian motion. We evaluate the function $u(x,t)$ through the following stochastic code:

$$u_r^{\text{code}}(x,t) = \frac{1}{r}\sum_{i=1}^{r}\left(\frac{1}{s}\sum_{j=1}^{s} g(x+W_{t,i,j})\right) \tag{34}$$

where the number of replications r tunes the precision of the output, $s = 30$ and $(W_{t,i,j})_{\substack{i=1,\ldots,r \\ j=1,\ldots,s}}$ are sampled from a Gaussian random variable of mean zero and variance t.

We note that there is a closed form expression for the solution of the considered heat equation, that will allow is to compute exactly the Sobol indices and to assess the quality of our estimate:

$$u(x,t) = \prod_{i=1}^{d}\left(\frac{\sigma_{g,i}^2}{\sigma_{g,i}^2+t}\right)^{1/2} \exp\left(-\frac{x_i^2}{2(\sigma_{g,i}^2+t)}\right) \tag{35}$$

6.1 Exact Sobol indices

Let us consider that x is a random variable X defined on $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ such that $X \sim \mathcal{N}\left(0, \sigma_\mu^2 \mathbf{I}\right)$. We are interested for the application in the first order Sobol indices, i.e. the contribution of $(X^j)_{j=1,\dots,d}$. By straightforward calculations it can be shown that:

$$S^{X^j} = \frac{V^{X^j}}{V} = \frac{\mathrm{var}_X(\mathbb{E}_X[u(X,t)|X^j])}{\mathrm{var}_X(u(X,t))} = \frac{B_j - 1}{\left(\prod_{i=1}^d B_i\right) - 1} \tag{36}$$

where X^j is the j^{th} component of the random vector X with $j = 1, \dots, d$ and

$$B_j = \sigma_\mu \left(\frac{2}{t} - \frac{2}{t^2}\left(\frac{1}{t} + \frac{1}{\sigma_{g,i}^2}\right)^{-1} + \frac{1}{\sigma_\mu^2}\right)^{-\frac{1}{2}} \left(\frac{1}{t} + \frac{1}{\sigma_\mu^2} - \frac{1}{t^2}\left(\frac{1}{t} + \frac{1}{\sigma_{g,i}^2}\right)^{-1}\right)$$

Therefore, the importance measure of the j^{th} input is directly linked with the dispersion parameter $\sigma_{g,i}^2$ of the function $g(x)$. Furthermore, when t tends to the infinity, the response $u(x,t)$ tends to zero as the variance of the main effect. In this section, we consider the response at $t = 1$.

6.2 Model selection

Let us consider a Gaussian process of covariance $k_u(x, \tilde{x})$ and mean m_u to surrogate $u(x,t)$ at $t = 1$. We consider the predictive mean and variance presented in equations (1) and (2). As the response $u(x,t)$ is smooth, we choose a squared exponential covariance kernel:

$$k_u(x, \tilde{x}) = \sigma^2 \exp\left(-\frac{1}{2}\sum_{i=1}^d \frac{(x^i - \tilde{x}^i)^2}{\theta_i^2}\right)$$

Furthermore, as $u(x,t)$ tends to zero when x tends to the infinity, we consider that $m_u = 0$. Indeed, we want that the model tends to zero when we move away from the design points.

The experimental design set $\mathbf{D}$ is composed of $n = 3000$ training points x_i^{train} sampled from the multivariate normal distribution $\mathcal{N}\left(0, \sigma_\mu^2 \mathbf{I}\right)$ with $\sigma_\mu = 2$ and $d = 5$. Furthermore, the initial budget is $T_0 = 3000$. It corresponds to a unique repetition $r_0 = 1$ at each point of $\mathbf{D}$. The n observations of $u_{r_0}^{\text{code}}(x, 1)$ at points in $\mathbf{D}$ are denoted by $\mathbf{u}^n$.

The hyper-parameters σ^2, θ and σ_ε^2 are estimated by maximizing the marginal Likelihood [16]:

$$-\frac{1}{2}\left(\mathbf{u}^n\right)' \left(\sigma^2 \mathbf{K} + \sigma_\varepsilon \mathbf{I}\right)^{-1} \mathbf{u}^n - \frac{1}{2}\det\left(\sigma^2 \mathbf{K} + \sigma_\varepsilon \mathbf{I}\right)$$

where $\mathbf{K} = [k_u(x_i, x_j)]_{i,j=1,\dots,n}$. To solve the maximization problem, we have first randomly generated a set of 1,000 parameters $(\sigma^2, \theta, \sigma_\varepsilon)$ on the domain $(0,10) \times (0,2)^d \times (0,1)$ and we have started a quasi-Newton based maximization from the 10 best parameters using the BFGS method. We obtain the following parameter estimations.

- $\hat{\theta} = \begin{pmatrix} 1.01 & 1.02 & 1.03 & 1.00 & 1.07 \end{pmatrix}$
- $\hat{\sigma}^2 = 1.46$
- $\hat{\sigma}_\varepsilon^2 = 6.74.10^{-2}$

Furthermore, the dispersion term of $g(x)$ are set to:

- $(\sigma_{g,i}^2)_{i=1,\dots,d} = (5, 3, 2, 1, 1)$

6.3 Convergence of IMSE_T

As presented in Subsection 3.2 and Section 4, the asymptotic normality of the Sobol index estimator is closely related to the convergence of the generalization error IMSE_T (12). Therefore, in order to effectively estimate the confidence intervals of the estimators, we have to characterize this convergence. Especially, we have to take into account the initial budget used to select the model. The value of IMSE_{T_0} where T_0 corresponds to the initial budget allocated to $\mathbf{D}$ is estimated to $\text{IMSE}_{T_0} = 6.06.10^{-1}$. According to (30), we have the following convergence rate for IMSE_T with respect to T:

$$\text{IMSE}_T \sim \left(\sigma_\varepsilon^2/T\right) \log\left(T/\sigma_\varepsilon^2\right)/\xi_d$$

Therefore, from an initial budget T_0 we expect that IMSE_T as a function of T decays as:

$$\text{IMSE}_T = \text{IMSE}_{T_0} \frac{T_0 \log\left(T/\sigma_\varepsilon^2\right)}{T \log\left(T_0/\sigma_\varepsilon^2\right)}$$

The critical ratio $mB_T^2 = 1$ presented in Section 5 leads to the following budget:

$$T = \frac{m}{C} \log\left(\frac{m}{C\sigma_\varepsilon^2}\right) \tag{37}$$

with $C = \log\left(T_0/\sigma_\varepsilon^2\right)/(T_0 \text{IMSE}_{T_0})$.

