Estimating Sobol' indices combining Monte Carlo estimators and Latin hypercube sampling

Jean-Yves Tissot, Clémentine Prieur

Estimating Sobol' indices combining Monte Carlo estimators and Latin hypercube sampling

Jean-Yves Tissot* and Clémentine Prieur

Université de Grenoble LJK/INRIA

51, rue des Mathématiques
Campus de Saint Martin d'Hères, BP53
38041 Grenoble cedex 09 (France)

(jeanyvestissot@free.fr)
(clementine.prieur@imag.fr)

Phone: +33 (0)4 76 63 54 47
Fax: +33 (0)4 76 63 12 63

March 14, 2013

Abstract

In variance-based sensitivity analysis, the method of Sobol' (Sobol', 1993) allows to compute Sobol' indices using Monte Carlo integration. One of the main drawbacks of this approach is that the estimation of Sobol' indices requires the use of several distinct samples. For example, in a d-dimensional space, the estimation of all the first- or second-order Sobol' indices basically requires $d+1$ or $d(d+1)/2$ distinct samples, respectively. Some interesting combinatorial results have been introduced to weaken this defect, in particular by Saltelli (2002) and more recently by Owen (2012b), but the quantities they estimate still require $O(d)$ samples. In this paper, we introduce a new approach to estimate for any k all the k-th order Sobol' indices by using only two distinct samples. We establish theoretical properties of such a method for the first-order Sobol' indices and discuss the generalization to higher-order indices. As an illustration, we propose to apply this new approach to a marine ecosystem model of the Ligurian sea (northwestern Mediterranean) in order to study the relative importance of its several parameters. The calibration process of this kind of chemical simulators is well-known to be quite intricate, and a rigorous and robust — i.e. valid without strong regularity assumptions — sensitivity analysis, as the method of Sobol' provides, could be of great help.

Keywords: sensitivity analysis, variance-based sensitivity indices, numerical integration, orthogonal arrays

*Corresponding author.

Contents

1 Introduction and notation

Sobol' indices (SI) (Sobol', 1993) are quantities defined by normalizing parts of variance in an ANOVA decomposition (Hoeffding, 1948; Efron and Stein, 1981; Sobol', 1993). They are an important tool to study the sensitivity of a model output subject to the input parameters since they allow to quantify the relative importance of input factors of a function over their entire range of values. As they essentially consist of integrals, their computation can become rapidly expensive when the number of factors increases. Many techniques have been proposed to estimate these indices including Fast Amplitude Sensitivity Test (FAST) due to Cukier et al. (1973) — see also the review paper (Cukier et al., 1978) — and further developed by Saltelli et al. (1999), Random Balance Design (RBD) due to Tarantola et al. (2006), polynomial chaos expansion (PCE)-based estimators developed by Sudret (2008) and Blatman and Sudret (2010) and the method of Sobol'. A recent review of these methods can be found in Saltelli et al. (2008), and more specifically a new introduction to FAST and RBD has been recently provided by Tissot and Prieur (2012).

Until now, spectral methods — as FAST, RBD or PCE-based methods — which exploit the spectral decomposition of the model with respect to a particular multivariate basis, are generally preferred to the method of Sobol' because the latter is too expensive. However, spectral methods provide good estimations of SIs only under strong assumptions on the spectral decomposition of the model itself such as a decay of the spectrum sufficiently fast, the negligibility of high-order spectral coefficients, etc. As a result, these methods are not robust to complex phenomena as high-frequency variations or discontinuities, and so the method of Sobol' appears as the main method one can trust when no strong a priori knowledge on the model of interest is available.

The ANOVA decomposition proceeds as follows. Let f be a real square integrable function defined on $\mathbb{R}^d$ and $\mathbf{X} = (X_1, \ldots, X_d)$ a random vector with independent components arbitrarily distributed on $\mathbb{R}$. Then let us consider the real random variable $Y = f(\mathbf{X})$ and for any $\mathfrak{u} \subseteq \{1, \ldots, d\}$, denote by $\mathbf{X}_\mathfrak{u}$ the random vector with components X_i, $i \in \mathfrak{u}$. The ANOVA decomposition states that Y can be uniquely decomposed into summands of increasing dimensions

$$Y = \sum_{\mathfrak{u} \subseteq \{1,\ldots,d\}} f_\mathfrak{u}(\mathbf{X}_\mathfrak{u})$$

where $f_\emptyset = \mathbb{E}[Y]$ and the other components have mean zero and are mutually uncorrelated. In particular, the sum of functions

$$f_\emptyset + f_1(X_1) + f_2(X_2) + \cdots + f_d(X_d) \tag{1}$$

is the so-called additive part of f, where the constant $f_\emptyset$ and the random variables $f_i(X_i)$ are the components of lower complexity. The components $f_\mathfrak{u}(\mathbf{X}_\mathfrak{u})$ are explicitely known in terms of conditional expectation. As already mentioned, we have $f_\emptyset = \mathbb{E}[Y]$ and the other terms are recursively defined by

$$f_\mathfrak{u}(\mathbf{X}_\mathfrak{u}) = \mathbb{E}[Y|\mathbf{X}_\mathfrak{u}] - \sum_{\mathfrak{v} \subset \mathfrak{u}} f_\mathfrak{v}(\mathbf{X}_\mathfrak{v}).$$

Then denoting $\mathrm{Var}[Y]$ and any term $\mathrm{Var}[f_{\mathfrak{u}}(\mathbf{X}_{\mathfrak{u}})]$ by σ^2 and $\sigma^2_{\mathfrak{u}}$, respectively, the equation in Eq. (1) gives

$$\sigma^2 = \sum_{\mathfrak{u}\subseteq\{1,\ldots,d\}} \sigma^2_{\mathfrak{u}}$$

and the SIs — also known as global sensitivity indices — are defined as

$$S_{\mathfrak{u}} = \frac{\sigma^2_{\mathfrak{u}}}{\sigma^2}, \quad \mathfrak{u} \subseteq \{1,\ldots,d\}.$$

If $\mathfrak{u} = \{i\}$, $S_{\mathfrak{u}}$ quantifies the main effect due to the factor X_i, and if $\mathrm{Card}(\mathfrak{u}) > 1$, $S_{\mathfrak{u}}$ quantifies the interaction effect between the factors X_i, $i \in \mathfrak{u}$. Note that $S_{\emptyset}$ is trivially equal to zero. Similarly one can consider the quantities

$$\underline{S}_{\mathfrak{u}} = \frac{\underline{\tau}^2_{\mathfrak{u}}}{\sigma^2} = \frac{\mathrm{Var}\big[\mathbb{E}[Y|\mathbf{X}_{\mathfrak{u}}]\big]}{\mathrm{Var}[Y]} = \frac{\sum_{\mathfrak{v}\subseteq\mathfrak{u}} \mathrm{Var}[f_{\mathfrak{v}}(\mathbf{X}_{\mathfrak{v}})]}{\mathrm{Var}[Y]}, \quad \mathfrak{u} \subseteq \{1,\ldots,d\}$$

known as lower SIs or closed sensitivity indices. If $\mathfrak{u} = \{i\}$ then $\underline{S}_{\mathfrak{u}} = S_{\mathfrak{u}}$, and if $\mathrm{Card}(\mathfrak{u}) > 1$, $\underline{S}_{\mathfrak{u}}$ quantify the sum of all the main or interaction effects due to any group of factors $\mathfrak{v} \subseteq \mathfrak{u}$. Note that for any $1 \leq r \leq d$, the knowledge of $\{\underline{S}_{\mathfrak{u}},\ \mathrm{Card}(\mathfrak{u}) \leq r\}$ or $\{S_{\mathfrak{u}},\ \mathrm{Card}(\mathfrak{u}) \leq r\}$ are strictly equivalent since, on the one hand, we have

$$\underline{S}_{\mathfrak{u}} = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} S_{\mathfrak{v}}$$

and on the other hand, the Möbius inversion formula (see, e.g., Stanley, 2012) gives

$$S_{\mathfrak{u}} = \sum_{\mathfrak{v}\subseteq\mathfrak{u}} (-1)^{|\mathfrak{u}|-|\mathfrak{v}|} \underline{S}_{\mathfrak{v}} \ .$$

In practice global sensitivity analysis focuses on the first- or second-order — i.e. $\mathrm{Card}(\mathfrak{u}) \leq 1$ or 2, respectively — terms. In the present paper, we only focus on the indices $\underline{S}_{\mathfrak{u}}$ which are the quantities estimated by the method of Sobol'.

The paper proceeds as follows. Section 2 provides a short review of the method of Sobol', gives some notation and explains the main idea of the new method we propose in the present paper. In Section 3, theory is presented including asymptotic properties related to both Latin hypercube sampling (LHS) and replicated Latin hypercube sampling (RLHS). In this section, we also discuss potential generalizations to using orthogonal array(OA)-based Latin hypercubes (see Owen, 1992; Tang, 1993). Numerical illustrations are provided in Section 4, and Section 5 has conclusions. Proofs are given in the appendices.

2 Background

As any numerical integration technique, the method of Sobol' can be viewed as the particular combination of a design of experiments (DoE) — i.e. the location the information is collected — and an estimator — i.e. the way the collected information is processed. In this section, we first describe the designs of experiments of interest and we then come to the definitions of some currently used estimators in the method of Sobol'. At last, we introduce the concept of RLHS and we explain why this kind of DoE is of great importance in the issue of estimating first-order SIs.

From now on, we only consider uniformly distributed inputs, i.e. $X_1, \ldots, X_d$ are independent random variables uniformly distributed on $[0,1]$. It does not consist of a restrictive assumption since, thanks to the *inversion method* — see, e.g., Devroye (1986) — a model formulation $Y = f(X_1, \ldots X_d)$ where the X_is are arbitrarily distributed can always come down to the formulation $Y = f(U_1, \ldots, U_d)$ where the U_is are random variables uniformly distributed on $[0,1]$, defined by $U_i = F_i(X_i)$ with F_i the cumulative distribution function of X_i.

2.1 Designs of experiments

In the present paper a design of experiments refer to a finite subset of $[0,1]^d$. We do not consider deterministic constructions of DoEs but only random ones, i.e. as in a Monte Carlo method, DoEs consist of realizations of a set of random d-dimensional vectors.

For any non-empty subset $\mathfrak{u}$ of $\{1, \ldots, d\}$, there exist several estimators of the lower SI $\underline{S}_\mathfrak{u}$ and they all require n double, triple or more evaluations of the model (see Owen (2012a) for a recent survey). In the present paper, we focus on the most basic estimators that only require n double evaluations. In this case, the first of any double evaluation is a realization of the random variable $Y = f(X_1, \ldots, X_d)$, and the complementary evaluation is obtained from the first one by resampling its components indexed by the elements of $\mathfrak{u}^c$. In other words, the complementary evaluation is a realization of the random variable denoted by $Y_\mathfrak{u}$ and defined by

$$Y_\mathfrak{u} = f(\mathbf{X}_\mathfrak{u} : \mathbf{Z}_{\mathfrak{u}^c}) \tag{2}$$

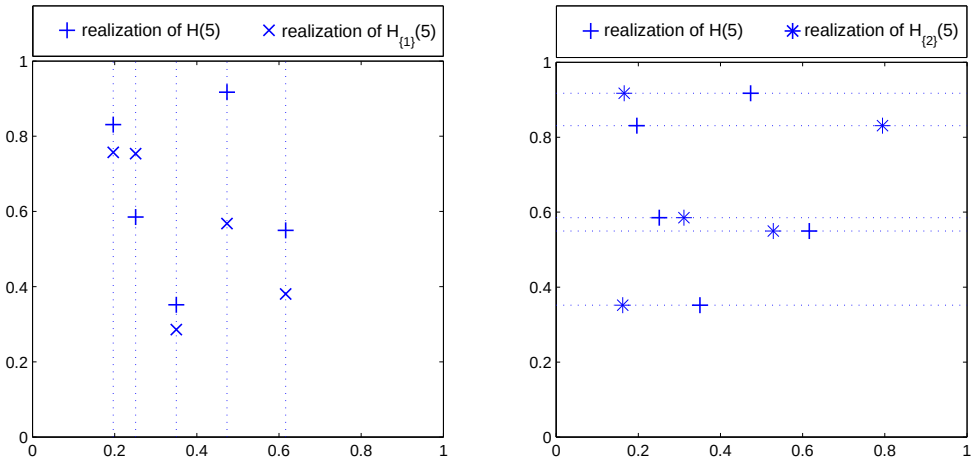
where $\mathbf{Z}$ is a d-dimendional vector uniformly distributed on $[0,1]^d$, and for all i in $\{1, \ldots, d\}$

$$(\mathbf{X}_\mathfrak{u} : \mathbf{Z}_{\mathfrak{u}^c})_i = \begin{cases} X_i & \text{if } i \in \mathfrak{u} \\ Z_i & \text{otherwise.} \end{cases} \tag{3}$$

Hence the definition of the design of experiments of size $N = 2n$ for estimating $\underline{S}_\mathfrak{u}$, denoted by $D_\mathfrak{u}(N)$, proceeds as follows. First let $(\mathbf{X}^j)_{j=1..n}$ and $(\mathbf{Z}^j)_{j=1..n}$ be independent replications of the random vectors $\mathbf{X}$ and $\mathbf{Z}$, respectively. Then denote the two halves of the main DoE by

$$\begin{aligned} H(n) &= \left\{\mathbf{X}^j,\ 1 \le j \le n\right\} \\ H_\mathfrak{u}(n) &= \left\{\mathbf{X}^j_\mathfrak{u} : \mathbf{Z}^j_{\mathfrak{u}^c},\ 1 \le j \le n\right\} \quad \text{(see the definition in Eq.(3))}, \end{aligned}$$

and define $D_\mathfrak{u}(N)$ to be the resulting union of both sets. Figure 1 shows illustrations of such DoEs for estimating first-order SIs in a 2-dimensional space. Note that in this figure $D_{\{1\}}(10)$ and $D_{\{2\}}(10)$ contain two points per level of the first and second axis, respectively; this consists of the main property — or constraint — of the DoE in the issue of estimating SIs using the method of Sobol'.



(a) realization of $D_{\{1\}}(10)$ (b) realization of $D_{\{2\}}(10)$

Figure 1: Designs of experiments of size $N = 10$ for estimating the first-order SIs $\underline{S}_1$ (subfigure (a)) and $\underline{S}_2$ (subfigure (b)).

2.2 Estimators

Using the previous notation, we now consider, for any j in $\{1, \ldots, n\}$, the output observations

$$Y^j = f(\mathbf{X}^j) \tag{4}$$

$$Y_{\mathrm{u}}^j = f(\mathbf{X}_{\mathrm{u}}^j : \mathbf{Z}_{\mathrm{u}^c}). \tag{5}$$

In the method of Sobol', the estimation of the index $\underline{S}_{\mathrm{u}}$ consists in applying a Monte Carlo (MC) method to both the numerator and denominator of

$$\underline{S}_{\mathrm{u}} = \frac{\mathrm{Var}\left[\mathbb{E}[Y|\mathbf{X}_{\mathrm{u}}]\right]}{\mathrm{Var}[Y]}$$

that can be rewritten, using the notation in Eq. (2), as

$$\underline{S}_{\mathrm{u}} = \frac{\mathrm{Cov}(Y, Y_{\mathrm{u}})}{\mathrm{Var}[Y]} = \frac{\mathbb{E}[YY_{\mathrm{u}}] - \mathbb{E}[Y]\mathbb{E}[Y_{\mathrm{u}}]}{\mathrm{Var}[Y]}. \tag{6}$$

As already mentioned in this section, several estimators have been introduced to perform this numerical integration. In the present paper, we only consider the natural estimator coming from (6), and an other one due to Monod et al. (2006). Using the notation in (4) and (5), they are defined by

$$\widetilde{\underline{S}}_{\mathrm{u},n} = \frac{\widetilde{\underline{\tau}}_{\mathrm{u},n}^2}{\widetilde{\sigma}_n^2} = \frac{\frac{1}{n}\sum_{j=1}^n Y^j Y_{\mathrm{u}}^j - \left(\frac{1}{n}\sum_{j=1}^n Y^j\right)\left(\frac{1}{n}\sum_{j=1}^n Y_{\mathrm{u}}^j\right)}{\frac{1}{n}\sum_{j=1}^n \left(Y^j\right)^2 - \left(\frac{1}{n}\sum_{j=1}^n Y^j\right)^2} \tag{7}$$

and

$$\widehat{\underline{S}}_{\mathrm{u},n} = \frac{\widehat{\underline{\tau}}_{\mathrm{u},n}^2}{\widehat{\sigma}_n^2} = \frac{\frac{1}{n}\sum_{j=1}^n Y^j Y_{\mathrm{u}}^j - \left(\frac{1}{2n}\sum_{j=1}^n Y^j + Y_{\mathrm{u}}^j\right)^2}{\frac{1}{2n}\sum_{j=1}^n \left(\left(Y^j\right)^2 + \left(Y_{\mathrm{u}}^j\right)^2\right) - \left(\frac{1}{2n}\sum_{j=1}^n Y^j + Y_{\mathrm{u}}^j\right)^2}, \tag{8}$$

respectively.

2.3 Motivation to use replicated Latin hypercube sampling

Following the previous description of both the designs of experiments and the estimators used in the method of Sobol', it is easy to understand that estimating all the first-order SIs using this technique requires $n(d+1)$ evaluations of the model, i.e. $d+1$ DoEs each containing n points. More generally, for any k in $\{1,\dots,d\}$, estimating all the k-th-order SIs basically requires $n\left(\binom{n}{k}+1\right)$ evaluations.

Some interesting combinatorial results have been introduced to weaken this defect. For instance, Saltelli (2002) shows how to estimate all the second-order SIs by using only $n(2d+2)$ evaluations. In a more recent work, Owen (2012b) deeply studies these combinatorial aspects in the issue of estimating SIs. However the quantities of interest in both these papers still require $O(d)$ evaluations of the model, and this dependency with respect to the dimension d appears as a major drawback in the sensitivity analysis of high-dimensional models potentially depending on more than one hundred input parameters.