6.4 Confidence intervals for the Sobol index estimations

According to Theorem 1, if T follows the relation in (37), the Sobol index estimator presented in Subsection 4.1 is asymptotically distributed with respect to a Gaussian random variable centered on the true index and with variance given in (20). We use this property to build 90% confidence intervals on the estimations of $(S^j)_{j=1,\dots,d}$ (36). The exact values of the Sobol indices (36) are given by:

$$(S^j)_{j=1,\dots,d} = (0.052, 0.088, 0.124, 0.194, 0.194)$$

Remember that m represents the number of particles for the Monte-Carlo integrations and T is the budget used to construct the surrogate model $\hat{z}_T(x)$. In order to illustrate the relevance of (37), we consider the following equation:

$$T = \sigma_\varepsilon^2 \frac{m^\alpha}{C} \log\left(\frac{m}{C}\right)$$

with different values of α - the right value being $\alpha = 1$ - and different values of m. For each combination (α, m), we estimate the Sobol indices with the estimator (18) and from 500 different Monte-Carlo samples $(x_i^{\text{MC}})_{i=1,\dots,m}$. For each sample we evaluate the 90% confidence intervals thanks to (20) and we check if the estimations are covered or not. The result of the procedure is presented in Table 1.

m	α	S^1	S^2	S^3	S^4	S^5
1,000	0.8	88.00	86.20	87.60	88.20	86.40
1,000	0.9	89.00	91.80	89.60	86.20	86.00
1,000	1.0	88.40	87.00	89.40	87.60	90.80
1,000	1.1	88.00	89.40	88.80	87.00	88.60
1,000	1.2	90.00	91.00	86.60	88.80	89.00
3,000	0.8	88.00	87.60	86.60	87.80	87.20
3,000	0.9	89.80	87.80	87.40	88.60	88.00
3,000	1.0	89.40	90.40	89.20	89.40	89.60
3,000	1.1	90.40	90.60	91.00	91.60	90.80
3,000	1.2	92.00	91.80	92.00	91.40	91.40
5,000	0.8	87.60	86.20	87.40	88.20	86.40
5,000	1.0	89.20	89.40	90.80	89.80	89.60
5,000	1.2	92.00	91.40	92.80	90.60	92.20

Table 1: Coverage rates for $(S^j)_{j=1,\dots,d}$ in percentage. The confidence intervals are built from the variance presented in (20) in Theorem 1. The theoretical rates is 90% and the estimations is performed from 500 different Monte-Carlo samples.

We see in Table 1 that the asymptotic behavior is not reached for $m = 1,000$ Monte-Carlo particles since the coverage is globally too low in this case for every α. Furthermore, for $m = 3,000$ and $m = 5,000$, we see that the coverage is globally better for $\alpha = 1$ than for the other values. Indeed, the covering rate is underestimated for $\alpha < 1$ and often overestimated for $\alpha > 1$ whereas it is always around 90% for $\alpha = 1$. Furthermore, the confidence intervals seem to be well evaluated either for large values of S^j with S^4 and S^5, for intermediate values of S^j with S^3 or for small values of S^j with S^1 and S^2. Therefore, this example emphasizes the relevance of the asymptotic normality for the Sobol index estimators presented in Theorem 1.

7 Conclusion

This paper focuses on the estimation of the Sobol indices to perform global sensitivity analysis for stochastic simulators. We suggest an index estimator which combines a Monte-Carlo scheme to estimate the integrals involved in the index definition and a Gaussian process regression to surrogate the stochastic simulator. The surrogate model is necessary since the Monte-Carlo integrations require an important number of simulations.

In a stochastic simulator framework, for a fixed computational budget the observation noise variance is inversely proportional to the number of simulations. In this paper, we consider the special case of a large number of observations with an important uncertainty on the output. This choice allows us to consider an idealized version of the regression problem from which we can define a surrogate model which is tractable for our purpose.

In particular we aim to build confidence intervals for the index estimator taking into account both the uncertainty due to the Monte-Carlo integrations and the one due to the surrogate modelling. To handle this point, we present a theorem providing sufficient conditions to ensure the asymptotic normality of the suggested estimator. The proof of the theorem is the main point of this paper. It gives a closed form expression for the variance of the asymptotic

distribution of the estimator. From it we can easily estimate the desired confidence intervals. Furthermore, a strength of the suggested theorem is that it gives the relation between the number of particles for the Monte-Carlo integrations and the computational budget allocated to the surrogate model so that they have the same contribution on the error of the Sobol index estimations.

8 Aknowledgments

The author is grateful to his supervisor Dr. Josselin Garnier for his fruitful guidance and constructive suggestions.

A Proof of Theorem 1

Let us denote by $S_{T_m}^{X^1} = \mathrm{cov}_X\left(\hat{z}_{T_m}(X), \hat{z}_{T_m}(\tilde{X})\right)/\mathrm{var}_X\left(\hat{z}_{T_m}(X)\right)$ the variance of the main effect of X^1 for the surrogate model $\hat{z}_{T_m}(x)$ (4). The random variables S^{X^1} and $S_{T_m}^{X^1}$ are defined on the probability space $(\Omega_Z, \mathcal{F}_Z\ , \mathbb{P}_Z)$ and the random variables $S_{T_m,m}^{X^1}$, $\hat{z}_{T_m}(X)$ and $f(X)$ are defined on the product probability space $(\Omega_Z \times \Omega_X, \sigma(\mathcal{F}_Z \times \mathcal{F}_X), \mathbb{P}_Z \otimes \mathbb{P}_X)$.

Let us consider the following decomposition:

$$S_{T_m,m}^{X^1} - S^{X^1} = S_{T_m,m}^{X^1} - S_{T_m}^{X^1} + S_{T_m}^{X^1} - S^{X^1} \tag{38}$$

In a first hand we deal with the convergence of $\sqrt{m}\left(S_{T_m,m}^{X^1} - S_{T_m}^{X^1}\right)$. We handle this problem thanks to the Skorokhod's representation theorem, the Lindeberg-Feller theorem and the Delta method. In a second hand, we study the convergence of $\sqrt{m}\left(S_{T_m}^{X^1} - S^{X^1}\right)$ through the Skorokhod's representation theorem.

In the forthcoming developments, we consider that $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} 0$. Therefore, there exists $g(T_m)$ such that $g(T_m) \overset{m\to\infty}{\longrightarrow} 0$ and $mB_{T_m}^2 g^{-2}(T_m) \overset{m\to\infty}{\longrightarrow} 0$. The function $g(T_m)$ considered in the remainder of this section satisfies this property.