In order to overcome this issue of dimensionality, we propose a new approach based on a non-common notion of replication. More precisely, for any DoE

$$D = \left\{\mathbf{x}^j = (x_1^j,\dots,x_d^j),\ 1 \leq j \leq d\right\},$$

we say that a DoE D' is replicated from D if there exist d independent random permutations of $\{1,\dots,n\}$ — i.e. d independent random variables uniformly distributed on the set of the permutations of $\{1,\dots,n\}$ — denoted by $\pi_1,\dots,\pi_d$, such that

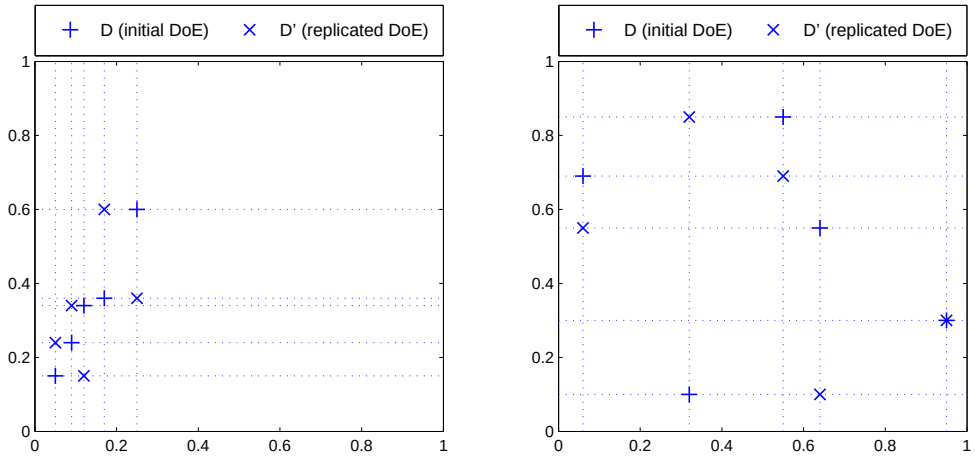
$$D' = \left\{\mathbf{x}'^j = (x_1^{\pi_1(j)},\dots,x_d^{\pi_d(j)}),\ 1 \leq j \leq n\right\}.$$

Note that this notation is clearly symmetric, so we can also say that D is replicated from D' or even that D and D' are replicated from each other. As a result, for any axis k in $\{1,\dots,d\}$, the union of a DoE D and one of its replicate D' contains two points per level of the k-th coordinate (see Figure 2(a)). Hence, this composite DoE fulfills the requirement for estimating any first-order SI using the method of Sobol', and we thus shows provide a DoE that allows to estimate all the first-order SIs using only $2n$ evaluations of the model.

However, as we can observe on Figure 2(a), replication can have a bad effect on the space-filling property of the composite DoE $D \cup D'$, and it could be desirable to take care to replicate only good space-filling DoEs. Then LHS appears as the most natural candidate, and this leads to the notion of RLHS (see Figure 2(b)) already used by McKay (1995) in the issue of estimating first-order Sobol' indices.

3 Theory

Both the estimators introduced in the previous section, $\widetilde{\underline{S}}_{\mathrm{u},n}$ and $\widehat{\underline{S}}_{\mathrm{u},n}$, have strong statistical properties. It is easy to prove that they are *strongly consistent* — i.e. they converge almost surely to the theoretical value $\underline{S}_{\mathrm{u}}$



(a) D: arbitrary DoE

(b) D: Latin hypercube

Figure 2: Composite DoE for estimating all the first-order SIs.

— and that the *biases* of both their numerator and denominator are $O(n)$:

$$\begin{aligned}\mathbb{E}\big[\widetilde{\tau}^2_{\mathrm{u},n}\big] &= \tau^2_{\mathrm{u}} - \frac{1}{n}\tau^2_{\mathrm{u}} \\ \mathbb{E}\big[\widetilde{\sigma}^2_{n}\big] &= \sigma^2 - \frac{1}{n}\sigma^2\end{aligned}$$

and, as shown by Owen (2012b),

$$\begin{aligned}\mathbb{E}\big[\widehat{\tau}^2_{\mathrm{u},n}\big] &= \tau^2_{\mathrm{u}} - \frac{1}{2n}(\sigma^2 + \tau^2_{\mathrm{u}}) \\ \mathbb{E}\big[\widehat{\sigma}^2_{n}\big] &= \sigma^2 - \frac{1}{2n}(\sigma^2 + \tau^2_{\mathrm{u}}).\end{aligned}$$

But the most important property, proved by Janon et al. (2012), is that they are *asymptotically normal*. In particular this allows to derive asymptotic confidence intervals which provide probabilistic bounds on the estimation error.

Now the question that naturally arises is

"Does the new approach based on RLHS we proposed for estimating all the first-order Sobol' indices still has these strong statistical properties?"

The answer is affirmative to the three previous points of interest and we prove it in this section. First we begin by stating important results of convergence on LHS — see Proposition 1 in Section 3.1 — and on RLHS — see Proposition 2 in Section 3.2 — and we then formulate the main result on the estimation of all the first-order SIs using RLHS in Theorem 1 in Section 3.2. In a last subsection, we address the issue of potential generalizations to OA-based LHS.

3.1 On Latin hypercube sampling

We now assume that the $\mathbf{X}^j$s and the $\mathbf{Z}^j$s are no longer independent replications of $\mathbf{X}$ and $\mathbf{Z}$, but we consider that $\{\mathbf{X}^j,\ 1 \leq j \leq n\}$ and $\{\mathbf{Z}^j,\ 1 \leq j \leq n\}$ are two independent Latin hypercubes of size n. We then prove

that both the estimators introduced in the previous section still have the statistical properties presented above. We first introduce the definition of a Latin hypercube:

Definition 1. *Let d and n in $\mathbb{N}^*$, and consider Π_n the set of all the permutations of $\{1,\dots,n\}$. We say that $(\mathbf{X}^j)_{j=1..n}$ is a Latin hypercube of size n in $[0,1]^d$ — and we denote $(\mathbf{X}^j)_j \sim \mathcal{LH}(n,d)$ — if for all $j \in \{1,\dots,n\}$,*

$$\mathbf{X}^j = \left(\frac{\pi_1(j) - U_{1,\pi_1(j)}}{n}, \dots, \frac{\pi_d(j) - U_{d,\pi_d(j)}}{n} \right) \tag{9}$$

where the π_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_n and $[0,1]$, respectively.

Now let $(\dot{\mathbf{X}}^j)_j \sim \mathcal{LH}(n,d)$ and $(\dot{\mathbf{Z}}^j)_j \sim \mathcal{LH}(n,d)$ in place of the independent replications $\mathbf{X}^j$s and $\mathbf{Z}^j$s, and for any $\mathfrak{u}$ in $\{1,\dots,d\}$ denote

$$\begin{aligned} \dot{Y}^j &= f(\dot{\mathbf{X}}^j) \\ \dot{Y}^j_{\mathfrak{u}} &= f(\dot{\mathbf{X}}^j_{\mathfrak{u}} : \dot{\mathbf{Z}}^j_{\mathfrak{u}^c}). \end{aligned}$$

The resulting estimators are now denoted $\widetilde{\underline{S}}^{LHS}_{\mathfrak{u},n} = \widetilde{\underline{\tau}}^{2,LHS}_{\mathfrak{u},n} / \widetilde{\sigma}^{2,LHS}_n$ and $\widehat{\underline{S}}^{LHS}_{\mathfrak{u},n} = \widehat{\underline{\tau}}^{2,LHS}_{\mathfrak{u},n} / \widehat{\sigma}^{2,LHS}_n$, respectively. Their statistical properties are gathered in the following result:

Proposition 1.

(i) *If f^4 is integrable then $\widetilde{\underline{S}}^{LHS}_{\mathfrak{u},n}$ et $\widehat{\underline{S}}^{LHS}_{\mathfrak{u},n}$ are strongly consistent.*

(ii) *If f^6 is integrable then $\sqrt{n}(\widetilde{\underline{S}}^{LHS}_{\mathfrak{u},n} - \underline{S}_{\mathfrak{u}})$ and $\sqrt{n}(\widehat{\underline{S}}^{LHS}_{\mathfrak{u},n} - \underline{S}_{\mathfrak{u}})$ converge in law to a zero-mean normal distribution with lower variance than the respective variance given in the central limit theorem (CLT) for the basic estimators $\widetilde{\underline{S}}_{\mathfrak{u},n}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}$.*

(iii) *We have*

$$\begin{aligned} \mathbb{E}\big[\widetilde{\underline{\tau}}^{2,LHS}_{\mathfrak{u},n}\big] &= \underline{\tau}^2_{\mathfrak{u}} + B_{n,1} \\ \mathbb{E}\big[\widetilde{\sigma}^{2,LHS}_n\big] &= \sigma^2 + B_{n,2} \\ \mathbb{E}\big[\widehat{\underline{\tau}}^{2,LHS}_{\mathfrak{u},n}\big] &= \underline{\tau}^2_{\mathfrak{u}} + B_{n,3} \\ \mathbb{E}\big[\widehat{\sigma}^{2,LHS}_n\big] &= \sigma^2 + B_{n,3} \end{aligned}$$

where

$$\begin{aligned} -\frac{1}{n-1}\underline{\tau}^2_{\mathfrak{u}} \le\ & B_{n,1} \le 0 \\ -\frac{1}{n-1}\sigma^2 \le\ & B_{n,2} \le 0 \\ -\frac{1}{2(n-1)}(\sigma^2 + \underline{\tau}^2_{\mathfrak{u}}) \le\ & B_{n,3} \le 0\ . \end{aligned}$$

Proof.

(i) This is a consequence of the strong law of large numbers for LHS given in Theorem 3 in Loh (1996).

(ii) The proof consists in translating the original proof, given for simple random sampling — see Proposition 2.2 in Janon et al. (2012) — for LHS. The proof is given in Appendix A.

(iii) See Appendix A. □

Remark 1. *Due to their intricate structure, the biases of the estimators* $\widetilde{\underline{\tau}}^{2,LHS}_{\mathfrak{u},n}$, $\widetilde{\sigma}^{2,LHS}_{n}$, $\widehat{\underline{\tau}}^{2,LHS}_{\mathfrak{u},n}$ *and* $\widehat{\sigma}^{2,LHS}_{n}$ *can't be easily reduced. Nevertheless we can note that these biases are asymptotically negligible, with a rate of convergence in* $O(n^{-1})$ *larger than the rate of convergence of the estimators — to their theoretical values — themselves, which is in* $O(n^{-1/2})$.

To conclude, we have proven that the estimation of SIs combining MC estimators and Latin hypercube sampling has strong statistical properties. In particular, estimators in such a method have similar — and potentially smaller — bias than the classic method based on simple random sampling and asymptotically smaller variance.

3.2 On replicated Latin hypercube sampling

We now come to alternative estimators based on RLHS. First in Proposition 2, we present a technical result showing that such estimators still have the strong statistical properties of the former estimators. We then prove that only two RLHS are necessary to estimate all the first-order Sobol' indices and that the estimators defined in this efficient strategy have the same statistical properties as in Proposition 2.

We begin with the definition of a replicated Latin hypercube:

Definition 2. *Let* d *and* n *in* $\mathbb{N}^*$, *and consider* Π_n *the set of all the permutations of* $\{1,\dots,n\}$. *We say that* $(\mathbf{X}^j)_{j=1..n}$ *and* $(\mathbf{X}'^j)_{j=1..n}$ *are two replicated Latin hypercubes of size* n *in* $[0,1]^d$ — *and we denote* $(\mathbf{X}^j,\mathbf{X}'^j)_j \sim \mathcal{RLH}(n,d)$ — *if for all* $j \in \{1,\dots,n\}$,

$$\mathbf{X}^j = \left(\frac{\pi_1(j)-U_{1,\pi_1(j)}}{n},\dots,\frac{\pi_d(j)-U_{d,\pi_d(j)}}{n}\right)$$

and

$$\mathbf{X}'^j = \left(\frac{\pi'_1(j)-U_{1,\pi'_1(j)}}{n},\dots,\frac{\pi'_d(j)-U_{d,\pi'_d(j)}}{n}\right)$$

where the π_i*'s, the* π'_i*'s and the* $U_{i,j}$*'s are independent random variables uniformly distributed on* Π_n, Π_n *and* $[0,1]$, *respectively.*

Now let $(\dot{\mathbf{X}}^j)_j \sim \mathcal{LH}(n,d)$ and $(\ddot{\mathbf{Z}}^j,\ddot{\mathbf{Z}}'^j)_j \sim \mathcal{RLH}(n,d)$, and for any $\mathfrak{u}$ in $\{1,\dots,d\}$ denote

$$\ddot{Y}^j = f(\dot{\mathbf{X}}^j_{\mathfrak{u}} : \ddot{\mathbf{Z}}^j_{\mathfrak{u}^c}) \tag{10}$$

$$\ddot{Y}^j_{\mathfrak{u}} = f(\dot{\mathbf{X}}^j_{\mathfrak{u}} : \ddot{\mathbf{Z}}'^j_{\mathfrak{u}^c}). \tag{11}$$

Note that $\ddot{Y}^j$ now depends on $\mathfrak{u}$, but the essential constraint stating that the random vectors $\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j$ and $\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j}$ have the same components in their i-th coordinate, for all $i \in \mathfrak{u}$ is still checked. However for convenience, we keep the notation unchanged.

The resulting estimators are now denoted by $\widetilde{\underline{S}}_{\mathfrak{u},n}^{RLHS} = \widetilde{\underline{\tau}}_{\mathfrak{u},n}^{2,RLHS}/\widetilde{\sigma}_n^{2,RLHS}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}^{RLHS} = \widehat{\underline{\tau}}_{\mathfrak{u},n}^{2,RLHS}/\widehat{\sigma}_n^{2,RLHS}$, respectively. Note that estimators of SIs based on r replicated Latin hypercubes have already been introduced by McKay (1995) — see also the summarized presentation by Saltelli et al. (2000) — but the main drawback of these estimators is that they converge to their corresponding analytical values only as r tends to $+\infty$. The statistical properties of $\widetilde{\underline{S}}_{\mathfrak{u}}^{RLHS}$ and $\widehat{\underline{S}}_{\mathfrak{u}}^{RLHS}$ are gathered in the following result:

Proposition 2.

(i) *If f^4 is integrable then $\widetilde{\underline{S}}_{\mathfrak{u},n}^{RLHS}$ and $\widehat{\underline{S}}_{\mathfrak{u},n}^{RLHS}$ are strongly consistent.*

(ii) *If f^6 is integrable then $\sqrt{n}(\widetilde{\underline{S}}_{\mathfrak{u},n}^{RLHS} - \underline{S}_{\mathfrak{u}})$ and $\sqrt{n}(\widehat{\underline{S}}_{\mathfrak{u},n}^{RLHS} - \underline{S}_{\mathfrak{u}})$ converge in law to a zero-mean normal distribution with the same respective variance given in CLT for the estimators $\widetilde{\underline{S}}_{\mathfrak{u},n}^{LHS}$ and $\widetilde{\underline{S}}_{\mathfrak{u},n}^{LHS}$.*

(iii) *We have*

$$\begin{aligned}
\mathbb{E}\big[\widetilde{\underline{\tau}}_{\mathfrak{u},n}^{2,RLHS}\big] &= \underline{\tau}_{\mathfrak{u}}^2 - \frac{1}{n}\underline{\tau}_{\mathfrak{u}}^2 + B_{n,1} + B_{|\mathfrak{u}|,n} \\
\mathbb{E}\big[\widetilde{\sigma}_n^{2,RLHS}\big] &= \sigma^2 + B_{n,3} \\
\mathbb{E}\big[\widehat{\underline{\tau}}_{\mathfrak{u},n}^{2,RLHS}\big] &= \underline{\tau}_{\mathfrak{u}}^2 - \frac{1}{2n}\underline{\tau}_{\mathfrak{u}}^2 + B_{n,1} + B_{n,2} + B_{|\mathfrak{u}|,n} \\
\mathbb{E}\big[\widehat{\sigma}_n^{2,RLHS}\big] &= \sigma^2 - \frac{1}{2n}\underline{\tau}_{\mathfrak{u}}^2 + B_{n,1} + B_{n,2}
\end{aligned}$$

where

$$\begin{aligned}
|B_{n,1}| &\leq \left(\frac{d+1}{n} + 2\right)\left(\frac{d+1}{n}\right)\mathbb{E}[Y^2] \\
|B_{n,2}| &\leq \frac{\sigma^2}{2n} \\
-\frac{1}{n-1}\sigma^2 \leq \quad B_{n,3} &\leq 0 \\
|B_{|\mathfrak{u}|,n}| &\leq \left(\frac{d-|\mathfrak{u}|+1}{n} + 2\right)\left(\frac{d-|\mathfrak{u}|+1}{n-1}\right)\mathbb{E}[Y^2] \ .
\end{aligned}$$

Proof. See Appendix B. □

Using this technical result, it is possible to estimate all the first-order Sobol' indices with only two RLH and in addition the new estimators inherit the strong statistical properties stated in Proposition 2. The main idea consists in defining, for any $i \in \{1, \ldots, d\}$, the two samples Y^j and $Y_{\{i\}}^j$ — necessary to estimate $\underline{S}_{\{i\}}$ — by only considering two RLH. To this end, we first consider $(\ddot{\mathbf{W}}^j, \ddot{\mathbf{W}}^{'j})_j \sim \mathcal{RLH}(n, d)$ defined by

$$\ddot{\mathbf{W}}^j = \left(\frac{\pi_1(j) - U_{1,\pi_1(j)}}{n}, \ldots, \frac{\pi_d(j) - U_{d,\pi_d(j)}}{n}\right)$$

and

$$\ddot{\mathbf{W}}^{'j} = \left(\frac{\pi_1'(j) - U_{1,\pi_1'(j)}}{n}, \ldots, \frac{\pi_d'(j) - U_{d,\pi_d'(j)}}{n}\right).$$

Secondly, let $\pi \in \Pi_n$ be a random permutation independent from the π_is and the π'_is, and for any $i \in \{1,\dots,d\}$ and $j \in \{1,\dots,n\}$ consider

$$Y^j = f(\ddot{\mathbf{W}}^{\pi_i^{-1}\circ\pi(j)}) \tag{12}$$

$$Y^j_{\{i\}} = f(\ddot{\mathbf{W}}^{'\pi_i^{'-1}\circ\pi(j)}). \tag{13}$$

Then we have the following result.