A.1 The Skorokhod's representation theorem

Let us consider the following random variables defined on the probability space $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$:

$$a_{T_m}(x) = (\hat{z}_{T_m}(x) - f(x))B_{T_m}^{-1} g(T_m) \tag{39}$$

$$b_{T_m}(x) = (\hat{z}_{T_m}(x) - f(x))g(T_m)^{1/3} B_{T_m}^{-1/3} \tag{40}$$

Markov's inequality and (14) give us $\forall \delta > 0$:

$$\mathbb{P}_Z(||a_{T_m}(x)||_{L_\mu^2}^2 > \delta) \leq \mathbb{E}_Z(||a_{T_m}(x)||_{L_\mu^2}^2)/\delta \leq g(T_m)^2/\delta$$

Therefore, we have the following convergence in probability in $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$:

$$\lim_{m\to\infty} ||a_{T_m}(x)||_{L_\mu^2}^2 = 0$$

and the inequalities in (14) ensure the following one:

$$||a_{T_m}(x)||_{L_\mu^2}^2 \geq g(T_m)^2/2 \tag{41}$$

Furthermore, the following equality stands since $f(x)$ is a Gaussian process:

$$\mathbb{E}_Z[(\hat{z}_{T_m}(x) - f(x))^6] = 15\sigma^6_{T_m}(x)$$

Cauchy-Schwarz inequality leads to:

$$\mathbb{E}_Z[||\hat{z}_{T_m}(x) - f(x)||^6_{L^6_\mu}] \leq 15 \int \sigma^6_{T_m}(x)\, d\mu(x) \leq 15 B^2_{T_m} \sup_x k^2(x,x)$$

Therefore, thanks to Markov's inequality we have:

$$\mathbb{P}_Z(||b_{T_m}(x)||^6_{L^6_\mu} > \delta) \leq 15 g(T_m)^2 \sup_x k^2(x,x)/\delta$$

and the following convergence stands in probability in $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$:

$$\lim_{m\to\infty} ||b_{T_m}(x)||^6_{L^6_\mu} = 0$$

Therefore, we have the following convergences in probability in $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$ when $m \to \infty$:

$$\begin{cases} f(x) \\ a_{T_m}(x) = (\hat{z}_{T_m}(x) - f(x)) g(T_m) B^{-1}_{T_m} \\ b_{T_m}(x) = (\hat{z}_{T_m}(x) - f(x)) g(T_m)^{1/3} B^{-1/3}_{T_m} \end{cases} \overset{\substack{L^6_\mu \times L^2_\mu \times L^6_\mu \\ \longrightarrow \\ m\to\infty}}{} \begin{pmatrix} f(x) \\ 0 \\ 0 \end{pmatrix}$$

As $L^6_\mu \times L^2_\mu \times L^6_\mu$ is separable we can use the Skorokhod's representation theorem [20] presented below.

Theorem 2 (Skorokhod's representation theorem)**.** *Let μ_n, $n \in \mathbb{N}$ be a sequence of probability measures on a topological space S; suppose that μ_n converges weakly to some probability measure μ on S as $n \to \infty$. Suppose also that the support of μ is separable. Then there exist random variables X_n and X defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that:*

(i) μ_n is the distribution of X_n

(ii) μ is the distribution of X

(iii) $X_n(\omega) \to X(\omega)$ as $n \to \infty$ for every $\omega \in \Omega$.

Therefore, there is a probability space denoted by $(\tilde{\Omega}_Z, \tilde{\mathcal{F}}_Z, \tilde{\mathbb{P}}_Z)$ such that

$$(\tilde{f}_{T_m}(x), \tilde{a}_{T_m}(x), \tilde{b}_{T_m}(x)) \overset{\mathcal{L}}{=} (f(x), a_{T_m}(x), b_{T_m}(x)), \quad \forall m \tag{42}$$

with $(\tilde{f}_{T_m}(x), \tilde{a}_{T_m}(x), \tilde{b}_{T_m}(x))$, $\tilde{f}(x)$ defined on $(\tilde{\Omega}_Z, \tilde{\mathcal{F}}_Z, \tilde{\mathbb{P}}_Z)$ and $(f(x), a_{T_m}(x), b_{T_m}(x))$ defined on $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$ - and $\forall \tilde{\omega}_Z \in \tilde{\Omega}_Z$ the following convergence holds for $m \to \infty$:

$$(\tilde{f}_{T_m}(x), \tilde{a}_{T_m}(x), \tilde{b}_{T_m}(x)) \overset{\substack{L^6_\mu \times L^2_\mu \times L^6_\mu \\ \longrightarrow \\ m\to\infty}}{} (\tilde{f}(x), 0, 0) \tag{43}$$

First, let us build below the analogous of $z_{T_m}(x)$ in $(\tilde{\Omega}_Z, \tilde{\mathcal{F}}_Z, \tilde{\mathbb{P}}_Z)$. For a fixed $T_m > 0$, we have the equality $a_{T_m}(x) g(T_m)^{-1} B_{T_m} = b_{T_m}(x) g(T_m)^{-1/3} B^{1/3}_{T_m}$. Therefore, we have

$$||a_{T_m}(x) g(T_m)^{-1} B_{T_m} - b_{T_m}(x) g(T_m)^{-1/3} B^{1/3}_{T_m}||_{L^2_\mu} = 0$$

and

$$\mathbb{P}_Z\left(||a_{T_m}(x)g(T_m)^{-1}B_{T_m} - b_{T_m}(x)g(T_m)^{-1/3}B_{T_m}^{1/3}||_{L^2_\mu} = 0\right) = 1$$

The equality $(\tilde{a}_{T_m}(x), \tilde{b}_{T_m}(x)) \stackrel{\mathcal{L}}{=} (a_{T_m}(x), b_{T_m}(x))$ $\forall T_m$ leads to the following one

$$\tilde{\mathbb{P}}_Z\left(||\tilde{a}_{T_m}(x)g(T_m)^{-1}B_{T_m} - \tilde{b}_{T_m}(x)g(T_m)^{-1/3}B_{T_m}^{1/3}||_{L^2_\mu} = 0\right) = 1$$

Thus, for almost every $\tilde{\omega_Z}$ in $\tilde{\Omega}_Z$, we have

$$||\tilde{a}_{T_m}(x)g(T_m)^{-1}B_{T_m} - \tilde{b}_{T_m}(x)g(T_m)^{-1/3}B_{T_m}^{1/3}||_{L^2_\mu} = 0 \tag{44}$$

If we consider such a $\tilde{\omega}_Z$ we have the equality $\tilde{a}_{T_m}(x)g(T_m)^{-1}B_{T_m} = \tilde{b}_{T_m}(x)g(T_m)^{-1/3}B_{T_m}^{1/3}$ for μ-almost every x.

Let us denote by

$$\tilde{z}_{T_m}(x) = \tilde{f}_{T_m}(x) + g(T_m)^{-1}B_{T_m}\tilde{a}_{T_m}(x),$$

$\tilde{z}_{T_m}(x)$ is defined on $(\tilde{\Omega}_Z, \tilde{\mathcal{F}}_Z, \tilde{\mathbb{P}}_Z)$. For $\tilde{\omega}_Z$ such that (44) holds we have the equality $\tilde{z}_{T_m}(x) = \tilde{f}_{T_m}(x) + g(T_m)^{-1/3}B_{T_m}^{1/3}\tilde{b}_{T_m}(x)$ for μ-almost every x.

A.2 Convergences with a fixed $\tilde{\omega}_Z \in \tilde{\Omega}_Z$

Let us consider a fixed $\tilde{\omega}_Z \in \tilde{\Omega}_Z$ such that (44) holds. We aim to study the convergence of $\sqrt{m}\left(\tilde{S}^{X^1}_{T_m,m} - \tilde{S}^{X^1}_{T_m}\right)$ and $\sqrt{m}\left(\tilde{S}^{X^1}_{T_m} - \tilde{S}^{X^1}\right)$ in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ with:

$$\tilde{S}^{X^1} = \mathrm{cov}_X(\tilde{f}(X), \tilde{f}(\tilde{X}))/\mathrm{var}_X(\tilde{f}(X)), \tag{45}$$

$$\tilde{S}^{X^1}_{T_m} = \mathrm{cov}_X(\tilde{z}_{T_m}(X), \tilde{z}_{T_m}(\tilde{X}))/\mathrm{var}_X(\tilde{z}_{T_m}(X)) \tag{46}$$

and

$$\tilde{S}^{X^1}_{T_m,m} = \frac{m^{-1}\sum_{i=1}^n \tilde{z}_{T_m}(X_i)\tilde{z}_{T_m}(\tilde{X}_i) - m^{-2}\sum_{i,j=1}^n \tilde{z}_{T_m}(X_i)\tilde{z}_{T_m}(\tilde{X}_j)}{m^{-1}\sum_{i=1}^n \tilde{z}^2_{T_m}(X_i) - m^{-2}(\sum_{i=1}^n \tilde{z}_{T_m}(X_i))^2} \tag{47}$$

A.2.1 Convergence of $\sqrt{m}\left(\tilde{S}^{X^1}_{T_m,m} - \tilde{S}^{X^1}_{T_m}\right)$ in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$

Let us denote by $Y_{T_m,i} = \tilde{z}_{T_m}(X_i)$, $Y^{X^1}_{T_m,i} = \tilde{z}_{T_m}(\tilde{X}_i)$ and

$$\begin{aligned} U_{T_m,i} = \quad & \Big((Y_{T_m,i} - \mathbb{E}_X[Y_{T_m,i}])(Y^{X^1}_{T_m,i} - \mathbb{E}_X[Y_{T_m,i}]), \\ & Y_{T_m,i} - \mathbb{E}_X[Y_{T_m,i}], Y^{X^1}_{T_m,i} - \mathbb{E}_X[Y_{T_m,i}], (Y_{T_m,i} - \mathbb{E}_X[Y_{T_m,i}])^2\Big) \end{aligned} \tag{48}$$

Since $\tilde{\omega}_Z \in \tilde{\Omega}_Z$ is fixed, $Y_{T_m,i}$, $Y^{X^1}_{T_m,i}$ and $U_{T_m,i}$ are defined on the probability space $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$. For each m, $(U_{T_m,i}/\sqrt{m})_{i=1,\ldots,m}$ is a sequence of independent random vectors such that for any $\varepsilon > 0$:

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}_X\left[||U_{T_m,i}||^2/m\mathbf{1}_{\{||U_{T_m,i}|| > \varepsilon\sqrt{m}\}}\right] &= \mathbb{E}_X\left[||U_{T_m,1}||^2\mathbf{1}_{\{||U_{T_m,1}|| > \varepsilon\sqrt{m}\}}\right] \\ &\leq \mathbb{E}_X\left[||U_{T_m,1}||^3\right]/(\varepsilon\sqrt{m}) \end{aligned}$$

since $||U_{T_m,1}|| > \varepsilon\sqrt{m}$.

We aim below to find an upper bound for $\sup_{T_m} \mathbb{E}_X\left[||U_{T_m,i}||^3\right]$. First, for any m, let us consider the component $(Y_{T_m,i}-\mathbb{E}_X[Y_{T_m}])(Y_{T_m,i}^{X^1}-\mathbb{E}_X[Y_{T_m}])$. We have the following inequality:

$$\mathbb{E}_X\left[|(Y_{T_m,i}-\mathbb{E}[Y_{T_m,i}])(Y_{T_m,i}^{X^1}-\mathbb{E}[Y_{T_m,i}])|^3\right] \leq C\mathbb{E}_X\left[|Y_{T_m,i}|^6\right]$$

with $C>0$ a constant. Minkowski inequality and the equality $\tilde{z}_{T_m}(x) = (\tilde{f}_{T_m}(x)+g(T_m)^{-1/3}B_{T_m}^{1/3}\tilde{b}_{T_m}(x))$ for μ-almost every x give that there exists $C, C'>0$ such that:

$$\mathbb{E}_X\left[|Y_{T_m,i}|^6\right] \leq C||\tilde{f}_{T_m}(x)||_{L^6_\mu}^6 + C'B_{T_m}^2 g(T_m)^{-2}||\tilde{b}_{T_m}(x)||_{L^6_\mu}^6$$

The convergence $(\tilde{f}_{T_m}(x),\tilde{b}_{T_m}(x)) \overset{L^6_\mu\times L^6_\mu}{\underset{m\to\infty}{\longrightarrow}} (\tilde{f}(x),0)$ implies that there exists $C>0$ such that for any m:

$$\mathbb{E}_X\left[|(Y_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}])(Y_{T_m,i}^{X^1}-\mathbb{E}_X[Y_{T_m,i}])|^3\right] \leq C \tag{49}$$

Second, following the same guideline, we find that there exists $C, C', C''>0$ such that for any m:

$$\mathbb{E}_X\left[|(Y_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}])^2|^3\right] \leq C \tag{50}$$

$$\mathbb{E}_X\left[|Y_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}]|^3\right] \leq C' \tag{51}$$

$$\mathbb{E}_X\left[|Y_{T_m,i}^{X^1}-\mathbb{E}_X[Y_{T_m,i}]|^3\right] \leq C' \tag{52}$$

Third, the inequalities (49), (51), (51) and (52) give that $\sup_{T_m}\mathbb{E}_X\left[||U_{T_m}||^3\right]<\infty$.

The inequality $\sum_{i=1}^m \mathbb{E}_X\left[||U_{T_m,i}||^2/m\mathbf{1}_{\{||U_{T_m,i}||>\varepsilon\sqrt{m}\}}\right] \leq \mathbb{E}_X\left[||U_{T_m,1}||^3\right]/(\varepsilon\sqrt{m})$ and the uniform boundedness of $\mathbb{E}_X\left[||U_{T_m}||^3\right]$ lead to the following convergence $\forall\varepsilon>0$ when $m\to\infty$:

$$\sum_{i=1}^m \mathbb{E}_X\left[||U_{T_m,i}||^2/m\mathbf{1}_{\{||U_{T_m,i}||>\varepsilon\sqrt{m}\}}\right] = \mathbb{E}_X\left[||U_{T_m,i}||^2\mathbf{1}_{\{||U_{T_m,i}||>\varepsilon\sqrt{m}\}}\right] \overset{m\to\infty}{\longrightarrow} 0 \tag{53}$$

and thus $||U_{T_m,i}||^2$ is uniformly integrable.