Theorem 1. *Consider $(\ddot{\mathbf{W}}^j, \ddot{\mathbf{W}}^{'j})_j \sim \mathcal{RLH}(n,d)$ and denote $D(n)$ the DoE defined as the union of both these replicated Latin hypercubes:*

$$D(n) = \{\ddot{\mathbf{W}}^j,\ 1 \le j \le n\} \cup \{\ddot{\mathbf{W}}^{'j},\ 1 \le j \le n\}.$$

For any $i \in \{1,\dots,d\}$, consider the evaluations of the function f on $D(n)$ denoted by $(Y^j)_{j=1..n}$ and $(Y^j_{\{i\}})_{j=1..n}$ as defined in (12) and (13). Then the estimator $\underline{\widetilde{S}}_{\{i\}}$ as defined in (7) — (resp. $\underline{\widehat{S}}_{\{i\}}$ as defined in (8)) — built with Y^j and $Y^j_{\{i\}}$ is strongly consistent, asymptotically normal and has numerator and denominator with bias as stated in (iii) of Proposition 2.

Proof. We have $\ddot{\mathbf{W}}^{\pi_i^{-1}\circ\pi(j)} = \dot{\mathbf{X}}^j_{\{i\}} : \ddot{\mathbf{Z}}^j_{\{i\}^c}$ and $\ddot{\mathbf{W}}^{'\pi_i^{'-1}\circ\pi(j)} = \dot{\mathbf{X}}^j_{\{i\}} : \ddot{\mathbf{Z}}^{'j}_{\{i\}^c}$ where

$$\begin{aligned}
\dot{\mathbf{X}}^j_{\{i\}} &= \frac{\pi(j) - U_{i,\pi(j)}}{n} \\
\ddot{\mathbf{Z}}^j_{\{i\}^c} &= \left(\frac{\breve{\pi}_1(j) - U_{1,\breve{\pi}_1(j)}}{n}, \dots, \frac{\breve{\pi}_{i-1}(j) - U_{i-1,\breve{\pi}_{i-1}(j)}}{n}, \frac{\breve{\pi}_{i+1}(j) - U_{i+1,\breve{\pi}_{i+1}(j)}}{n}, \dots, \frac{\breve{\pi}_d(j) - U_{d,\breve{\pi}_d(j)}}{n}\right) \\
\ddot{\mathbf{Z}}^{'j}_{\{i\}^c} &= \left(\frac{\breve{\pi}'_1(j) - U_{1,\breve{\pi}'_1(j)}}{n}, \dots, \frac{\breve{\pi}'_{i-1}(j) - U_{i-1,\breve{\pi}'_{i-1}(j)}}{n}, \frac{\breve{\pi}'_{i+1}(j) - U_{i+1,\breve{\pi}'_{i+1}(j)}}{n}, \dots, \frac{\breve{\pi}'_d(j) - U_{d,\breve{\pi}'_d(j)}}{n}\right)
\end{aligned}$$

with $\breve{\pi}_k = \pi_k \circ \pi_i^{-1} \circ \pi$ and $\breve{\pi}'_k = \pi'_k \circ \pi_i^{'-1} \circ \pi$. Then note that π, the $\breve{\pi}_k$s and the $\breve{\pi}'_k$s are independent random permutations, and deduce that Y^j and $Y^j_{\{i\}}$ given in (12) and (13) are defined as in (10) and (11). Thus Proposition 2 applies and the conclusion follows. □

3.3 Potential generalization to orthogonal array-based Latin hypercubes

The main question that arises given the result stated in Theorem 1 is: can we estimate all the k-th order Sobol' indices with only two RLHS? On the one hand, the most straightforward answer is clearly negative since RLHS do not have the required strcuture to handle these higher-order Sobol' indices. On the other hand, we have to observe that such a well-suited structure can be built by using orthogonal arrays. So we first begin with the definition of an orthogonal array (OA):

Definition 3. *An OA in dimension d, with q levels, strength $t \le d$ and index λ is a matrix with $n = \lambda q^t$ rows and d columns such that in every n-by-t submatrix each of the q^t possible rows — i.e. the distinct t-tuples $(l_1,\dots,l_t)$ where the l_i's take their values in the set of the q levels — occurs exactly the same number λ of times.*

We now recall the definition of OA-based Latin hypercubes — see Owen (1992) — and introduce the general notion of replicated OA-based Latin hypercubes.

Definition 4. *Let $(A_i^j)_{i=1..d,j=1..n}$ be an OA in dimension d, with n points and q levels in $\{1,\dots,q\}$, and consider Π_q the set of all the permutations of $\{1,\dots,q\}$. We say that $(\mathbf{X}^j)_{j=1..n}$ is a Latin hypercube based on the OA $(\mathbf{A}^j)_{j=1..n}$ — and we denote $(\mathbf{X}^j)_j \sim \mathcal{LH}\big((\mathbf{A}^j)_j\big)$ — if for all $j \in \{1,\dots,n\}$,*

$$\mathbf{X}^j = \left(\frac{\pi_1(A_1^j) - U_{1,\pi_1(A_1^j)}}{q}, \dots, \frac{\pi_d(A_d^j) - U_{d,\pi_d(A_d^j)}}{q}\right)$$

where the π_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_q and $[0,1]$, respectively.

Definition 5. *Let $(A_i^j)_{i=1..d,j=1..n}$ an OA in dimension d, with n points and q levels in $\{1,\dots,q\}$, and consider Π_q the set of all the permutations of $\{1,\dots,q\}$. We say that $(\mathbf{X}^j)_{j=1..n}$ and $(\mathbf{X}'^j)_{j=1..n}$ are two replicated Latin hypercubes based on the orthogonal array $(\mathbf{A}^j)_{j=1..n}$ — and we denote $(\mathbf{X}^j, \mathbf{X}'^j)_j \sim \mathcal{RLH}\big((\mathbf{A}^j)_j\big)$ — if for all $j \in \{1,\dots,n\}$,*

$$\mathbf{X}^j = \left(\frac{\pi_1(A_1^j) - U_{1,\pi_1(A_1^j)}}{q}, \dots, \frac{\pi_d(A_d^j) - U_{d,\pi_d(A_d^j)}}{q}\right)$$

and

$$\mathbf{X}'^j = \left(\frac{\pi'_1(A_1^j) - U_{1,\pi'_1(A_1^j)}}{q}, \dots, \frac{\pi'_d(A_d^j) - U_{d,\pi'_d(A_d^j)}}{q}\right)$$

where the π_i's, the π'_i's and the $U_{i,j}$'s are independent random variables uniformly distributed on Π_q, Π_q and $[0,1]$, respectively.

It is interesting to note that in the particular case of the OA $(\mathbf{A}^j)_{j=1}$ with strength 1 and index unity defined by

$$\forall i \in \{1,\dots d\},\ \forall j \in \{1,\dots n\},\quad A_i^j = j,$$

these definitions are exactly Definitions 1 and 2.

Now consider $(\ddot{\mathbf{Z}}^j, \ddot{\mathbf{Z}}'^j)_j \sim \mathcal{RLH}((\mathbf{A}^j)_j)$ where $(A_i^j)_{i=1..d,j=1..n}$ is an OA in dimension d, with n points, of strength t and index unity, and denote $D(n)$ the DoE defined as the union of both these OA-based replicated Latin hypercubes:

$$D(n) = \{\ddot{\mathbf{Z}}^j,\ 1 \le j \le n\} \cup \{\ddot{\mathbf{Z}}'^j,\ 1 \le j \le n\}.$$

Then thanks to the definition of OA-based RLHS, it is easy to note that for any t indices, $1 \le i_1 < \dots < i_t \le d$, for all $\mathbf{W} \in D(n)$, there exists a unique element $\mathbf{W}'$ distinct from $\mathbf{W}$ such that $W_{i_m} = W'_{i_m}$, $1 \le m \le t$. This property allows to estimate all the t-th order Sobol' indices by using only two the DoE $D(n)$.

Remark 2. *Theoretical properties of the estimators for this generalization remain open issues and will consist of a further work. The first step for strong consistency will be to state a strong law of large numbers for OA-based Latin hypercubes with strength $t > 1$ since, as far as we know, such a result does not exist. Asymptotic normality has already been proved for OA-based Latin hypercube with strength $t = 2$ under smoothness conditions*

— *see* Loh (2008) — *but it is not sufficient to conclude in the case of replicated OA-based Latin hypercubes since formulas as in* (105) *and* (106) *are necessary. As for the biases of the estimators, it will be necessary to study covariances in OA-based Latin hypercubes with strength* $t > 1$ *in order to state formulas as in* (105) *and* (106) *as well.*

4 Numerical illustrations

4.1 Application to the Ishigami function

4.1.1 Main experiment

In this section, we apply the new method proposed in Section 4 to the Ishigami function (see Ishigami and Homma (1990)):

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

where the X_i's are independent random variable uniformly distributed on $[-\pi, \pi]$. Analytical values of SIs of this model are

$$\underline{S}_1 = 0.3139, \quad \underline{S}_2 = 0.4424, \quad \underline{S}_3 = 0, \quad \underline{S}_{12} = 0.7563, \quad \underline{S}_{23} = 0.4424, \quad \underline{S}_{13} = 0.5575 \text{ and } \underline{S}_{123} = 1.$$

We are interested in comparing the new method, with the classic one based on crude MC method and which need $d+1$ samples to estimate all the first-order SIs, and $2d+2$ samples to estimate all the second-order SIs (see Saltelli (2002)). Here, both methods are compared at the same sample size n in order to investigate the estimators themselves, but keep in mind that the new method is definitely more efficient since only two samples are needed to estimate all the first-order SIs or all the second-order Sobol' indices. In the experiment, we focus on the empirical coverage — i.e. the empirical proportion of confidence interval containing the analytical value of the SI — of both estimators at different sample size between 10^2 and 10^5, and for $r = 100000$ replicates. We first investigate estimators $\widehat{\underline{S}}_{\{i\},n}$ and $\widehat{\underline{S}}^{RLHS}_{\{i\},n}$, $i \in \{1, \ldots, d\}$ and in both cases, we provide asymptotic confidence intervals from the estimation of the asymptotic variance given in Janon et al. (2012) (see end of the proof of Prop. 2.2). Indeed, as we know that this asymptotic variance is:

$$\sigma^2_{IID,\mathfrak{u}} = \frac{\mathrm{Var}\big[(Y - \mathbb{E}[Y])(Y_{\mathfrak{u}} - \mathbb{E}[Y]) - \underline{S}_{\mathfrak{u}}/2\big((Y - \mathbb{E}[Y])^2\big)\big(Y_{\mathfrak{u}} - \mathbb{E}[Y]\big)\big]}{\mathrm{Var}[Y]^2} \geq \sigma^2_{RLHS,\mathfrak{u}}, \tag{14}$$

we can provide an estimator of the asymptotic confidence interval for the classic method

$$I_{IID,\mathfrak{u},\alpha} = \Big[\underline{S}_{\mathfrak{u}} - \frac{\sigma^2_{IID,\mathfrak{u}}\, u_{\alpha/2}}{\sqrt{n}}, \underline{S}_{\mathfrak{u}} + \frac{\sigma^2_{IID,\mathfrak{u}}\, u_{\alpha/2}}{\sqrt{n}}\Big]$$

and an other one for the new method

$$I_{RLHS,\mathfrak{u},\alpha} = \Big[\underline{S}_{\mathfrak{u}} - \frac{\sigma^2_{RLHS,\mathfrak{u}}\, u_{\alpha/2}}{\sqrt{n}}, \underline{S}_{\mathfrak{u}} + \frac{\sigma^2_{RLHS,\mathfrak{u}}\, u_{\alpha/2}}{\sqrt{n}}\Big]$$

where $u_{\alpha/2}$ is the normal quantile at the significance level α. By using the estimator of the asymptotic variance given in (14) in both cases, the confidence interval lengths of the classic and the new estimators are the same. More specifically, the estimated length of the new estimator is greater or equal than its optimal value. Thus the asymptotic value of the empirical coverage of the new method is greater or equal than the expected one. However at the moment, we do not know how to estimate correctly $\sigma^2_{RLHS,\mathrm{u}}$ because of its singular expression (see Proof of (ii) in Proposition 1 in Section 3.2). We just say few words about it in the next subsection and more fundamentally, it should consist of a further work.

We also investigate estimators $\widehat{\underline{S}}_{\{i,j\},n}$ and $\widehat{\underline{S}}^{\text{OA2-RLHS}}_{\{i,j\},n}$, $i \neq j \in \{1,\dots,d\}$, where the notation OA2-RLHS refers to the generalization to replicated latin hypercube based on OA of strength 2 presented in Section 4.4. In this case, we conjecture that the Central Limit theorem established in (ii) in Proposition 2 is also true under some smoothness assumption — note that, here, Ishigami function is $\mathcal{C}^\infty$. Results are gathered in Figures 3 to 6. For the second-order SIs, we can observe that the bivariate stratification has a bad effect on the new

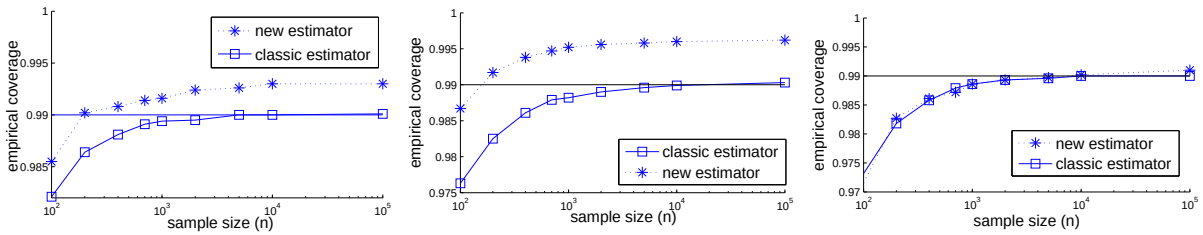


Figure 3: Empirical coverage of confidence intervals for $\underline{S}_1$ (left), $\underline{S}_2$ (center) and $\underline{S}_3$ (right).

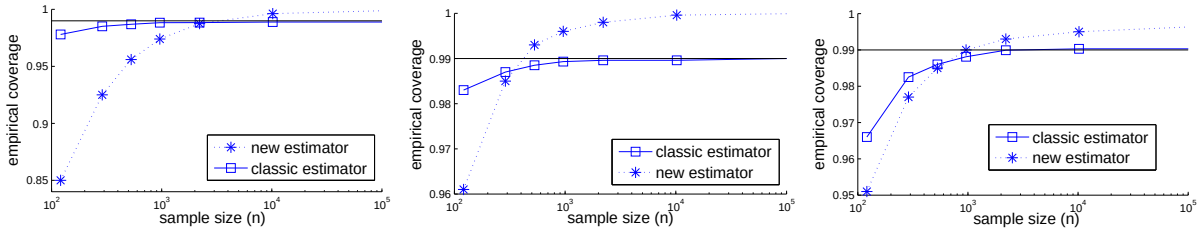


Figure 4: Empirical coverage of confidence intervals for $\underline{S}_{12}$ (left), $\underline{S}_{13}$ (center) and $\underline{S}_{23}$ (right).

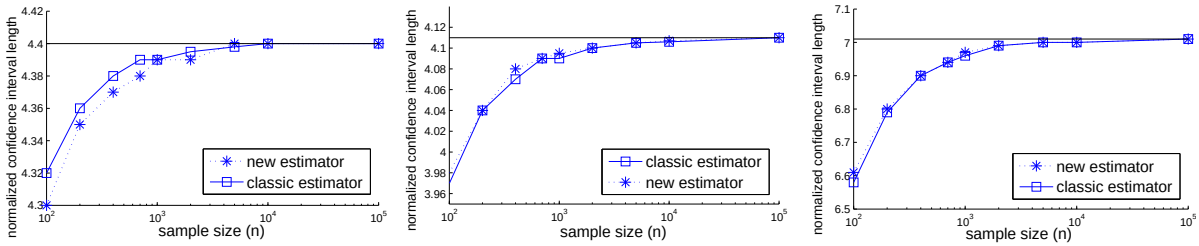


Figure 5: Normalized ($\times\sqrt{n}$) length of the empirical interval for $\underline{S}_1$ (left), $\underline{S}_2$ (center) and $\underline{S}_3$ (right).

estimator at very low sample size, but we can notice its good properties as the number of simulations increases.

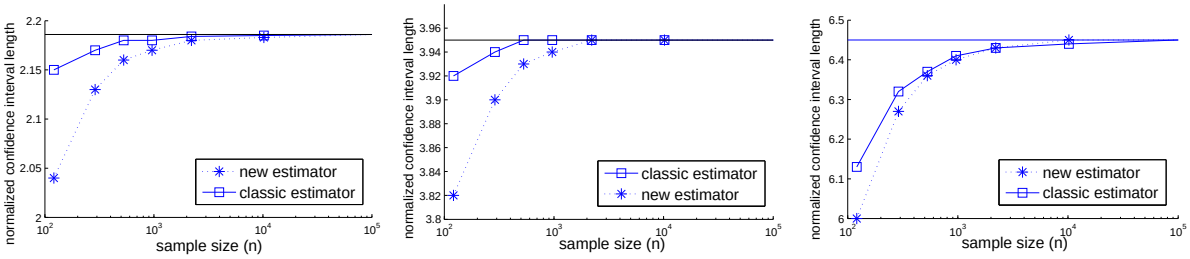


Figure 6: Normalized ($\times\sqrt{n}$) length of the empirical interval for $\underline{S}_{12}$ (left), $\underline{S}_{13}$ (center) and $\underline{S}_{23}$ (right).

	$\underline{S}_1$	$\underline{S}_2$	$\underline{S}_{12}$	$\underline{S}_{13}$	$\underline{S}_{23}$
estimated lengths using (14)	4.40	4.15	2.19	3.95	6.45
right lengths	3.96	3.28	1.53	2.37	5.16

Table 1: Comparison between confidence interval lengths estimated using (14) and the right lengths for $\underline{S}_1$, $\underline{S}_2$, $\underline{S}_{12}$, $\underline{S}_{13}$ and $\underline{S}_{23}$

4.1.2 Remark on the confidence interval length of the new estimator

Concerning the estimation of the right confidence interval length of the new estimators, note that if the asymptotic empirical coverage — estimated using Formula (14) — is $1-\alpha'$ instead of the expected value $1-\alpha$, then it means that the true asymptotic confidence interval should be $u_{\alpha/2}/u_{\alpha'/2}$ time as long, where $u_.$ denote the normal quantiles. More specifically in our first application, we obtain in this way the true asymptotic normalized ($\times\sqrt{n}$) confidence interval length of $\underline{S}_1$, $\underline{S}_2$, $\underline{S}_{12}$, $\underline{S}_{13}$ and $\underline{S}_{23}$; they are gathered in Table 1. Moreover considering these right normalized confidence interval lengths, we can observe on Figures 7 and 8 that the empirical coverage of the new estimator converges to the expected level 0.99 as n increases, and so we confirm the reliability of the empirical confidence intervals constructed with the true asymptotic length. Unfortunately, evaluating the true asymptotic confidence interval length is infeasible in practice since

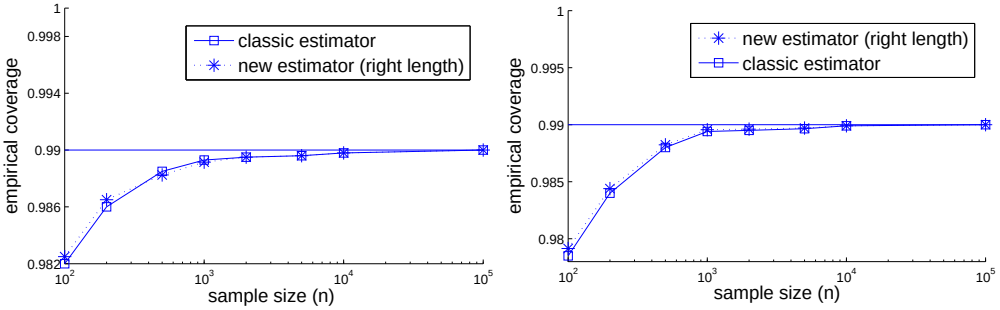


Figure 7: Empirical coverage of confidence intervals for $\underline{S}_1$ (left) and $\underline{S}_2$ (right).

it requires a lot of replications to estimate the empirical coverage. So the issue related to the construction of optimal confidence intervals remains open.

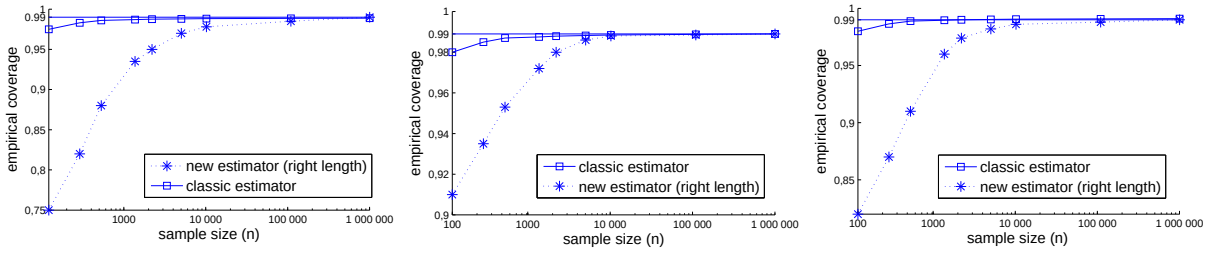


Figure 8: Empirical coverage of confidence intervals for $\underline{S}_{12}$ (left), $\underline{S}_{13}$ (center) and $\underline{S}_{23}$ (right).

4.2 Comparison between the new method and the quasi-Monte Carlo estimation

We now come to the comparison between the new method based on RLHS and the method of Sobol' performed with Sobol' sequences, also known as LP_τ sequences and that consist of (t, s)-sequences in base 2 — see, e.g., Niederreiter (1992) for (t, s)-sequences and Saltelli et al. (2010) for the application to the method of Sobol'. The numerical application we propose consists in comparing the estimation error — the mean squared error — in the computation of all the first-order Sobol' indices $\underline{S}_i$ between these two approaches using exactly the same total number of model evaluations. Low discrepancy sequences are known for their efficiency but keep in mind that the total number of model evaluations required to estimate all the first-order Sobol' indices is $(d+1)n$ when using (t, s)-sequences while only $2n$ are necessary when using RLHS. The analytical test-case is a multiplicative function known as the g-function — see Saltelli and Sobol' (1995). So we consider the model

$$Y = f_1(X_1) \times \cdots \times f_d(X_d)$$

where the X_is are independent random variables uniformly distributed on $[0, 1]$ and for any $i \in \{1, \ldots, d\}$

$$f_i(X_i) = \frac{|4X_i - 2| + a_i}{1 + a_i}, \quad a_i \geq 0.$$

The numerical test is divided into three cases:

Case i) in dimension $d = 3$ with g-function parameters $\mathbf{a} = (0, 1, 9)$

Case ii) in dimension $d = 12$ with g-function parameters $\mathbf{a} = (0, 0, 0, 0, 1, 1, 1, 1, 9, 9, 9, 9)$

Case iii) in dimension $d = 24$ with g-function parameters $\mathbf{a} = (\underbrace{0, \ldots, 0}_{8 \text{ times}}, \underbrace{1, \ldots, 1}_{8 \text{ times}}, \underbrace{9, \ldots, 9}_{8 \text{ times}})$.

The theoretical values of the Sobol' indices are the following:

Case i) $\underline{S}_1 = 0.742$, $\underline{S}_2 = 0.185$, $\underline{S}_3 = 0.007$

Case ii) $\underline{S}_1 = \cdots = \underline{S}_4 = 0.098$, $\underline{S}_5 = \cdots = \underline{S}_8 = 0.024$, $\underline{S}_9 = \cdots = \underline{S}_{12} = 0.001$,

Case iii) $\underline{S}_1 = \cdots = \underline{S}_8 = 0.018$, $\underline{S}_9 = \cdots = \underline{S}_{16} = 0.004$, $\underline{S}_{17} = \cdots = \underline{S}_{24} = 10^{-4}$.

The computation are performed at different total number of runs:

Case i) $N = 64, 512, 4096, 32768, 262144, 2097152$

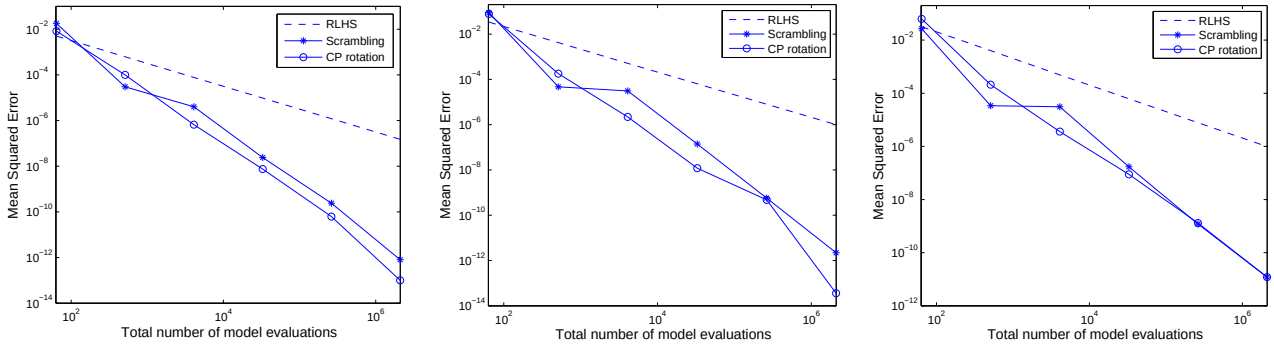
Case ii) $N = 208, 1664, 13312, 106496, 851968, 6815744$

Case iii) $N = 200, 1600, 12800, 102400, 819200, 6553600$.

In this numerical experiment, as we measure the error in term of mean squared error, we consider randomized Sobol' sequences. More precisely we use both the following well-known method of randomization:

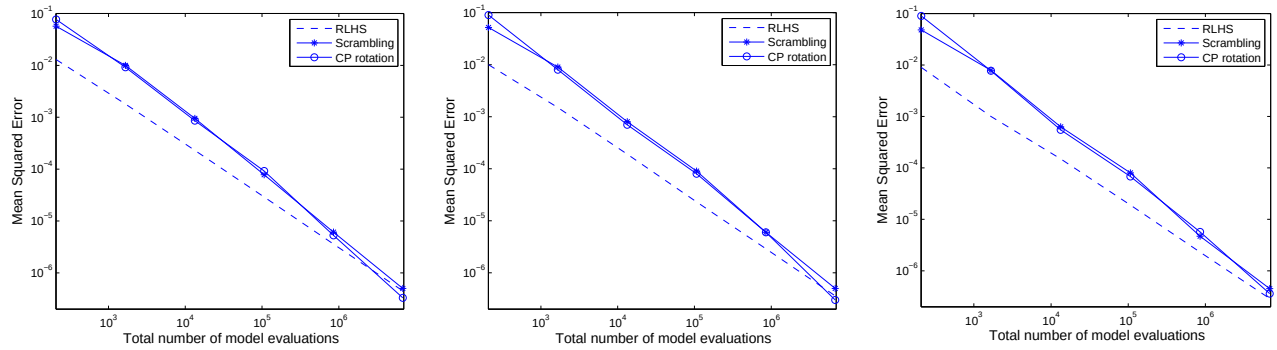
1. Cranley-Patterson rotation, that consists in adding a random vector to all the points of a DoE — where the addition is the componentwise addition modulo 1.
2. Owen's scrambling, that essentially consists in randomly permuting the levels of a (t, s)-sequence keeping the low discrepancy structure unchanged — see Owen (1995, 1997a,b).

The results are gathered in Figures (9–11).



(a) Estimation error of $\underline{S}_1$ (b) Estimation error of $\underline{S}_2$ (c) Estimation error of $\underline{S}_3$

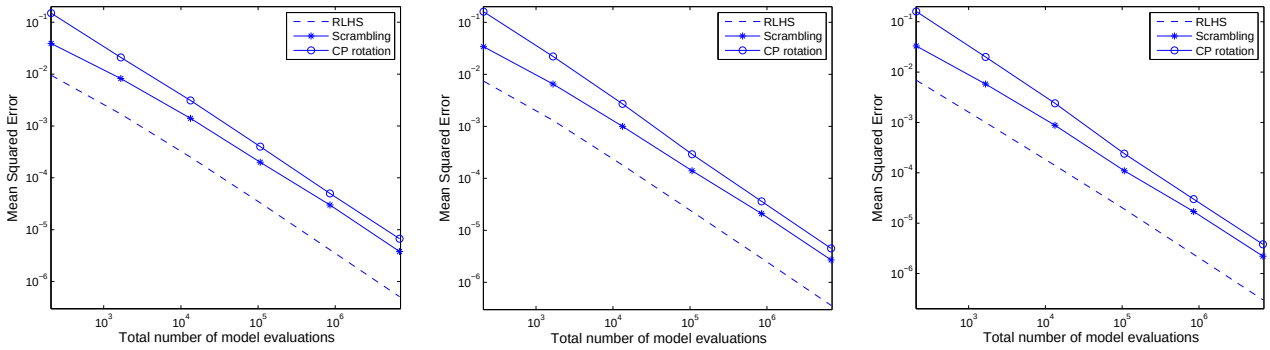
Figure 9: Plots of the estimation error of Sobol' indices for a g-function in dimension 3 with parameters $(0, 1, 9)$. NB: CP rotation stands for Cranley-Patterson rotation.



(a) Estimation error of $\underline{S}_1, \ldots, \underline{S}_4$ (b) Estimation error of $\underline{S}_5, \ldots, \underline{S}_8$ (c) Estimation error of $\underline{S}_9, \ldots, \underline{S}_{12}$

Figure 10: Plots of the estimation error of Sobol' indices for a g-function in dimension 12 with parameters $(0, 0, 0, 0, 1, 1, 1, 1, 9, 9, 9, 9)$. NB: CP rotation stands for Cranley-Patterson rotation.

First we can observe that the mean squared error (MSE) of all the estimates computed by using RLHS decreases with a rate of convergence of $O(n^{-1})$ in the three test-cases, while the rate of convergence of the low



(a) Estimation error of $\underline{S}_1, \ldots, \underline{S}_8$ (b) Estimation error of $\underline{S}_9, \ldots, \underline{S}_{16}$ (c) Estimation error of $\underline{S}_{17}, \ldots, \underline{S}_{24}$

Figure 11: Plots of the estimation error of Sobol' indices for a g-function in dimension 24 with parameters $(0, \ldots, 0, 1, \ldots, 1, 9, \ldots, 9)$. NB: CP rotation stands for Cranley-Patterson rotation.

discrepancy method — with scrambling as well as Cranley-Patterson rotation — is $O(n^{-2})$ in dimension 3 but only $O(n^{-1})$ in dimension 24. Second note that MSEs of the estimates computed by using RLHS at the lowest total number of model evaluations keep in the same range of values, while initial MSEs in low discrepancy methods become worse as the dimension increases. Consequently, the use of Sobol' sequences clearly leads to better results in dimension 3, but appears as a poor choice in the other cases. Indeed in dimension 12, we can observe that Sobol' sequences provide better results than RLHS only asymptotically — we can see in Figure 10 that at least 10^7 model evaluations are necessary — and in dimension 24, MSEs computed by using RLHS are always a factor 10 or 15 better than those of low discrepancy methods.

4.3 Application to a marine ecosystem simulator

We now illustrate the new method to a one-dimensional coupled hydrodynamical– biological model developed and applied to the Ligurian Sea (northwestern Mediterranean). This ecosystem simulator, MODèle d'ÉCOsystème du GHER et du LOBEPM[1] (MODECOGeL), combines a 1D (vertical) version of the 3D GHER model which takes into account momentum and heat surface fluxes computed from a real meteorological data set, and a biogeochemical model defined by a nitrogen cycle of 12 biological state variables (see Figure 12) controlled by 87 input parameters, see Lacroix and Nival (1998). Here we focus on the chlorophyll-a concentration which is defined as a function of time and depth

$$\texttt{chla}(t,z) = 1.59 * \big(\texttt{pp}(t,z) + \texttt{np}(t,z) + \texttt{mp}(t,z)\big)$$

where `pp`, `np` and `mp` are the phyto-, nano- and microphytoplankton biomasses, respectively. The behavior of these three state variables are modeled by the following reaction-diffusion and reaction-advection-diffusion

[1]GHER: GeoHydrodynamics and Environment Research, Université de Liège, Belgium. LOBEPM: Laboratoire d'Océanologie Biologique et d'Écologie du Plancton Marin, Université Pierre et Marie Curie, France

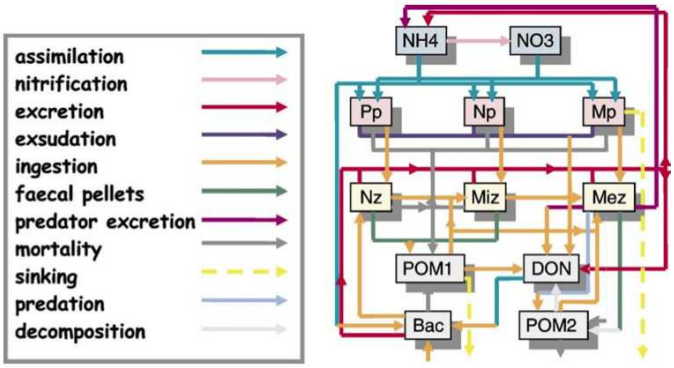


Figure 12: Biogeochemical model (NH4: Ammonium; NH3: nitrate; Pp, Np, Mp: pico-, nano-, microphytoplankton; Nz, Miz, Mez: nano-, micro-, mesozooplankton; PON1, PON2: type 1 and 2 particulate organic nitrogen; Bac: bacteria; DON: dissolved organic nitrogen).

equation

$$
\begin{aligned}
\frac{\partial\, \mathtt{pp}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial\, \mathtt{pp}}{\partial z}\right) + \big((1-exud_{\mathtt{pp}})\mu_{\mathtt{pp}} - mort_{\mathtt{pp}}\big)\mathtt{pp} - ing_{\mathtt{pp},\mathtt{nz}}\mathtt{nz} \\
\frac{\partial\, \mathtt{np}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial\, \mathtt{np}}{\partial z}\right) + \big((1-exud_{\mathtt{np}})\mu_{\mathtt{np}} - mort_{\mathtt{np}}\big)\mathtt{np} - ing_{\mathtt{np},\mathtt{miz}}\mathtt{miz} \\
\frac{\partial\, \mathtt{mp}}{\partial t} &= \frac{\partial}{\partial z}\left(\lambda \frac{\partial\, \mathtt{mp}}{\partial z}\right) + \big((1-exud_{\mathtt{mp}})\mu_{\mathtt{mp}} - mort_{\mathtt{mp}}\big)\mathtt{mp} - ing_{\mathtt{mp},\mathtt{mez}}\mathtt{mez} - sin_{\mathtt{mp}}\frac{\partial \mathtt{mp}}{\partial z}
\end{aligned}
$$

where `nz`, `miz` and `mez` are the nano-, micro- and mesozooplankton biomasses, respectively, and the other notations are

λ	vertical turbulent diffusivity ($\mathrm{m}^2.\mathrm{s}^{-1}$)
$exud_A$	exudation of A (percentage)
μ_A	growth rate of A (day^{-1})
$mort_A$	mortality rate of A (day^{-1})
$ing_{A,B}$	ingestion rate of A by predator B (mgChl)
$sin_{\mathtt{mp}}$	sinking velocity of microphytoplankton ($\mathrm{m.day}^{-1}$)

In our experiment, we focus on two different outputs: the annual maximum of chlorophyll-a concentration in surface water Y_{surf} and the annual maximum of the mean of cholorphyll-a concentration between 20 and 50 meters in depth Y_{depth}. These are practical indicators of biological activity. We are interested in the influence of eight parameters among the 87 input factors. On the one hand, we consider 6 a priori sensitive parameters $\mu_{max\mathtt{pp}}$, $\mu_{max\mathtt{np}}$, $\mu_{max\mathtt{mp}}$, $I_{opt\mathtt{pp}}$, $I_{opt\mathtt{np}}$ and $I_{opt\mathtt{mp}}$ where μ_{maxA} and I_{optA} denote the maximum growth rate of A and the optimum insolation for A, respectively. These input factors are directly related to the growth rate of A, μ_A (see details in Appendix C). On the other hand, we consider the maximum growth rate of bacteria $\mu_{max\mathtt{bac}}$ and the sinking velocity of particulate organic nitrogen (type 1) $sin_{\mathtt{pon1}}$ which have a priori a negligible effect on chlorophyll-a concentration since they do not act directly on `pp`, `np` and `mp` but on the state variables `bac` and `pon1`. We take these eight parameters to be independent gamma distributed random variables with parameters given in Table 2. We estimate all first- and second-order SIs of both outputs Y_{surf} and Y_{depth} by using the estimators defined in Sections 4.3 and 4.4 with sample sizes $n = 65536$ and $n = 66049$, respectively.