Now, we aim to show the convergence in probability of $U_{T_m,i} \overset{m\to\infty}{\longrightarrow} U_i$ in $(\Omega_X,\mathcal{F}_X,\mathbb{P}_X)$. Let us denote by

$$U_i = \Big((Y_i-\mathbb{E}_X[Y_i])(Y_i^{X^1}-\mathbb{E}_X[Y_i]), Y_i-\mathbb{E}_X[Y_i], Y_i^{X^1}-\mathbb{E}_X[Y_i], (Y_i-\mathbb{E}_X[Y_i])^2\Big)$$

with $Y_i=\tilde{f}(X_i)$ and $Y_i^{X^1}=\tilde{f}(\tilde{X}_i)$. The random variables U_i, Y_i and $Y_i^{X^1}$ are defined on $(\Omega_X,\mathcal{F}_X,\mathbb{P}_X)$ since $\tilde{\omega}_Z\in\tilde{\Omega}_Z$ is fixed.

First, we study the term $\mathbb{E}_X\left[\left|U_{T_m,i}^{(1)}-U_i^{(1)}\right|\right]$ where $U_i^{(1)}=(Y_i-\mathbb{E}_X[Y_i])(Y_i^{X^1}-\mathbb{E}_X[Y_i])$ and $U_{T_m,i}^{(1)}=(Y_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}])(Y_{T_m,i}^{X^1}-\mathbb{E}_X[Y_{T_m,i}])$. We have the following equality:

$$\begin{aligned}\mathbb{E}_X\left[\left|U_{T_m,i}^{(1)}-U_i^{(1)}\right|\right] &= \mathbb{E}_X\Big[\Big|(Y_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}])\Big((Y_{T_m,i}^{X^1}-\mathbb{E}_X[Y_{T_m,i}])-(Y_i^{X^1}-\mathbb{E}_X[Y_i])\Big)\\ &+ (Y_i^{X^1}-\mathbb{E}_X[Y_i])\Big((Y_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}])-(Y_i-\mathbb{E}_X[Y_i])\Big)\Big|\Big]\end{aligned}$$

from which we deduce the inequality:

$$\begin{aligned}\mathbb{E}_X\left[\left|U^{(1)}_{T_m,i}-U^{(1)}_i\right|\right] &\leq \mathbb{E}_X\left[\left|(Y_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}])\left((Y^{X^1}_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}])-(Y_i^{X^1}-\mathbb{E}_X[Y_i])\right)\right|\right]\\ &+ \mathbb{E}_X\left[\left|(Y_i^{X^1}-\mathbb{E}_X[Y_i])\Big((Y_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}])-(Y_i-\mathbb{E}_X[Y_i])\Big)\right|\right]\end{aligned}$$

and from Cauchy-Schwarz inequality there exists $C, C', C'' > 0$ such that:

$$\begin{aligned}\mathbb{E}_X\left[\left|U^{(1)}_{T_m,i}-U^{(1)}_i\right|\right] &\leq C\mathbb{E}_X\left[(Y_{T_m,i}-\mathbb{E}_X[Y_{T_m,i}])^2\right]^{1/2}\mathbb{E}_X\left[(Y^{X^1}_{T_m,i}-Y_i^{X^1})^2\right]^{1/2}\\ &+ C'\mathbb{E}_X\left[(Y_i^{X^1}-\mathbb{E}_X[Y_i])^2\right]^{1/2}\mathbb{E}_X\left[(Y_{T_m,i}-Y_i)^2\right]^{1/2}\\ &\leq C''\mathbb{E}_X\left[(Y_{T_m,i}-Y_i)^2\right]^{1/2}\left(\mathbb{E}_X\left[(Y_i^{X^1})^2\right]^{1/2}+\mathbb{E}_X\left[(Y_{T_m,i})^2\right]^{1/2}\right)\end{aligned}$$

The equality $Y_{T_m,i}-Y_i = g(T_m)^{-1}B_{T_m}\tilde{a}_{T_m}(X_i)$ for $\mathbb{P}_X$-almost every $\omega_X \in \Omega_X$ implies that $\mathbb{E}_X\left[(Y_{T_m,i}-Y_i)^2\right]^{1/2} = g(T_m)^{-1}B_{T_m}\mathbb{E}_X\left[(\tilde{a}_{T_m}(X_i))^2\right]^{1/2}$. Since $\tilde{a}_{T_m}(x) \overset{m\to\infty}{\longrightarrow} 0$ in L^2_μ, we have the convergence $\mathbb{E}_X\left[(Y_{T_m,i}-Y_i)^2\right]^{1/2} \overset{m\to\infty}{\longrightarrow} 0$.

Furthermore, there exists $C, C' > 0$ such that $\mathbb{E}_X\left[(Y_i^{X^1})^2\right]^{1/2} < C$ and $\mathbb{E}_X\left[(Y_{T_m,i})^2\right]^{1/2} < C'$ since $\tilde{z}_{T_m}(x) = \tilde{f}_{T_m}(x) + g(T_m)^{-1}B_{T_m}\tilde{a}_{T_m}(x)$, $\tilde{f}_{T_m}(x) \overset{m\to\infty}{\longrightarrow} \tilde{f}(x)$ in L^6_μ and $\tilde{a}_{T_m}(x) \overset{m\to\infty}{\longrightarrow} 0$ in L^2_μ. Therefore, we have the following convergence:

$$\mathbb{E}_X\left[\left|U^{(1)}_{T_m,i}-U^{(1)}_i\right|\right] \overset{m\to\infty}{\longrightarrow} 0 \tag{54}$$

Then, if we consider the terms $U^{(4)}_i = (Y_i - \mathbb{E}_X[Y_i])^2$ and $U^{(4)}_{T_m,i} = (Y_{T_m,i} - \mathbb{E}_X[Y_{T_m,i}])^2$. Following the same guideline we find the convergence:

$$\mathbb{E}_X\left[\left|U^{(4)}_{T_m,i}-U^{(4)}_i\right|\right] \overset{m\to\infty}{\longrightarrow} 0 \tag{55}$$