The first-order SIs are estimated by using nested replicated latin hypercubes following Qian's construction Qian (2009). They allow to visualize empirical convergence of the estimated indices as shown in Figure 13. The estimated indices at the biggest sample size ($n = 65536$) are reported in Tables 3 and 4; we can notice that both outputs do not define an additive model since in both cases, the sum of the first-order SIs are less than sixty percents. We also notice that $\mu_{max\mathtt{pp}}$ is important in both outputs, while three other a priori important parameters — $\mu_{max\mathtt{np}}$, $I_{opt\mathtt{np}}$ and $I_{opt\mathtt{mp}}$— have actually no effect. At last, it is surprising to observe that the parameter $\mu_{max\mathtt{bac}}$, which does not act directly on both outputs, has non-zero values.

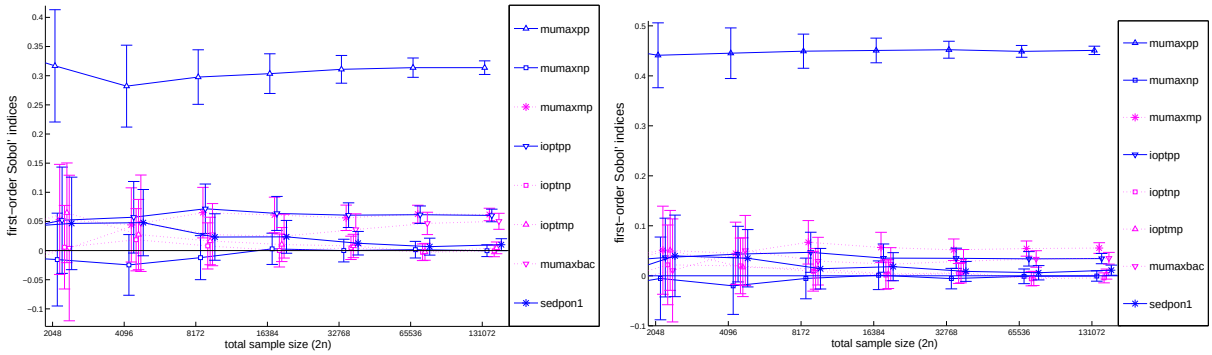


Figure 13: Plots of first-order SIs with error bars — 99% confidence interval — for both outputs Y_{surf} (left) and Y_{depth} (right).

The second-order SIs are estimated by using a replicated latin hypercube based on an orthogonal array with 257 levels, index 1 and strength 2 — i.e. $n = 66049$ — following Bush's construction, see Bose (1938). The results are reported in Tables 5 and 6; they confirm that $\mu_{max\mathtt{pp}}$ has the main role in both outputs since the non-negligible second-order SIs are all related to the latter. As a conclusion, we can notice that both outputs are extremely complex and contain, without any doubt, interactions of order more than or equal to 3. Such an analysis with the MC estimator of SIs would be less efficient without the new approach we proposed in this paper. More precisely, both order 1 and order 2 analysis using the classic MC estimator — i.e. estimating all the SIs of order 1 or 2 — only could use a sample size of 30000 instead of 132000 since this classic approach needs 9 independent samples while the new one only needs 2 for the order 1 analysis and 18 independent samples while the new one only needs 4 for the order 2 analysis, see Saltelli (2002).

5 Conclusion

We have introduced a new method to estimate all the k-th order SIs by using only 2 samples, for any k. This outperforms existing methods including the combinatorial results established by Saltelli (2002). We derive theoretical results in the particular case of first-order SIs from the work by Janon et al. (2012) on asymptotical properties of SIs and from the work by Loh (1996) on asymptotical properties of LHS. Further works will consist in deriving these theoretical results to higher-order SIs and in improving the method by studying how

	label	k	θ	mean	standard deviation
$\mu_{max\mathtt{pp}}$ (day^{-1})	1	9	0.33	3	1
$\mu_{max\mathtt{np}}$ (day^{-1})	2	9	0.28	2.5	0.83
$\mu_{max\mathtt{mp}}$ (day^{-1})	3	9	0.22	2	0.67
$I_{opt\mathtt{pp}}$ ($\mathrm{W.m}^{-2}$)	4	9	1.11	10	3.33
$I_{opt\mathtt{np}}$ ($\mathrm{W.m}^{-2}$)	5	9	1.67	15	5
$I_{opt\mathtt{mp}}$ ($\mathrm{W.m}^{-2}$)	6	9	2.22	20	6.67
$\mu_{max\mathtt{bac}}$ (day^{-1})	7	9	0.22	2	0.67
$sin_{\mathtt{pon1}}$ ($\mathrm{m.day}^{-1}$)	8	9	0.17	1.5	0.5

Table 2: Distributions of variables using gamma density $f(x;k,\theta) = x^{k-1}\exp(-x/\theta)/(\Gamma(\theta)\theta^k)$, where $\Gamma(\cdot)$ is the gamma function.

	$\underline{S}_{\{1\}}$	$\underline{S}_{\{2\}}$	$\underline{S}_{\{3\}}$	$\underline{S}_{\{4\}}$	$\underline{S}_{\{5\}}$	$\underline{S}_{\{6\}}$	$\underline{S}_{\{7\}}$	$\underline{S}_{\{8\}}$
estimated index	0.314	0	0.061	0.060	0	0.003	0.051	0.010
estimated error	0.010	0.011	0.012	0.011	0.010	0.010	0.013	0.012

Table 3: Estimation of first-order SIs for the output Y_{surf}. The estimated error is the radius of the 99% confidence interval.

	$\underline{S}_{\{1\}}$	$\underline{S}_{\{2\}}$	$\underline{S}_{\{3\}}$	$\underline{S}_{\{4\}}$	$\underline{S}_{\{5\}}$	$\underline{S}_{\{6\}}$	$\underline{S}_{\{7\}}$	$\underline{S}_{\{8\}}$
estimated index	0.451	0	0.055	0.034	0	0	0.035	0.011
estimated error	0.009	0.010	0.010	0.010	0.010	0.010	0.012	0.010

Table 4: Estimation of first-order SIs for the output Y_{depth}. The estimated error is the radius of the 99% confidence interval.

	$\underline{S}_{\{1,2\}}$	$\underline{S}_{\{1,3\}}$	$\underline{S}_{\{1,4\}}$	$\underline{S}_{\{1,5\}}$	$\underline{S}_{\{1,6\}}$	$\underline{S}_{\{1,7\}}$	$\underline{S}_{\{1,8\}}$	$\underline{S}_{\{2,3\}}$	$\underline{S}_{\{2,4\}}$	$\underline{S}_{\{2,5\}}$
estimated index	0.374	0.479	0.424	0.339	0.324	0.400	0.318	0.069	0.066	0.016
estimated error	0.012	0.011	0.013	0.011	0.011	0.010	0.011	0.011	0.011	0.011

	$\underline{S}_{\{2,6\}}$	$\underline{S}_{\{2,7\}}$	$\underline{S}_{\{2,8\}}$	$\underline{S}_{\{3,4\}}$	$\underline{S}_{\{3,5\}}$	$\underline{S}_{\{3,6\}}$	$\underline{S}_{\{3,7\}}$	$\underline{S}_{\{3,8\}}$	$\underline{S}_{\{4,5\}}$	$\underline{S}_{\{4,6\}}$
estimated index	0.015	0.069	0.015	0.125	0.074	0.075	0.128	0.072	0.077	0.070
estimated error	0.010	0.015	0.010	0.011	0.011	0.011	0.013	0.011	0.011	0.011

	$\underline{S}_{\{4,7\}}$	$\underline{S}_{\{4,8\}}$	$\underline{S}_{\{5,6\}}$	$\underline{S}_{\{5,7\}}$	$\underline{S}_{\{5,8\}}$	$\underline{S}_{\{6,7\}}$	$\underline{S}_{\{6,8\}}$	$\underline{S}_{\{7,8\}}$
estimated index	0.121	0.066	0.017	0.055	0.014	0.056	0.009	0.050
estimated error	0.013	0.011	0.010	0.015	0.010	0.014	0.010	0.015

Table 5: Estimation of second-order SIs for the output Y_{surf}. The estimated error is the radius of the 99% confidence interval.

	$\underline{S}_{\{1,2\}}$	$\underline{S}_{\{1,3\}}$	$\underline{S}_{\{1,4\}}$	$\underline{S}_{\{1,5\}}$	$\underline{S}_{\{1,6\}}$	$\underline{S}_{\{1,7\}}$	$\underline{S}_{\{1,8\}}$	$\underline{S}_{\{2,3\}}$	$\underline{S}_{\{2,4\}}$	$\underline{S}_{\{2,5\}}$
estimated index	0.506	0.593	0.510	0.455	0.450	0.515	0.447	0.056	0.034	0.005
estimated error	0.010	0.009	0.010	0.009	0.009	0.008	0.009	0.011	0.011	0.011

	$\underline{S}_{\{2,6\}}$	$\underline{S}_{\{2,7\}}$	$\underline{S}_{\{2,8\}}$	$\underline{S}_{\{3,4\}}$	$\underline{S}_{\{3,5\}}$	$\underline{S}_{\{3,6\}}$	$\underline{S}_{\{3,7\}}$	$\underline{S}_{\{3,8\}}$	$\underline{S}_{\{4,5\}}$	$\underline{S}_{\{4,6\}}$
estimated index	0.008	0.055	0.009	0.087	0.057	0.064	0.109	0.063	0.041	0.043
estimated error	0.010	0.014	0.010	0.011	0.011	0.011	0.013	0.011	0.010	0.010

	$\underline{S}_{\{4,7\}}$	$\underline{S}_{\{4,8\}}$	$\underline{S}_{\{5,6\}}$	$\underline{S}_{\{5,7\}}$	$\underline{S}_{\{5,8\}}$	$\underline{S}_{\{6,7\}}$	$\underline{S}_{\{6,8\}}$	$\underline{S}_{\{7,8\}}$
estimated index	0.082	0.041	0.009	0.040	0.007	0.046	0.006	0.041
estimated error	0.013	0.010	0.010	0.014	0.010	0.014	0.010	0.014

Table 6: Estimation of second-order SIs for the output Y_{depth}. The estimated error is the radius of the 99% confidence interval.

we can estimate correctly the asymptotic variance of the new estimator.

Acknowledgments

The authors are grateful to Eric Blayo, Jean-Michel Brankart and Pierre Brasseur for valuable discussions on the simulator MODECOGeL and more generally on marine ecosystem models. They also thank Art Owen for his helpful comments. This work has been partially supported by French National Research Agency (ANR) through COSINUS program (project COSTA-BRAVA n° ANR-09-COSI-015).

A Proof of Proposition 1

NB: the integration set is generally omitted for integrals defined over a unit hypercube $[0,1]^s$ with any $s \leq d$,

A.1 Technical lemmas

Let $\dot{\mathbf{X}}^1$ and $\dot{\mathbf{X}}^2$ two distinct points of a Latin hypercube of size n in $[0,1]^d$. For any function f defined on $[0,1]^d$, consider $\dot{Y}^1 = f(\dot{\mathbf{X}}^1)$ et $\dot{Y}^2 = f(\dot{\mathbf{X}}^2)$. In Theorem 1 in Stein (1987), Stein gives the following result

Theorem 2. *If f is a square integrable function then as n tends to $+\infty$, we have*

$$\mathrm{Cov}(\dot{Y}^1, \dot{Y}^2) = -\frac{1}{n}\sum_{i=1}^{d} \sigma_i^2 + o(n^{-1}) \ . \tag{15}$$

In this subsection, we prove an analogous result with more general settings and without the asymptotic assumption on n (see Lemma 4).

Notation and defintions For s and n in $\mathbb{N}^*$, define the partition of $[0,1)^s$ in elementary hypercubes of side $1/n$,

$$\mathcal{Q}_s(n) = \left\{ Q \subseteq [0,1)^s \mid Q = \prod_{i=1}^{s} [\alpha_i, \beta_i),\ \alpha_i \in \left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\right\},\ \beta_i = \alpha_i + \frac{1}{n} \right\}. \tag{16}$$

For any square integrable function g defined on $[0,1)^s$, $s \leq d$, define the sequence with general term

$$u_n(g) = n^s \sum_{Q \in \mathcal{Q}_s(n)} \left(\int_Q g(\mathbf{x}) d\mathbf{x} \right)^2,\ n \in \mathbb{N}. \tag{17}$$

Outline The first lemma is the analogous result for Lebesgue integrability of a result given in Equation (A.4) in Stein (1987) for Riemann integrability. The second lemma gives an important inequality which allows to work without asymptotic assumption on n. The last one consists in simplifying integrals under LHS using the ANOVA decomposition. Lemma 4 provides the expected inequalities.

Lemma 1. *If g is a square integrable function, the sequence $\big(u_n(g)\big)$ converges to $\int g^2(\mathbf{x})d\mathbf{x}$ as n tends to $+\infty$.*

Proof. Noting that

$$u_n(g) = \int g_n(\mathbf{x}) d\mathbf{x} \tag{18}$$

where

$$\forall \mathbf{x} \in [0,1)^s,\ g_n(\mathbf{x}) = \sum_{Q \in \mathcal{Q}_s(n)} \left(n^s \int_Q g(\mathbf{y}) d\mathbf{y} \right)^2 \mathbf{1}_Q(\mathbf{x}) \tag{19}$$

Lemma 1 is a straightforward consequence of the dominated convergence theorem. So let us prove that there exists an integrable function h such that for all $n \in \mathbb{N}^*$, $|g_n| \leq h$ almost surely, and g_n converges pointwise to g^2, and the conclusion will follow.

First since g is a square integrable function, we have $|g(\mathbf{x})| \leq M$ a.s., and by their definition, the g_n's are as well. Hence their exists an integrable function ($h : \mathbf{x} \mapsto M$) such that $|g_n| \leq h$ almost surely. Concerning the pointwise convergence, let us prove that for any $\mathbf{x} \in [0,1)^s$,

$$\forall \varepsilon > 0,\ \exists\ N > 0,\ \forall n \geq N,\ \left| \left(n^s \int_{Q_{\mathbf{x}}} g(\mathbf{y}) d\mathbf{y} \right)^2 - g^2(\mathbf{x}) \right| < \varepsilon \tag{20}$$

where $Q_{\mathbf{x}}$ is the set Q in $\mathcal{Q}_s(n)$ containing $\mathbf{x}$. This is obvious if g^2 is a simple function and we easily generalize the result to any g^2 since any measurable function is a pointwise limit of simple functions. □

Lemma 2. *The sequence $\big(u_n(g)\big)$ is dominated by $\int g^2(\mathbf{x})d\mathbf{x}$.*

Proof. Let $n \in \mathbb{N}^*$, the result is proved by showing that the sequence of general term $v_k(g) = u_{2^k n}(g)$ is increasing. In this case, by Lemma 1, we have $\lim v_k(g) = \int g^2(\mathbf{x})d\mathbf{x}$, and since v_k is increasing, all the terms of this sequence are dominated by $\int g^2(\mathbf{x})d\mathbf{x}$, hence $v_0(g) = u_n(g) \leq \int g^2(\mathbf{x})d\mathbf{x}$. To prove that the sequence $\big(v_k(g)\big)$ is increasing, note that

$$\begin{aligned} v_{k+1}(g) &= (2^{k+1}n)^s \sum_{Q \in \mathcal{Q}_s(2^{k+1}n)} \left(\int_Q g(\mathbf{x}) d\mathbf{x} \right)^2 \\ &= (2^k n)^s \sum_{Q \in \mathcal{Q}_s(2^k n)} \left(2^s \sum_{P \in \mathcal{P}(Q, 2^{k+1}n)} \left(\int_P g(\mathbf{x}) d\mathbf{x} \right)^2 \right) \end{aligned} \tag{21}$$

where $\mathcal{P}(Q, 2^{k+1}n) = \mathcal{Q}(2^{k+1}n) \cap Q$. Then by Jensen inequality, we have

$$2^s \sum_{P \in \mathcal{P}(Q,2^{k+1}n)} \left(\int_P g(\mathbf{x}) d\mathbf{x} \right)^2 \geq \left(\int_Q g(\mathbf{x}) d\mathbf{x} \right)^2 \tag{22}$$

and we conclude that $v_{k+1}(g) \geq v_k(g)$. □

For $0 \leq x_1,\ x_2 \leq 1$ define

$$r_n(x_1, x_2) = \begin{cases} 1 & \text{if } \lfloor nx_1 \rfloor = \lfloor nx_2 \rfloor \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

where $\lfloor \cdot \rfloor$ is the floor function. We now end with the following result

Lemma 3. *Let $\mathfrak{v}$ be a subset of $\{1, \ldots, d\}$, we have*

$$\int f(\mathbf{x}_1) f(\mathbf{x}_2) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 = \int \sum_{\mathfrak{w} \subseteq \mathfrak{v}} f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}}) f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}}) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 \ . \tag{24}$$

Proof. By the ANOVA decomposition — see (1) — we have

$$\int f(\mathbf{x}_1) f(\mathbf{x}_2) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 = \int \sum_{\mathfrak{w}_1 \subseteq \{1,..,d\}} \sum_{\mathfrak{w}_2 \subseteq \{1,..,d\}} f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 \ . \tag{25}$$

Then note that a certain number of terms in the member on the right-hand side vanishe. If $(\mathfrak{w}_1 \cap \mathfrak{v}^c) \cup (\mathfrak{w}_2 \cap \mathfrak{v}^c) \neq \emptyset$ then suppose without loss of generality that there exists $k \in \mathfrak{w}_1 \setminus \mathfrak{v}$; we have

$$\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 = \int \underbrace{\left(\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) dx_{1k} \right)}_{I_1} f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_{1\{k\}^c} d\mathbf{x}_2 \tag{26}$$

and note that, by a basic property of the ANOVA decomposition, $I_1 = 0$. If $(\mathfrak{w}_1 \cap \mathfrak{v}^c) \cup (\mathfrak{w}_2 \cap \mathfrak{v}^c) = \emptyset$ and $\mathfrak{w}_1 \neq \mathfrak{w}_2$, then suppose without loss of generality that there exists $k \in \mathfrak{w}_1 \setminus \mathfrak{w}_2$. In this case, we have

$$\begin{aligned} &\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \prod_{i \in \mathfrak{v}} r_n(x_{1i}, x_{2i}) d\mathbf{x}_1 d\mathbf{x}_2 \\ &\quad = \int \underbrace{\left(\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) r_n(x_{1k}, x_{2k}) dx_{1k} dx_{2k} \right)}_{I_2} f_{\mathfrak{w}_2}(\mathbf{x}_{2\mathfrak{w}_2}) \left(\prod_{i \in \mathfrak{v} \setminus \{k\}} r_n(x_{1i}, x_{2i}) \right) d\mathbf{x}_{1\{k\}^c} d\mathbf{x}_{2\{k\}^c} \end{aligned} \tag{27}$$

and note that by the definition of r_n, we have

$$\int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) r_n(x_{1k}, x_{2k}) dx_{1k} dx_{2k} = \int f_{\mathfrak{w}_1}(\mathbf{x}_{1\mathfrak{w}_1}) dx_{1k} \tag{28}$$

and thus $I_2 = 0$. The conclusion of the lemma follows. □

Let $\mathfrak{u}$ be a non-empty subset of $\{1, \ldots d\}$ and consider $(\dot{\mathbf{X}}^j)_j \sim \mathcal{LH}(n, d)$ and $(\dot{\mathbf{Z}}^j)_j \sim \mathcal{LH}(n, d)$. For any function f defined on $[0, 1]^d$, consider

$$\dot{Y}^1 = f(\dot{\mathbf{X}}^1) \text{ and } \dot{Y}^2_{\mathfrak{u}} = f(\dot{\mathbf{X}}^2_{\mathfrak{u}} : \dot{\mathbf{Z}}^2_{\mathfrak{u}^c}) \ . \tag{29}$$

We have the following result

Lemma 4. *If f is a square integrable function then we have*

$$-\sum_{\substack{\emptyset\neq\mathfrak{w}\subseteq\mathfrak{u}\\ |\mathfrak{w}|\ odd}}\frac{\sigma^2_{\mathfrak{w}}}{(n-1)^{|\mathfrak{w}|}}\leq \mathrm{Cov}(\dot{Y}^1,\dot{Y}^2_{\mathfrak{u}})\leq \sum_{\substack{\emptyset\neq\mathfrak{w}\subseteq\mathfrak{u}\\ |\mathfrak{w}|\ even}}\frac{\sigma^2_{\mathfrak{w}}}{(n-1)^{|\mathfrak{w}|}}\ . \tag{30}$$

Proof. Recall that for $0\leq x_1,\ x_2\leq 1$,

$$r_n(x_1,x_2)=\begin{cases}1 & \text{if } \lfloor nx_1\rfloor=\lfloor nx_2\rfloor\\ 0 & \text{otherwise,}\end{cases} \tag{31}$$

where $\lfloor\cdot\rfloor$ is the floor function. For $\mathbf{x}_1=(x_{11},\ldots,x_{1d})$ in $[0,1)^d$, define $\mathbf{x}_{1\mathfrak{v}}=(x_{1i_1},\ldots,x_{1i_{|\mathfrak{v}|}})$ where $\mathfrak{v}=\{i_1,\ldots,i_{|\mathfrak{v}|}\}$. Due to the joint density of $(\dot{\mathbf{X}}^1_{\mathfrak{u}},\dot{\mathbf{X}}^2_{\mathfrak{u}})$ under LHS — see McKay et al. (1979) or Stein (1987) — and by Lemma 3, we have

$$\begin{aligned}
\mathrm{Cov}(\dot{Y}^1,\dot{Y}^2_{\mathfrak{u}})+\Big(\int f(\mathbf{x})d\mathbf{x}\Big)^2 &= \int f(\mathbf{x}_1)f(\mathbf{x}_2)\Big(\frac{n}{n-1}\Big)^{|\mathfrak{u}|}\prod_{i\in\mathfrak{u}}\big(1-r_n(x_{1i},x_{2i})\big)d\mathbf{x}_1d\mathbf{x}_2\\
&= \Big(\frac{n}{n-1}\Big)^{|\mathfrak{u}|}\sum_{\mathfrak{v}\subseteq\mathfrak{u}}(-1)^{|\mathfrak{v}|}\int f(\mathbf{x}_1)f(\mathbf{x}_2)\prod_{i\in\mathfrak{v}}r_n(x_{1i},x_{2i})d\mathbf{x}_1d\mathbf{x}_2\\
&= \Big(\frac{n}{n-1}\Big)^{|\mathfrak{u}|}\sum_{\mathfrak{v}\subseteq\mathfrak{u}}(-1)^{|\mathfrak{v}|}\int \sum_{\mathfrak{w}\subseteq\mathfrak{v}}f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{v}}r_n(x_{1i},x_{2i})d\mathbf{x}_1d\mathbf{x}_2\\
&= \Big(\frac{n}{n-1}\Big)^{|\mathfrak{u}|}\sum_{\mathfrak{v}\subseteq\mathfrak{u}}(-1)^{|\mathfrak{v}|}\sum_{\mathfrak{w}\subseteq\mathfrak{v}}\Big(\frac{1}{n}\Big)^{|\mathfrak{v}|-|\mathfrak{w}|}\int f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}
\end{aligned} \tag{32}$$

Then note that for any function of $\mathfrak{w}$ denoted by $A(\mathfrak{w})$, we have

$$\begin{aligned}
\sum_{\mathfrak{v}\subseteq\mathfrak{u}}(-1)^{|\mathfrak{v}|}\sum_{\mathfrak{w}\subseteq\mathfrak{v}}\Big(\frac{1}{n}\Big)^{|\mathfrak{v}|-|\mathfrak{w}|}A(\mathfrak{w}) &= \sum_{\mathfrak{v}\subseteq\mathfrak{u}}\Big(-\frac{1}{n}\Big)^{|\mathfrak{v}|}\sum_{\mathfrak{w}\subseteq\mathfrak{v}}\Big(\frac{1}{n}\Big)^{-|\mathfrak{w}|}A(\mathfrak{w})\\
&= \sum_{\mathfrak{w}\subseteq\mathfrak{u}}\left(\sum_{k=0}^{|\mathfrak{u}|-|\mathfrak{w}|}\binom{|\mathfrak{u}|-|\mathfrak{w}|}{k}\Big(-\frac{1}{n}\Big)^{k+|\mathfrak{w}|}\right)\Big(\frac{1}{n}\Big)^{-|\mathfrak{w}|}A(\mathfrak{w})\\
&= \sum_{\mathfrak{w}\subseteq\mathfrak{u}}\Big(\frac{n-1}{n}\Big)^{|\mathfrak{u}|-|\mathfrak{w}|}(-1)^{|\mathfrak{w}|}A(\mathfrak{w})
\end{aligned} \tag{33}$$

Hence, we deduce that

$$\mathrm{Cov}(\dot{Y}^1,\dot{Y}^2_{\mathfrak{u}})=\sum_{\substack{\mathfrak{w}\subseteq\mathfrak{u}\\ \mathfrak{w}\neq\emptyset}}\Big(\frac{n}{n-1}\Big)^{|\mathfrak{w}|}(-1)^{|\mathfrak{w}|}\int f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}\ . \tag{34}$$

Finally by the definition of r_n, we have

$$0\leq\int f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}\leq\sum_{Q\in\mathcal{Q}_{|\mathfrak{w}|}(n)}\Big(\int_Q f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})d\mathbf{x}_{1\mathfrak{w}}\Big)^2 \tag{35}$$

and by Lemma 2, this gives

$$0\leq\int f_{\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}\leq\frac{\sigma^2_{\mathfrak{w}}}{n^{|\mathfrak{w}|}}\ . \tag{36}$$

The latter inequalities and (34) lead to Lemma 4. □

Remark 3. *Note that if $\mathfrak{u}=\{1,\ldots,d\}$, the resulting inequalities are*

$$-\sum_{\substack{\mathfrak{w}\neq\emptyset\\ |\mathfrak{w}|\ odd}}\frac{\sigma^2_{\mathfrak{w}}}{(n-1)^{|\mathfrak{w}|}}\leq \mathrm{Cov}(\dot{Y}^1,\dot{Y}^2)\leq \sum_{\substack{\mathfrak{w}\neq\emptyset\\ |\mathfrak{w}|\ even}}\frac{\sigma^2_{\mathfrak{w}}}{(n-1)^{|\mathfrak{w}|}}\ . \tag{37}$$

This consists of a non-asymptotic equivalent of the theorem due to Stein presented at the beginning of this section.

A.2 Proof of Proposition 1

Proof.

(ii) Concerning $\widetilde{S}_{\mathrm{u},n}^{LHS}$, it is easy to show that

$$\widetilde{S}_{\mathrm{u},n}^{LHS} = \Phi(\overline{\mathbf{V}}_n) \tag{38}$$

where

$$\overline{\mathbf{V}}_n = \frac{1}{n}\sum_{j=1}^{n} \mathbf{V}_j \tag{39}$$

$$\mathbf{V}_j = \Big(\big(\dot{Y}^j - \mathbb{E}[Y]\big)\big(\dot{Y}_{\mathrm{u}}^j - \mathbb{E}[Y]\big) \ , \ \dot{Y}^j - \mathbb{E}[Y] \ , \ \dot{Y}_{\mathrm{u}}^j - \mathbb{E}[Y] \ , \ \big(\dot{Y}^j - \mathbb{E}[Y]\big)^2 \Big)^T \tag{40}$$

and

$$\Phi(x,y,z,t) = \frac{x - yz}{t - y^2} \ . \tag{41}$$

Then we deduce from Theorem 2 in Loh (1996) that

$$\sqrt{n}\big(\overline{\mathbf{V}}_n - \mu\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_4(0,\Gamma) \tag{42}$$

where $\mu = (\underline{\tau}_{\mathrm{u}}^2, 0, 0, \sigma^2)^T$ and Γ is the covariance matrix of $\mathbf{R}_1 = \mathbf{V}_1 - \mathbf{A}_1$ — see details in Eq. (3) in Loh (1996) — defined by

$$\forall i \in \{1,\dots,4\}, \ A_{1i} \text{ is the additive part — see (1) — of } V_{1i} \ . \tag{43}$$

Thus the Delta method — see Theorem 3.1 in Van der Vaart (1998) — gives

$$\sqrt{n}\big(\widetilde{S}_{\mathrm{u},n}^{LHS} - \underline{S}_{\mathrm{u}}\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_1(0, g^T\Gamma g) \tag{44}$$

where $g = \nabla\Phi(\mu)$. Developing the term $g^T\Gamma g$ does not seem to provide any useful information. However, denoting σ_{LHS}^2 this term, and σ_{IID}^2 the analogous quantity in the CLT for simple random sampling, we can show that $\sigma_{LHS}^2 \leq \sigma_{IID}^2$. Indeed we first note that, for simple random sampling, the variance given in Janon et al. (2012) reads

$$\sigma_{IID}^2 = \frac{\mathrm{Var}\big[V_{11} - \underline{S}_{\mathrm{u}} V_{14}\big]}{\sigma^2} \tag{45}$$

and for LHS, it is easy to show that

$$\sigma_{LHS}^2 = \frac{\mathrm{Var}\big[R_{11} - \underline{S}_{\mathrm{u}} R_{14}\big]}{\sigma^2} \ . \tag{46}$$

Hence,

$$\sigma_{IID}^2 = \sigma_{LHS}^2 + \frac{\mathrm{Var}\big[A_{11} - \underline{S}_{\mathrm{u}} A_{14}\big]}{\sigma^2} \tag{47}$$

and the conclusion of (ii) for $\widetilde{S}_{\mathrm{u},n}^{LHS}$ follows. Concerning $\widehat{S}_{\mathrm{u},n}^{LHS}$, the proof follows the same lines — see Proof of (10) in Janon et al. (2012) for details.

(iii) First we have

$$\begin{aligned}\mathbb{E}\big[\widetilde{\underline{\tau}}^{2,LHS}_{\mathrm{u},n}\big] &= \frac{n-1}{n^2}\sum_{j=1}^{n}\mathbb{E}\big[\dot{Y}^j\dot{Y}^j_{\mathrm{u}}\big] - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\substack{l=1\\l\neq j}}^{n}\mathbb{E}\big[\dot{Y}^j\dot{Y}^l_{\mathrm{u}}\big] \\ &= \frac{n-1}{n}\Big(\mathbb{E}[Y]^2+\underline{\tau}^2_{\mathrm{u}}\Big) - \frac{n-1}{n}\Big(\mathrm{Cov}(\dot{Y}^1,\dot{Y}^2_{\mathrm{u}})+\mathbb{E}[Y]^2\Big) \qquad (48)\end{aligned}$$

and thanks to Lemma 4 in Appendix A, it gives

$$\mathbb{E}\big[\widetilde{\underline{\tau}}^{2,LHS}_{\mathrm{u},n}\big] = \underline{\tau}^2_{\mathrm{u}} + B_{n,1} \qquad (49)$$

with

$$-\frac{1}{n-1}\underline{\tau}^2_{\mathrm{u}} \leq B_{n,1} \leq 0\ . \qquad (50)$$

Concerning $\widetilde{\sigma}^{2,LHS}_n$, we have

$$\begin{aligned}\mathbb{E}\big[\widetilde{\sigma}^{2,LHS}_{\mathrm{u},n}\big] &= \frac{n-1}{n^2}\sum_{j=1}^{n}\mathbb{E}\big[(\dot{Y}^j)^2\big] - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\substack{l=1\\l\neq j}}^{n}\mathbb{E}\big[\dot{Y}^j\dot{Y}^l\big] \\ &= \frac{n-1}{n}\mathbb{E}[Y^2] - \frac{n-1}{n}\Big(\mathrm{Cov}(\dot{Y}^1,\dot{Y}^2)+\mathbb{E}[Y]^2\Big) \qquad (51)\end{aligned}$$

and noting that

$$\mathrm{Cov}(\dot{Y}^1,\dot{Y}^2) = \mathrm{Cov}\big(\dot{Y}^1,\dot{Y}^2_{\{1,\ldots,d\}}\big) \qquad (52)$$

we conclude that

$$\mathbb{E}\big[\widetilde{\sigma}^{2,LHS}_n\big] = \sigma^2 + B_{n,2} \qquad (53)$$

with

$$-\frac{1}{n-1}\sigma^2 \leq B_{n,2} \leq 0\ . \qquad (54)$$

As for $\widehat{\underline{\tau}}^{2,LHS}_{\mathrm{u},n}$ and $\widehat{\sigma}^{2,LHS}_n$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\dot{Y}^j+\dot{Y}^j_{\mathrm{u}}}{2}\right)^2\right] &= \frac{1}{4n}\mathbb{E}\big[(\dot{Y}^1+\dot{Y}^1_{\mathrm{u}})^2\big] + \frac{1}{4n^2}\sum_{j=1}^{n}\sum_{\substack{l=1\\l\neq j}}^{n}\mathbb{E}\big[(\dot{Y}^j+\dot{Y}^j_{\mathrm{u}})(\dot{Y}^l+\dot{Y}^l_{\mathrm{u}})\big] \\ &= \frac{1}{2n}\Big(\mathbb{E}[Y^2]+\underline{\tau}^2_{\mathrm{u}}+\mathbb{E}[Y]^2\Big) + \frac{n-1}{n}\mathbb{E}[Y]^2 + \frac{n-1}{2n}\Big(\mathrm{Cov}(\dot{Y}^1,\dot{Y}^2)+\mathrm{Cov}(\dot{Y}^1,\dot{Y}^2_{\mathrm{u}})\Big) \\ &= \frac{1}{2n}\Big(\sigma^2+\underline{\tau}^2_{\mathrm{u}}\Big) + \mathbb{E}[Y]^2 + \frac{n-1}{2n}\Big(\mathrm{Cov}(\dot{Y}^1,\dot{Y}^2)+\mathrm{Cov}(\dot{Y}^1,\dot{Y}^2_{\mathrm{u}})\Big)\ . \qquad (55)\end{aligned}$$

Then it is easy to conclude that

$$\mathbb{E}\big[\widehat{\underline{\tau}}^{2,LHS}_{\mathrm{u},n}\big] = \underline{\tau}^2_{\mathrm{u}} + B_{n,3} \qquad (56)$$

$$\mathbb{E}\big[\widehat{\sigma}^{2,LHS}_n\big] = \sigma^2 + B_{n,3} \qquad (57)$$

with

$$-\frac{1}{2(n-1)}(\sigma^2+\underline{\tau}^2_{\mathrm{u}}) \leq B_{n,3} \leq 0\ . \qquad (58)$$

□

B Proof of Proposition 2

We first give three lemmas. The proof of Proposition 2 is given in Section B.2.