Furthermore, denoting by $U^{(2)}_i = (Y_i - \mathbb{E}_X[Y_i])$, $U^{(2)}_{T_m,i} = (Y_{T_m,i} - \mathbb{E}_X[Y_{T_m,i}])$, $U^{(3)}_i = (Y_i^{X^1} - \mathbb{E}_X[Y_i])$ and $U^{(3)}_{T_m,i} = (Y^{X^1}_{T_m,i} - \mathbb{E}_X[Y_{T_m,i}])$, we have the following inequalities:

$$\mathbb{E}_X\left[\left|U^{(2)}_{T_m,i}-U^{(2)}_i\right|\right] \leq C\mathbb{E}_X\left[(Y_{T_m,i}-Y_i)^2\right]^{1/2}$$

$$\mathbb{E}_X\left[\left|U^{(3)}_{T_m,i}-U^{(3)}_i\right|\right] \leq C'\mathbb{E}_X\left[(Y^{X^1}_{T_m,i}-Y_i^{X^1})^2\right]^{1/2}$$

with C, C' positive constants. The convergences $\tilde{f}_{T_m}(x) \overset{L^6_\mu}{\to} \tilde{f}(x)$ and $\tilde{a}_{T_m}(x) \overset{L^6_\mu}{\to} 0$ when $m \to \infty$ ensure that:

$$\mathbb{E}_X\left[\left|U^{(2)}_{T_m,i}-U^{(2)}_i\right|\right] \overset{m\to\infty}{\longrightarrow} 0 \tag{56}$$

and

$$\mathbb{E}_X\left[\left|U^{(3)}_{T_m,i}-U^{(3)}_i\right|\right] \overset{m\to\infty}{\longrightarrow} 0 \tag{57}$$

Finally, the convergences presented in (54), (55), (56) and (57) imply the desired one:

$$\mathbb{E}_X\left[||U_{T_m,i}-U_i||\right] \overset{m\to\infty}{\longrightarrow} 0 \tag{58}$$

Markov's inequality gives $\forall \delta > 0$:

$$\mathbb{P}_X\left(||U_{T_m,i} - U_i|| \geq \delta\right) \leq \mathbb{E}_X\left[||U_{T_m,i} - U_i||\right]/\delta \tag{59}$$

The equations (58) and (59) imply the convergence $U_{T_m,i} \overset{m\to\infty}{\longrightarrow} U_i$ in probability in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$.

This convergence in probability and the uniform integrability of $||U_{T_m,i}||^2$ implies that $U_{T_m,i} \overset{m\to\infty}{\longrightarrow} U_i$ in $L^2(\Omega_X)$ and thus $\mathrm{cov}_X(U_{T_m,i}) \overset{m\to\infty}{\longrightarrow} \mathrm{cov}_X(U_i) = \boldsymbol{\Sigma}$. We note that we have also the convergence $\mathbb{E}_X[U_{T_m,i}] \to \mathbb{E}_X[U_i] = \boldsymbol{\mu}$ since the convergence in $L^2(\Omega_X)$ implies the one in $L^1(\Omega_X)$.

The condition (53) and the convergence $\sum_{i=1}^m \mathrm{cov}_X(U_{T_m,i})/m = \mathrm{cov}_X(U_{T_m,i}) \overset{m\to\infty}{\longrightarrow} \boldsymbol{\Sigma}$ allow for using the Lindeberg-Feller Theorem (see [15]) which ensures the following convergence in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$:

$$\begin{aligned}\sum_{i=1}^m (U_{T_m,i}/\sqrt{m} - \mathbb{E}_X[U_{T_m,i}/\sqrt{m}]) \quad &= \quad \sqrt{m}\left(\sum_{i=1}^m (U_{T_m,i})/m - \mathbb{E}_X[U_{T_m,i}]\right) \\ &\underset{m\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \quad \mathcal{N}(0, \boldsymbol{\Sigma})\end{aligned}$$

Furthermore, we have the following equality:

$$\tilde{S}^{X^1}_{T_m,m} = \Phi(\bar{U}_{T_m})$$

where $\bar{U}_{T_m} = \sum_{i=1}^m U_{T_m,i}/m$ and $\Phi(x,y,z,t) = (x - yz)/(t - y^2)$. Therefore, the Delta method gives that in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$:

$$\sqrt{m}\left(\tilde{S}^{X^1}_{T_m,m} - \tilde{S}^{X^1}_{T_m}\right) \underset{m\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}\left(0, \nabla\Phi^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\nabla\Phi(\boldsymbol{\mu})\right) \tag{60}$$

where $\boldsymbol{\mu} = \mathbb{E}_X[U_i] = \left(\mathrm{cov}_X(Y_i, Y_i^{X^1}), 0, 0, \mathrm{var}_X(Y_i)\right)$. We note that the assumption $\mathrm{var}_X(Y_i) \neq 0$ justifies the use of the Delta method. A simple calculation gives that:

$$\nabla\Phi^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\nabla\Phi(\boldsymbol{\mu}) = \frac{\mathrm{var}_X\left((Y_i - \mathbb{E}_X[Y_i])\left(Y_i^{X^1} - \mathbb{E}_X[Y_i] - \tilde{S}^{X^1}Y_i + \tilde{S}^{X^1}\mathbb{E}_X[Y_i]\right)\right)}{(\mathrm{var}_X(Y_i))^2} \tag{61}$$

with $\tilde{S}^{X^1} = \mathrm{cov}_X(Y_i, Y_i^{X^1})/\mathrm{var}_X(Y_i) = \mathrm{var}_X(\mathbb{E}_X[Y_i|X^1])/\mathrm{var}_X(Y_i)$.

A.2.2 Convergence of $\sqrt{m}\left(\tilde{S}^{X^1}_{T_m} - \tilde{S}^{X^1}\right)$ in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$

Analogously to [12], we have the equality:

$$\tilde{S}^{X^1}_{T_m} - \tilde{S}^{X^1} \quad = \quad \frac{\mathrm{var}_X(\tilde{\delta}_{T_m,i})^{1/2} C_{\tilde{\delta}_{T_m,i}}}{\mathrm{var}_X(Y_i) + 2\mathrm{cov}_X(Y_i, \tilde{\delta}_{T_m,i}) + \mathrm{var}_X(\tilde{\delta}_{T_m,i})}$$

where $\tilde{\delta}_{T_m}(x) = g(T_m)^{-1} B_{T_m} \tilde{a}_{T_m}(x)$,

$$\begin{aligned}C_{\tilde{\delta}_{T_m,i}} \quad = \quad & 2\mathrm{var}_X(Y_i)^{1/2}(\mathrm{cor}_X(Y_i, \tilde{\delta}_{T_m,i}) - \mathrm{cor}_X(Y_i, Y_i^{X^1})\mathrm{cor}_X(Y_i, \tilde{\delta}_{T_m,i})) \\ & + \mathrm{var}_X(\tilde{\delta}_{T_m,i})^{1/2}(\mathrm{cor}_X(\tilde{\delta}_{T_m,i}, \tilde{\delta}^{X^1}_{T_m,i}) - \mathrm{cor}_X(Y_i, Y_i^{X^1}))\end{aligned} \tag{62}$$