B.1 Technical lemmas

Lemma 5. *Let $d \in \mathbb{N}^*$, if $n \geq \frac{d^2}{2}$ then*

$$\left(1+\frac{1}{n}\right)^d - 1 \leq \frac{d+1}{n} \ . \tag{59}$$

Proof. If $d = 1$, the result is obvious. Otherwise, for any $x > 0$, consider the function g_d defined by

$$g_d(x) = \left(1+\frac{1}{x}\right)^d - 1 - \frac{d+1}{x} \ . \tag{60}$$

We show that

(1) if there exists $x_0 > 0$ such that $g_d(x_0) \leq 0$ then for all $x \geq x_0$, $g_d(x) \leq 0$

(2) $g_d(d^2/2) \leq 0$

and the conclusion follows. Concerning (1) note that

$$g_d(x) = 1 + \frac{d}{x} + O(x^{-2}) - 1 - \frac{d}{x} - \frac{1}{x} = -\frac{1}{x} + O(x^{-2}) \tag{61}$$

and then that g_d is negative as x tends to $+\infty$. Moreover for any $d > 1$, g_d is first decreasing and then increasing. Indeed, we have

$$g_d'(x) = -\frac{d}{x^2}\left(1+\frac{1}{x}\right)^{d-1} + \frac{d+1}{x^2} \tag{62}$$

and we deduce that $g_d'(x_0) = 0$ with

$$x_0 = \frac{1}{\left(\frac{d+1}{d}\right)^{1/(d-1)} - 1} > 0 \tag{63}$$

and is negative on the left side and positive on the right side. The conclusion of (1) follows. Concerning (2), it is easy to check that it is true for $d = 1$ and 2, and for $d \geq 3$ we have

$$\begin{aligned}
g_d\left(\frac{d^2}{2}\right) &= \sum_{k=0}^{d} \binom{d}{k}\left(\frac{2}{d^2}\right)^k - 1 - \frac{2}{d} - \frac{2}{d^2} \\
&= -\frac{2}{d^3} + \sum_{k=3}^{d} \binom{d}{k}\left(\frac{2}{d^2}\right)^k \\
&\leq -\frac{2}{d^3} + \sum_{k=3}^{d} \frac{1}{k!}\left(\frac{2}{d}\right)^k \\
&\leq -\frac{2}{d^3} + \frac{1}{d^3} + \frac{1}{3d^3} + \frac{2}{3}\sum_{k=4}^{d} \frac{1}{d^k} \\
&\leq -\frac{2}{d^3} + \sum_{k=3}^{d} \frac{1}{d^k} + \frac{1}{3d^3}\left(1 - \sum_{k=1}^{d-3} \frac{1}{d^k}\right) \\
&\leq -\frac{2}{d^3} + \sum_{k=3}^{d} \frac{1}{d^k} \\
&\leq -\frac{2}{d^3} + \frac{2}{d^3}
\end{aligned} \tag{64}$$

and the conclusion follows. □

Let $\mathfrak{u}$ be a non-empty subset of $\{1,\dots d\}$ and consider $(\dot{\mathbf{X}}^j)_j \sim \mathcal{LH}(n,d)$. We have the following result

Lemma 6. *If f is a square integrable function, we have*

$$\mathbb{E}\big[f(\dot{\mathbf{X}}^1)f(\dot{\mathbf{X}}^1_{\mathfrak{u}}:\dot{\mathbf{X}}^2_{\mathfrak{u}^c})\big] = \mathbb{E}[Y]^2 + \underline{\tau}^2_{\mathfrak{u}} + B_{\mathfrak{u},n} \tag{65}$$

where

$$-\mathbb{E}[Y^2]\sum_{\substack{\emptyset\neq\mathfrak{v}\subseteq\mathfrak{u}^c\\ |v|\ odd}}\frac{1}{(n-1)^{|v|}} \leq B_{\mathfrak{u},n} \leq \mathbb{E}[Y^2]\sum_{\substack{\emptyset\neq\mathfrak{v}\subseteq\mathfrak{u}^c\\ |v|\ even}}\frac{1}{(n-1)^{|v|}}\ . \tag{66}$$

Proof. First, due to the joint density of $(\dot{\mathbf{X}}^1_{\mathfrak{u}^c},\dot{\mathbf{X}}^2_{\mathfrak{u}^c})$ under LHS — see McKay et al. (1979) or Stein (1987) — we have

$$\begin{aligned}\mathbb{E}\big[f(\dot{\mathbf{X}}^1)f(\dot{\mathbf{X}}^1_{\mathfrak{u}}:\dot{\mathbf{X}}^2_{\mathfrak{u}^c})\big] &= \int f(\mathbf{x},\mathbf{x}_1)f(\mathbf{x},\mathbf{x}_2)\Big(\frac{n}{n-1}\Big)^{d-|\mathfrak{u}|}\prod_{i\in\mathfrak{u}^c}\big(1-r_n(x_{1i},x_{2i})\big)d\mathbf{x}d\mathbf{x}_1d\mathbf{x}_2\\ &= \Big(\frac{n}{n-1}\Big)^{d-|\mathfrak{u}|}\int\Big(\underbrace{\sum_{\mathfrak{v}\subseteq\mathfrak{u}^c}(-1)^{|\mathfrak{v}|}\int f(\mathbf{x},\mathbf{x}_1)f(\mathbf{x},\mathbf{x}_2)\prod_{i\in\mathfrak{v}}r_n(x_{1i},x_{2i})d\mathbf{x}_1d\mathbf{x}_2}_{I(\mathbf{x})}\Big)d\mathbf{x}\ .\end{aligned} \tag{67}$$

We now denote $f_{\mathbf{x}}:\mathbf{y}\mapsto f(\mathbf{x},\mathbf{y})$ and then by (32) and (33) we have

$$\begin{aligned}I(\mathbf{x}) &= \sum_{\mathfrak{v}\subseteq\mathfrak{u}^c}(-1)^{|\mathfrak{v}|}\int f_{\mathbf{x}}(\mathbf{x}_1)f_{\mathbf{x}}(\mathbf{x}_2)\prod_{i\in\mathfrak{v}}r_n(x_{1i},x_{2i})d\mathbf{x}_1d\mathbf{x}_2\\ &= \sum_{\mathfrak{v}\subseteq\mathfrak{u}^c}(-1)^{|\mathfrak{v}|}\sum_{\mathfrak{w}\subseteq\mathfrak{v}}\Big(\frac{1}{n}\Big)^{|\mathfrak{v}|-|\mathfrak{w}|}\int f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}\\ &= \sum_{\mathfrak{w}\subseteq\mathfrak{u}^c}(-1)^{|\mathfrak{w}|}\Big(\frac{n-1}{n}\Big)^{d-|\mathfrak{u}|-|\mathfrak{w}|}\int f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}\ .\end{aligned} \tag{68}$$

Hence by (36) we have for all $\mathfrak{w}\neq\emptyset$,

$$0\leq\int f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})f_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{2\mathfrak{w}})\prod_{i\in\mathfrak{w}}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\mathfrak{w}}d\mathbf{x}_{2\mathfrak{w}}\leq\frac{\int f^2_{\mathbf{x},\mathfrak{w}}(\mathbf{x}_{1\mathfrak{w}})d\mathbf{x}_{1\mathfrak{w}}}{n^{|\mathfrak{w}|}}\leq\frac{\int f^2_{\mathbf{x}}(\mathbf{x}_1)d\mathbf{x}_1}{n^{|\mathfrak{w}|}} \tag{69}$$

and note that

$$\int f_{\mathbf{x},\emptyset}(\mathbf{x}_{1\emptyset})f_{\mathbf{x},\emptyset}(\mathbf{x}_{2\emptyset})\prod_{i\in\emptyset}r_n(x_{1i},x_{2i})d\mathbf{x}_{1\emptyset}d\mathbf{x}_{2\emptyset}=f^2_{\mathbf{x},\emptyset}\ . \tag{70}$$

Finally, note that

$$\int f^2_{\mathbf{x},\emptyset}d\mathbf{x}=\underline{\tau}^2_{\mathfrak{u}}+\mathbb{E}[Y]^2 \tag{71}$$

and

$$\int\int f^2_{\mathbf{x}}(\mathbf{x}_1)d\mathbf{x}_1d\mathbf{x}=\mathbb{E}\big[Y^2\big] \tag{72}$$

and conclude that

$$\mathbb{E}\big[f(\dot{\mathbf{X}}^1)f(\dot{\mathbf{X}}^1_{\mathfrak{u}}:\dot{\mathbf{X}}^2_{\mathfrak{u}^c})\big]=\Big(\frac{n}{n-1}\Big)^{d-|\mathfrak{u}|}\int I(\mathbf{x})d\mathbf{x}=\underline{\tau}^2_{\mathfrak{u}}+\mathbb{E}[Y]^2+B_{\mathfrak{u},n} \tag{73}$$

with

$$-\mathbb{E}[Y^2]\sum_{\substack{\emptyset\neq\mathfrak{v}\subseteq\mathfrak{u}^c\\ |v|\ \text{odd}}}\frac{1}{(n-1)^{|v|}} \leq B_{\mathfrak{u},n} \leq \mathbb{E}[Y^2]\sum_{\substack{\emptyset\neq\mathfrak{v}\subseteq\mathfrak{u}^c\\ |v|\ \text{even}}}\frac{1}{(n-1)^{|v|}}\ . \tag{74}$$

□

Lemma 7. *The inequalities in Equation* (66) *imply that*

$$\Big|\sum_{\mathfrak{w}\subseteq u^c}\Big(\frac{1}{n}\Big)^{|\mathfrak{w}|}B_{\mathfrak{u}\cup\mathfrak{w},n}\Big| \le \Big(\frac{d-|\mathfrak{u}|+1}{n}+1\Big)\Big(\frac{d-|\mathfrak{u}|+1}{n-1}\Big)\mathbb{E}\big[Y^2\big]\ . \tag{75}$$

Proof. By (66), we have

$$\begin{aligned}
\sum_{\mathfrak{w}\subseteq \mathfrak{u}^c}\Big(\frac{1}{n}\Big)^{|\mathfrak{w}|}B_{\mathfrak{u}\cup\mathfrak{w},n} &\le \mathbb{E}\big[Y^2\big]\sum_{\mathfrak{w}\subseteq \mathfrak{u}^c}\Big(\frac{1}{n}\Big)^{|\mathfrak{w}|}\sum_{\emptyset\neq\mathfrak{v}\subseteq(\mathfrak{u}\cup\mathfrak{w})^c}\frac{1}{(n-1)^{|\mathfrak{w}|}}\\
&\le \mathbb{E}\big[Y^2\big]\sum_{\mathfrak{w}\subseteq \mathfrak{u}^c}\Big(\frac{1}{n}\Big)^{|\mathfrak{w}|}\Big(\Big(1+\frac{1}{n-1}\Big)^{d-|\mathfrak{u}|-|\mathfrak{w}|}-1\Big)\\
&\le \mathbb{E}\big[Y^2\big]\Big[\Big(1+\frac{1}{n-1}\Big)^{d-|\mathfrak{u}|}\sum_{\mathfrak{w}\subseteq \mathfrak{u}^c}\Big(\frac{1}{n}\Big)^{|\mathfrak{w}|}-\sum_{\mathfrak{w}\subseteq \mathfrak{u}^c}\Big(\frac{1}{n}\Big)^{|\mathfrak{w}|}\Big]\\
&\le \mathbb{E}\big[Y^2\big]\Big[\Big(1+\frac{1}{n-1}\Big)^{d-|\mathfrak{u}|}\Big(1+\frac{1}{n}\Big)^{d-|\mathfrak{u}|}-\Big(1+\frac{1}{n}\Big)^{d-|\mathfrak{u}|}\Big]
\end{aligned} \tag{76}$$

and the conclusion follows by applying twice Lemma 5. □

B.2 Proof of Proposition 2

Proof.

(i) The proof is divided into two parts. In the first one, we only consider continuous functions, and in the second one, we extend the result to the whole class of functions such that f^4 is integrable.

First part: Consistency is obvious as in Proposition 1, except for the term

$$\frac{1}{n}\sum_{j=1}^{n}\ddot{Y}^j\ddot{Y}^j_{\mathfrak{u}}\ . \tag{77}$$

So denote $\overline{\mathbf{Z}}^j$ the Latin hypercube defined by

$$\overline{\mathbf{Z}}^j = \frac{\lfloor n\ddot{\mathbf{Z}}'^j\rfloor + \mathbf{U}^j}{n} \tag{78}$$

where the $\mathbf{U}^j$'s are independent random vectors uniformly distributed in $[0,1[^d$ independent from all the permutations and shifts in the definition of $(\ddot{\mathbf{Z}}'^j)_j$, and $\lfloor\cdot\rfloor$ is the floor function. We can write

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}\ddot{Y}^j\ddot{Y}^j_{\mathfrak{u}} &= \frac{1}{n}\sum_{j=1}^{n}f(\dot{\mathbf{X}}^j_{\mathfrak{u}}:\ddot{\mathbf{Z}}^j_{\mathfrak{u}^c})f(\dot{\mathbf{X}}^j_{\mathfrak{u}}:\ddot{\mathbf{Z}}'^j_{\mathfrak{u}^c})\\
&= \frac{1}{n}\sum_{j=1}^{n}f(\dot{\mathbf{X}}^j_{\mathfrak{u}}:\ddot{\mathbf{Z}}^j_{\mathfrak{u}^c})f(\dot{\mathbf{X}}^j_{\mathfrak{u}}:\overline{\mathbf{Z}}^j_{\mathfrak{u}^c})+\frac{1}{n}\sum_{j=1}^{n}f(\dot{\mathbf{X}}^j_{\mathfrak{u}}:\ddot{\mathbf{Z}}^j_{\mathfrak{u}^c})\Big(f(\dot{\mathbf{X}}^j_{\mathfrak{u}}:\ddot{\mathbf{Z}}'^j_{\mathfrak{u}^c})-f(\dot{\mathbf{X}}^j_{\mathfrak{u}}:\overline{\mathbf{Z}}^j_{\mathfrak{u}^c})\Big)\ .
\end{aligned} \tag{79}$$

The first term on the right-hand side is an estimator as described in Section 3.1. since by the construction proposed in Eq. (78), $(\ddot{\mathbf{Z}}^j)_{j=1..n}$ and $(\overline{\mathbf{Z}}^j)_{j=1..n}$ are two independent LHS; so Proposition 1 states it converges to $\mathbb{E}[Y]^2+\underline{\tau}^2_{\mathfrak{u}}$ almost surely. The second term on the right-hand side converges to 0 since as f is bounded — by continuity on a compact — it is bounded by

$$\frac{\sup|f|}{n}\sum_{j=1}^{n}\Big|f(\dot{\mathbf{X}}^j_{\mathfrak{u}}:\ddot{\mathbf{Z}}'^j_{\mathfrak{u}^c})-f(\dot{\mathbf{X}}^j_{\mathfrak{u}}:\overline{\mathbf{Z}}^j_{\mathfrak{u}^c})\Big| \tag{80}$$

and by uniform continuity of f — due to Heine-Cantor theorem — this quantity tends to 0 as n tends to $+\infty$. Thus the sum in the right-hand side, i.e. $\frac{1}{n}\sum_{j=1}^{n} \ddot{Y}^j \ddot{Y}_{\mathfrak{u}}^j$, converges to $\mathbb{E}[Y]^2 + \underline{\tau}_{\mathfrak{u}}^2$ almost surely.

Second part: Since the space of continuous functions on $[0,1]^d$ — denoted $\mathcal{C}\big([0,1]^d\big)$ — is dense in $L^4\big([0,1]^d\big)$, let $(f_m)_{m\in\mathbb{N}^*}$ be a sequence in $\mathcal{C}\big([0,1]^d\big)$ such that $\mathbb{E}\big[|f_m(\mathbf{X}) - f(\mathbf{X})|^4\big]$ converges to 0 as m tends to $+\infty$, where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$.