$\tilde{\delta}_{T_m,i} = \tilde{\delta}_{T_m,i}(X_i)$ and $\tilde{\delta}^{X^1}_{T_m,i} = \tilde{\delta}_{T_m,i}(\tilde{X}_i)$. The random variables $\tilde{\delta}_{T_m,i}$ and $\tilde{\delta}^{X^1}_{T_m,i}$ are defined on the product space $(\tilde{\Omega}_Z \times \Omega_X, \sigma(\tilde{\mathcal{F}}_Z \times \mathcal{F}_X), \tilde{\mathbb{P}}_Z \otimes \mathbb{P}_X)$ and $\tilde{S}^{X^1}$, $\tilde{\delta}_{T_m}(x)$ and $C_{\tilde{\delta}_{T_m,i}}$ are defined on $(\tilde{\Omega}_Z, \tilde{\mathcal{F}}_Z, \tilde{\mathbb{P}}_Z)$. We still consider a fixed $\tilde{\omega}_Z \in \tilde{\Omega}_Z$ such that (44) holds. The assumption $\mathrm{var}_X(Y_i) \neq 0$ ensures that the denominator is not equal to zero and the convergences $\tilde{f}_{T_m}(x) \overset{L^6_\mu}{\to} \tilde{f}(x)$ and $\tilde{a}_{T_m}(x) \overset{L^2_\mu}{\to} 0$ give that $\sup_m C_{\tilde{\delta}_{T_m,i}} < \infty$. Furthermore, since $\tilde{a}_{T_m}(x) \overset{L^2_\mu}{\to} 0$ we have the following inequalities:

$$\mathrm{var}_X(\tilde{\delta}_{T_m,i}) \leq C\mathbb{E}_X[(B_{T_m} g(T_m)^{-1} \tilde{a}_{T_m}(X_i))^2] \leq C' g(T_m)^{-2} B^2_{T_m}$$

with C, C' positive constants.

Thanks to Slutsky's theorem, the convergence $m g(T_m)^{-2} B^2_{T_m} \overset{m}{\longrightarrow} 0$ ensures the following asymptotic normality when $m \to \infty$ in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$:

$$\sqrt{m}\left(\tilde{S}^{X^1}_{T_m,m} - \tilde{S}^{X^1}\right) \underset{m\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}\left(0, \nabla\Phi^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\nabla\Phi(\boldsymbol{\mu})\right) \tag{63}$$

A.2.3 The case $mB^2_{T_m} \overset{m\to\infty}{\longrightarrow} \infty$.

Let us suppose that $mB^2_{T_m} \overset{m\to\infty}{\longrightarrow} \infty$. We consider the convergences of

$$B^{-1}_{T_m}\left(\tilde{S}^{X^1}_{T_m,m} - \tilde{S}^{X^1}_{T_m}\right) \tag{64}$$

and

$$B^{-1}_{T_m}\left(\tilde{S}^{X^1}_{T_m} - \tilde{S}^{X^1}\right)$$

in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$ with a fixed $\tilde{\omega}_Z \in \tilde{\Omega}_Z$ such that (44) holds. We have the following equality:

$$B^{-1}_{T_m}\left(\tilde{S}^{X^1}_{T_m,m} - \tilde{S}^{X^1}_{T_m}\right) = (\sqrt{m}B_{T_m})^{-1}\sqrt{m}\left(\tilde{S}^{X^1}_{T_m,m} - \tilde{S}^{X^1}_{T_m}\right)$$

The convergence $(\sqrt{m}B_{T_m})^{-1} \overset{m\to\infty}{\longrightarrow} 0$ and the convergence in (60) (which does not depend on the convergence of the ratio between $B^{-2}_{T_m}$ and $\sqrt{m}$) imply the following one:

$$B^{-1}_{T_m}\left(\tilde{S}^{X^1}_{T_m,m} - \tilde{S}^{X^1}_{T_m}\right) \overset{m\to\infty}{\longrightarrow} 0$$

Finally, thanks to the inequality (41), there exists $C, C' > 0$ such that

$$\begin{aligned}
B^{-1}_{T_m}\left(\tilde{S}^{X^1}_{T_m} - \tilde{S}^{X^1}\right) &= B^{-1}_{T_m}\frac{g(T_m)^{-1}B_{T_m}\mathrm{var}_X(\tilde{a}_{T_m}(X_i))^{1/2}C_{\tilde{\delta}_{T_m,i}}}{\mathrm{var}_X(Y_i) + 2\mathrm{cov}_X(Y_i, \tilde{\delta}_{T_m,i}) + \mathrm{var}_X(\tilde{\delta}_{T_m,i})} \\
&\geq Cg(T_m)^{-1}\frac{g(T_m)C_{\tilde{\delta}_{T_m,i}}}{\mathrm{var}_X(Y_i) + 2\mathrm{cov}_X(Y_i, \tilde{\delta}_{T_m,i}) + \mathrm{var}_X(\tilde{\delta}_{T_m,i})} \\
&\geq C'C_{\tilde{\delta}_{T_m,i}}
\end{aligned}$$

Therefore, if we have $C_{\tilde{\delta}_{T_m,i}} > 0$, the asymptotic normality is not reached and the estimator is biased. Regarding the expression of $C_{\tilde{\delta}_{T_m,i}}$ in (62) and assuming that $\mathrm{var}_X(Y_i) \neq 0$, $C_{\tilde{\delta}_{T_m,i}} = 0$ could happen if:

- $\mathrm{cor}_X(Y_i, Y_i^{X^1}) = 1$, i.e. all the variability of $\tilde{f}(x)$ is explained by the variable X^1.
- $\mathrm{var}_X(\tilde{\delta}_{T_m,i}) = 0$, i.e. the surrogate model error is null.

A.3 Convergence in the probability space $(\Omega_Z \times \Omega_X, \sigma(\mathcal{F}_Z \times \mathcal{F}_X), \mathbb{P}_Z \otimes \mathbb{P}_X)$.

We have proved that for almost every $\tilde{\omega}_Z \in \tilde{\Omega}_Z$:

If $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} 0$, then

$$\forall I \in \mathbb{R},\ \mathbb{P}_X\left(\sqrt{m}\left(\tilde{S}_{T_m,m}^{X^1} - \tilde{S}^{X^1}\right) \in I\right) \overset{m\to\infty}{\longrightarrow} \int_I \tilde{g}(x)dx$$

If $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} \infty$, then

$$\exists C > 0 \,\text{s.t.}\, \mathbb{P}_X\left(B_{T_m}^{-1}\left(\tilde{S}_{T_m,m}^{X^1} - \tilde{S}^{X^1}\right) \geq C\right) \overset{m\to\infty}{\longrightarrow} 1$$

where $\tilde{g}(x)$ is the probability density function of a random Gaussian vector of zero mean and covariance $\nabla\Phi^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\nabla\Phi(\boldsymbol{\mu})$ (61). Therefore, in the probability space $(\tilde{\Omega}_Z \times \Omega_X, \sigma(\tilde{\mathcal{F}}_Z \times \mathcal{F}_X), \tilde{\mathbb{P}}_Z \otimes \mathbb{P}_X)$ we have

If $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} 0$, then

$$\forall I \in \mathbb{R}, \forall \delta > 0,\ \tilde{\mathbb{P}}_Z\left(\left|\mathbb{P}_X\left(\sqrt{m}\left(\tilde{S}_{T_m,m}^{X^1} - \tilde{S}^{X^1}\right) \in I\right) - \int_I \tilde{g}(x)dx\right| > \delta\right) \overset{m\to\infty}{\longrightarrow} 0$$

If $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} \infty$, then

$$\forall \delta > 0, \exists C > 0 \,\text{s.t.}\, \tilde{\mathbb{P}}_Z\left(\left|\mathbb{P}_X\left(B_{T_m}^{-1}\left(\tilde{S}_{T_m,m}^{X^1} - \tilde{S}^{X^1}\right) \geq C\right) - 1\right| > \delta\right) \overset{m\to\infty}{\longrightarrow} 0$$

and the equalities $(\tilde{f}_{T_m}(x), \tilde{a}_{T_m}(x), \tilde{b}_{T_m}(x)) \overset{\mathcal{L}}{=} (f(x), a_{T_m}(x), b_{T_m}(x))$ and $\tilde{f}(x) \overset{\mathcal{L}}{=} f(x)$ for all m give us in the probability space $(\Omega_Z \times \Omega_X, \sigma(\mathcal{F}_Z \times \mathcal{F}_X), \mathbb{P}_Z \otimes \mathbb{P}_X)$:

If $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} 0$, then

$$\forall I \in \Omega_X, \forall \delta > 0,\ \mathbb{P}_Z\left(\left|\mathbb{P}_X\left(\sqrt{m}\left(S_{T_m,m}^{X^1} - S^{X^1}\right) \in I\right) - \int_I g(x)dx\right| > \delta\right) \overset{m\to\infty}{\longrightarrow} 0$$

If $mB_{T_m}^2 \overset{m\to\infty}{\longrightarrow} \infty$, then

$$\forall \delta > 0, \exists C > 0 \,\text{s.t.}\, \mathbb{P}_Z\left(\left|\mathbb{P}_X\left(B_{T_m}^{-1}\left(S_{T_m,m}^{X^1} - S^{X^1}\right) \geq C\right) - 1\right| > \delta\right) \overset{m\to\infty}{\longrightarrow} 0$$

where $g(x)$ is the probability density function of a random Gaussian vector of zero mean and variance

$$\frac{\mathrm{var}_X\left(\left(f(X) - \mathbb{E}_X\left[f(X)\right]\right)\left(f(\tilde{X}) - \mathbb{E}_X\left[f(X)\right] - S^{X^1}f(X) + S^{X^1}\mathbb{E}_X\left[f(X)\right]\right)\right)}{\left(\mathrm{var}_X\left(f(X)\right)\right)^2}$$

This completes the proof.

References

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

[2] A. Saltelli, K. Chan, and S. E. M., *Sensitivity Analysis*. England: Wiley Series in Probability and Statistics, 2000.

[3] I. M. Sobol, "Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates," *Mathematics and Computers in Simulations*, vol. 55, pp. 271–280, 2001.

[4] D. G. Cacuci, M. Ionescu-Bujor, and I. M. Navon, *Sensitivity and Uncertainty Analysis, Volume II: Applications to Large-Scale Systems*, vol. 2. Chapman & Hall/CRC, 2005.

[5] W. Hoeffding, "A class of statistics with asymptotically normal distribution," *The annals of Mathematical Statistics*, vol. 19, no. 3, pp. 293–325, 1948.

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

[7] S. Da Veiga, F. Wahl, and F. Gamboa, "Local polynomial estimation for sensitivity analysis on models with correlated inputs," *Technometrics*, vol. 51, no. 4, pp. 452–463, 2009.

[8] T. Mara and S. Tarantola, "Variance-based sensitivity analysis of computer models with dependent inputs," *Reliability Engineering & System Safety*, vol. 107, pp. 115–121, 2012.

[9] G. Li, H. Rabitz, P. E. Yelvington, O. Oluwole, F. Bacon, K. C. E, and J. Schoendorf, "Global sensitivity analysis with independent and/or correlated inputs," *Journal of Physical Chemistry A*, vol. 114, pp. 6022–6032, 2010.

[10] G. Chastaing, F. Gamboa, and C. Prieur, "Generalized hoeffding-Sobol decomposition for dependent variables -application to sensitivity analysis," *Electronic Journal of Statistics*, vol. 6, pp. 2420–2448, 2012.

[11] I. Sobol, S. Tarantola, D. Gatelli, S. Kucherenko, and W. Mauntz, "Estimating the approximation error when fixing unessential factors in global sensitivity analysis," *Reliability Engineering & System Safety*, vol. 92, no. 7, pp. 957–960, 2007.

[12] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur, "Asymptotic normality and efficiency of two Sobol index estimators," 2012.

[13] G. Archer, A. Saltelli, and I. Sobol, "Sensitivity measures, anova-like techniques and the use of bootstrap," *Journal of Statistical Computation and Simulation*, vol. 58, no. 2, pp. 99–120, 1997.

[14] A. Janon, M. Nodet, C. Prieur, *et al.*, "Uncertainties assessment in global sensitivity indices estimation from metamodels," 2011.

[15] A. W. van der Vaart, *Asymptotic Statistics*. New-York: Cambridge University Press, 1998.

[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge: MIT Press, 2006.

[17] H. König, *Eigenvalue distribution of compact operators*. Birkhäuser Basel, 1986.

[18] J. Ferreira and V. Menegatto, "Eigenvalues of integral operators defined by smooth positive definite kernels," *Integral Equations and Operator Theory*, vol. 64, no. 1, pp. 61–81, 2009.

[19] L. Le Gratiet and J. Garnier, "Regularity dependence of the rate of convergence of the learning curve for Gaussian process regression," 2012. arXiv:1210.0686.

[20] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley Series in Probability and Statistics, 1999.

[21] M. L. Stein, *Interpolation of Spatial Data*. New York: Springer Series in Statistics, 1999.

[22] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1965.

[23] R. S. Pusev, "Small deviation asymptotics for Matèrn processes and fields under weighted quadratic norm," *Theory Probab. Appl.*, vol. 55, pp. 164–172, 2011.

[24] R. A. Todor, "Robust eigenvalue computation for smoothing operators," *SIAM J. Numer. Anal.*, vol. 44, pp. 865–878, 2006.

[25] H. Zhu, C. K. Williams, R. Rohwer, and M. Morciniec, *Gaussian regression and optimal finite dimensional linear models*. Berlin: Springer-Verlag, 1998.

[26] I. S. Gradshteyn, I. M. Ryzhik, A. Jeffrey, and D. Zwillinger, *Table of integrals, series, and products*. Academic press, 2007.