Now let $\varepsilon > 0$ and $M = M(\varepsilon) \in \mathbb{N}^*$ such that

$$\mathbb{E}\Big[\big(f_M(\mathbf{X}) - f(\mathbf{X})\big)^2\Big] < \frac{\varepsilon^2}{65\ \mathbb{E}[f^2(\mathbf{X})]}. \tag{81}$$

We can write

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} \ddot{Y}^j \ddot{Y}_{\mathfrak{u}}^j &= \frac{1}{n}\sum_{j=1}^{n} f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j) f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \overline{\mathbf{Z}}_{\mathfrak{u}^c}^j) + \frac{1}{n}\sum_{j=1}^{n} f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j)\Big(f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \overline{\mathbf{Z}}_{\mathfrak{u}^c}^j)\Big) \\
&+ \frac{1}{n}\sum_{j=1}^{n} f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j)\Big(f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j})\Big) + \frac{1}{n}\sum_{j=1}^{n} f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j)\Big(f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \overline{\mathbf{Z}}_{\mathfrak{u}^c}^j) - f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \overline{\mathbf{Z}}_{\mathfrak{u}^c}^j)\Big).
\end{aligned} \tag{82}$$

As noted in the proof of (i) in Proposition 1, the first term on the right-hand side of (82) converges to $\underline{\tau}_{\mathfrak{u}}^2 + \mathbb{E}[Y]^2$ almost surely as n tends to $+\infty$ i.e.

$$\mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_1 \in \mathbb{N}^*,\ \forall n > N_1,\ \Big|\frac{1}{n}\sum_{j=1}^{n} f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j) f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \overline{\mathbf{Z}}_{\mathfrak{u}^c}^j) - \underline{\tau}_{\mathfrak{u}}^2 - \mathbb{E}[Y]^2\Big| < \frac{\varepsilon}{4}\Big) = 1\ . \tag{83}$$

Since f_M is uniformly continuous on $[0,1]^d$, we have that

$$A_n = \sup_{1\le j\le n} \big|f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \overline{\mathbf{Z}}_{\mathfrak{u}^c}^j)\big| \tag{84}$$

converges almost surely to 0 as n tends to $+\infty$. Moreover, since f is integrable, we have that $\frac{1}{n}\sum_{j=1}^{n} |f(\ddot{\mathbf{X}}_{\mathfrak{u}}^j)|$ converges to $\mathbb{E}\big[|Y|\big]$ as n tends to $+\infty$. Hence

$$\begin{aligned}
&\mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_1 \in \mathbb{N}^*,\ \forall n > N_1,\ \Big|\frac{1}{n}\sum_{j=1}^{n} f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j)\Big(f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \overline{\mathbf{Z}}_{\mathfrak{u}^c}^j)\Big)\Big| < \frac{\varepsilon}{4}\Big) \\
&\qquad \ge\ \mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_2 \in \mathbb{N}^*,\ \forall n > N_2,\ A_n \frac{1}{n}\sum_{j=1}^{n} \big|f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j)\big| < \frac{\varepsilon}{4}\Big) \\
&\qquad =\ 1\ .
\end{aligned} \tag{85}$$

For the third and the fourth terms on the right-hand side of (82), we apply twice the same proof. First the Cauchy-Schwartz inequality gives

$$\mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_3 \in \mathbb{N}^*,\ \forall n > N_3,\ \Big|\frac{1}{n}\sum_{j=1}^{n} f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j)\Big(f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j})\Big)\Big| < \frac{\varepsilon}{4}\Big)$$

$$\ge\ \mathbb{P}\Big(\forall \varepsilon > 0,\ \exists N_3 \in \mathbb{N}^*,\ \forall n > N_3,\ \Big(\frac{1}{n}\sum_{j=1}^{n} f^2(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j)\Big)^{1/2}\Big(\frac{1}{n}\sum_{j=1}^{n} \big(f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j})\big)^2\Big)^{1/2} < \frac{\varepsilon}{4}\Big). \tag{86}$$

Then note that $\frac{1}{n}\sum_{j=1}^{n} f^2(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j)$ and $\frac{1}{n}\sum_{j=1}^{n} \big(f(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^j) - f_M(\dot{\mathbf{X}}_{\mathfrak{u}}^j : \ddot{\mathbf{Z}}_{\mathfrak{u}^c}^{'j})\big)^2$ converge almost surely to $\mathbb{E}[Y^2]$ and $\mathbb{E}[(f_M(\mathbf{X}) - f(\mathbf{X}))^2]$ — where $\mathbf{X}$ is uniformly distributed on $[0,1]^d$ — respectively. And deduce that

there exists $N_4 \in \mathbb{N}^*$ such that for all $n > N_4$, we have $\frac{1}{n}\sum_{j=1}^n f^2(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^j) < 2\,\mathbb{E}[Y^2]$ and $\frac{1}{n}\sum_{j=1}^n \big(f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^j)\big)^2 < 2\,\mathbb{E}[(f_M(\mathbf{X}) - f(\mathbf{X}))^2]$ almost surely. As a consequence, deduce from Eq. (81) that

$$\begin{aligned}\mathbb{P}\Big(\forall\varepsilon>0,\ \exists N_3\in\mathbb{N}^*,\ \forall n>N_3,\ \Big|\frac{1}{n}\sum_{j=1}^n f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^j)\Big(f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^{'j})\Big)\Big| < \frac{\varepsilon}{4}\Big) \\ \geq\ \mathbb{P}\Big(\forall\varepsilon>0,\ \exists N_3>N_4,\ \forall n>N_3,\ \varepsilon\sqrt{\frac{4}{65}} < \frac{\varepsilon}{4}\Big) \\ =\ 1 \end{aligned} \tag{87}$$

Finally, Eqs. (83–87) gives

$$\mathbb{P}\Big(\forall\varepsilon>0,\ \exists N\in\mathbb{N}^*,\ \forall n>N,\ \Big|\frac{1}{n}\sum_{j=1}^n \ddot{Y}^j\ddot{Y}_{\mathrm{u}}^j\Big| < \varepsilon\Big) = 1 \tag{88}$$

and we have the conclusion.

(ii) As in (i), the only term to treat is

$$\frac{1}{n}\sum_{j=1}^n \ddot{Y}^j\ddot{Y}_{\mathrm{u}}^j\ , \tag{89}$$

so asymptotic normality is shown in the same way by using the decomposition in (79). We always obtain the sum of a term already considered in Section 3 which converges in law to a normal distribution and a term which converges to 0 in probability, and the conclusion follows from Slutsky's lemma. We only detail the proof for $\underline{\widetilde{S}}_{\mathrm{u},n}^{RLHS}$, it is exactly the same for $\underline{\widehat{S}}_{\mathrm{u},n}^{RLHS}$. So note that following the proof of (ii) in Proposition 1 and the notation above, it is sufficient to show that

$$\sqrt{n}\Big(\frac{1}{n}\sum_{j=1}^n \big(f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^j) - \mathbb{E}[Y]\big)\big(f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \overline{\mathbf{Z}}_{\mathrm{u}^c}^j)\big)\Big) \xrightarrow[n\to\infty]{\mathbb{P}} 0 \tag{90}$$

to prove the asymptotic normality of $\underline{\widetilde{S}}_{\mathrm{u},n}^{RLHS}$.

So consider $\varepsilon, \eta > 0$ and prove that there exists $N \in \mathbb{N}^*$ such that for all $n > N$, the quantity

$$P = \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{j=1}^n \big(f(\ddot{\mathbf{X}}_{\mathrm{u}}^j) - \mathbb{E}[Y]\big)\big(f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \overline{\mathbf{Z}}_{\mathrm{u}^c}^j)\big)\Big| > \varepsilon\Big) \tag{91}$$

is less than η. First as f^6 is integrable, there exists a constant $K > 0$ such that $\mathbb{P}(|f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^j)| > K) < \eta/4$. Hence

$$\begin{aligned} P \ &\leq\ \mathbb{P}\Big(\Big(\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^n \big(f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^j) - \mathbb{E}[Y]\big)\big(f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \overline{\mathbf{Z}}_{\mathrm{u}^c}^j)\big)\Big| > \varepsilon\Big)\bigcap\big(|f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^j)| \leq K\big)\Big) \\ &\quad +\ \mathbb{P}\Big(\Big(\Big|\frac{1}{\sqrt{n}}\sum_{j=1}^n \big(f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^j) - \mathbb{E}[Y]\big)\big(f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \overline{\mathbf{Z}}_{\mathrm{u}}^j)\big)\Big| > \varepsilon\Big)\bigcap\big(|f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^j)| > K\big)\Big) \\ &<\ \mathbb{P}\Big(\frac{K + |\mathbb{E}[Y]|}{\sqrt{n}}\sum_{j=1}^n \big|f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \ddot{\mathbf{Z}}_{\mathrm{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathrm{u}}^j : \overline{\mathbf{Z}}_{\mathrm{u}^c}^j)\big| > \varepsilon\Big) + \frac{\eta}{4}. \end{aligned} \tag{92}$$

Now note that the space of continuous functions on $[0,1]^d$, denoted by $\mathcal{C}([0,1]^d)$, is dense in $L^6([0,1]^d)$ and let $(f_m)_{m\in\mathbb{N}^*}$ be a sequence in $\mathcal{C}([0,1]^d)$ such that $\mathbb{E}[|f_m(\mathbf{X}) - f(\mathbf{X})|^6]$ converges to 0 as m tends to $+\infty$ where $\mathbf{X}$ is

uniformly distributed on $[0,1]^d$. It is easy to note that there exists $M = M(n)$ such that $\mathbb{P}(|f_M(\mathbf{X}) - f(\mathbf{X})| > 1/n) < \eta/4$. Thus we get from Eq. (92) that

$$\begin{aligned} P &< \sum_{i=1}^{4} \mathbb{P}\Big(\Big(\frac{K + |\mathbb{E}[Y]|}{\sqrt{n}} \sum_{j=1}^{n} \big(|f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j)| + |f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j})| \\ &+ |f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j) - f(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j)| \big) \Big) \bigcap A_i \Big) + \frac{\eta}{4} \end{aligned} \tag{93}$$

where

$$A_1 = \big(|f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j) - f(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j)| > \frac{1}{n} \big) \cap \big(|f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j})| > \frac{1}{n} \big) \tag{94}$$

$$A_2 = \big(|f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j) - f(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j)| > \frac{1}{n} \big) \cap \big(|f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j})| < \frac{1}{n} \big) \tag{95}$$

and A_3 and A_4 are the complementary events of A_1 and A_2, respectively. So we deduce

$$\begin{aligned} P &< \mathbb{P}\Big(\Big(\frac{K + |\mathbb{E}[Y]|}{\sqrt{n}} \sum_{j=1}^{n} \big(|f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j)| + |f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j})| \\ &+ |f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j)| \big) \Big) \bigcap A_3 \Big) + \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_4) + \frac{\eta}{4} \\ &< \mathbb{P}\Big(\frac{K + |\mathbb{E}[Y]|}{\sqrt{n}} \Big(2 + \sum_{j=1}^{n} |f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j)| \Big) > \varepsilon \Big) + \eta. \end{aligned} \tag{96}$$

Now by an other density argument, note that there exists a sequence of Lipschitz continuous functions with constant 1, denoted $(f_{M,q})_{q \in \mathbb{N}^*}$, such that $\sup_{[0,1]^d} |f_{M,q}(\mathbf{x}) - f_M(\mathbf{x})|$ converges to 0 as q tends to $+\infty$. Then there exists $Q = Q(n) \in \mathbb{N}^*$ such that $\sup_{[0,1]^d} |f_{M,Q}(\mathbf{x}) - f_M(\mathbf{x})| < 1/n$ and deduce that

$$\begin{aligned} P &< \mathbb{P}\Big(\frac{K + |\mathbb{E}[Y]|}{\sqrt{n}} \Big(2 + \sum_{j=1}^{n} \big(|f_{M,Q}(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f_{M,Q}(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j)| + |f_{M,Q}(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) - f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j})| \\ &+ |f_{M,Q}(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j) - f_M(\dot{\mathbf{X}}_{\mathbf{u}}^j : \overline{\mathbf{Z}}_{\mathbf{u}^c}^j)| \big) \Big) > \varepsilon \Big) + \eta \\ &< \mathbb{P}\Big(\frac{5(K + |\mathbb{E}[Y]|)}{\sqrt{n}} > \varepsilon \Big) + \eta. \end{aligned} \tag{97}$$

and the conclusion follows.

(iii) First note that since $(\dot{\mathbf{X}}^j)_j \sim \mathcal{LH}(n,d)$ and $(\ddot{\mathbf{Z}}^j, \ddot{\mathbf{Z}}^{'j})_j \sim \mathcal{RLH}(n,d)$, we have

$$\ddot{Y}^j = f(\dot{\mathbf{X}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^j) \text{ and } \ddot{Y}_{\mathbf{u}}^j = f(\dot{\mathbf{X}}_{\mathbf{u}}^j : \ddot{\mathbf{Z}}_{\mathbf{u}^c}^{'j}) \tag{98}$$

with

$$\dot{\mathbf{X}}^j = \Big(\frac{\pi_1''(j) - U'_{1,\pi_1''(j)}}{n}, \ldots, \frac{\pi_d''(j) - U'_{d,\pi_d''(j)}}{n} \Big) \tag{99}$$

$$\ddot{\mathbf{Z}}^j = \Big(\frac{\pi_1(j) - U_{1,\pi_1(j)}}{n}, \ldots, \frac{\pi_d(j) - U_{d,\pi_d(j)}}{n} \Big) \tag{100}$$

and

$$\ddot{\mathbf{Z}}^{'j} = \Big(\frac{\pi_1'(j) - U_{1,\pi_1'(j)}}{n}, \ldots, \frac{\pi_d'(j) - U_{d,\pi_d'(j)}}{n} \Big) \tag{101}$$

where the π_is, the π'_is, the π''_is, the $U_{i,j}$s and the $U'_{i,j}$s are independent random variables uniformly distributed on Π_n — see Definition 1 —, Π_n, Π_n, $[0,1]$ and $[0,1]$, respectively. Moreover note that if for an index i in $\{1,\dots,d\}$, we have $\pi_i(j) = \pi'_i(j)$ then $\ddot{Z}_i^j = \ddot{Z}_i^{'j}$; and if $\pi_i(j) \neq \pi'_i(j)$ then $U_{i,\pi_i(j)}$ and $U_{i,\pi'_i(j)}$ are independent and therefore $\ddot{Z}_i^j$ and $\ddot{Z}_i^{'j}$ are two distinct points of a Latin hypercube of size n in $[0,1]$. For $j \in \{1,\dots,n\}$, denote by $\mathfrak{e}(j)$ the set of integers $i \in \mathfrak{u}^c$ such that $\pi_i(j) = \pi'_i(j)$. Thus we have

$$\mathbb{E}\big[\ddot{Y}^j\ddot{Y}^j_{\mathfrak{u}}\big] = \frac{1}{n^{2d-|\mathfrak{u}|}} \sum_{\mathfrak{w}\subseteq\mathfrak{u}^c} \sum_{\substack{\boldsymbol{\pi}''_{\mathfrak{u}}(j)\in\\ \{1,\dots,n\}^{|\mathfrak{u}|}}} \sum_{\substack{\boldsymbol{\pi}_{\mathfrak{u}^c}(j)\in\\ \{1,\dots,n\}^{d-|\mathfrak{u}|}}} \sum_{\substack{\boldsymbol{\pi}'_{\mathfrak{u}^c}(j)\in\\ \{1,\dots,n\}^{d-|\mathfrak{u}|}}} \mathbf{1}_{\{\mathfrak{e}(j)=\mathfrak{w}\}} \cdots$$

$$\int f\Big(\dots, \underbrace{\frac{\pi''_i(j)-u_{1i}}{n}}_{i\in\mathfrak{u}}, \dots, \underbrace{\frac{\pi_k(j)-u_{1k}}{n}}_{k\in\mathfrak{u}^c}, \dots\Big) f\Big(\dots, \underbrace{\frac{\pi''_i(j)-u_{1i}}{n}}_{i\in\mathfrak{u}}, \dots, \underbrace{\frac{\pi_k(j)-u_{1k}}{n}}_{k\in\mathfrak{u}^c\cap\mathfrak{w}}, \dots, \underbrace{\frac{\pi'_l(j)-u_{2l}}{n}}_{l\in\mathfrak{u}^c\cap\mathfrak{w}^c}, \dots\Big) d\mathbf{u}_1 d\mathbf{u}_{2(\mathfrak{u}\cup\mathfrak{w})^c} \tag{102}$$

where for all $i \in \mathfrak{u}\cup\mathfrak{e}(j)$, $\pi_i(j) = \pi'_i(j)$ and $u_{1i}(j) = u_{2i}(j)$. And noting that

$$\frac{1}{(n-1)^{d-|\mathfrak{u}|-|\mathfrak{w}|}n^d} \sum_{\substack{\boldsymbol{\pi}''_{\mathfrak{u}}(j)\in\\ \{1,\dots,n\}^{|\mathfrak{u}|}}} \sum_{\substack{\boldsymbol{\pi}_{\mathfrak{u}^c}(j)\in\\ \{1,\dots,n\}^{d-|\mathfrak{u}|}}} \sum_{\substack{\boldsymbol{\pi}'_{\mathfrak{u}^c}(j)\in\\ \{1,\dots,n\}^{d-|\mathfrak{u}|}}} \mathbf{1}_{\{\mathfrak{e}(j)=\mathfrak{w}\}} \cdots$$

$$\int f\Big(\dots, \underbrace{\frac{\pi''_i(j)-u_{1i}}{n}}_{i\in\mathfrak{u}}, \dots, \underbrace{\frac{\pi_k(j)-u_{1k}}{n}}_{k\in\mathfrak{u}^c}, \dots\Big) f\Big(\dots, \underbrace{\frac{\pi''_i(j)-u_{1i}}{n}}_{i\in\mathfrak{u}}, \dots, \underbrace{\frac{\pi_k(j)-u_{1k}}{n}}_{k\in\mathfrak{u}^c\cap\mathfrak{w}}, \dots, \underbrace{\frac{\pi'_l(j)-u_{2l}}{n}}_{l\in\mathfrak{u}^c\cap\mathfrak{w}^c}, \dots\Big) d\mathbf{u}_1 d\mathbf{u}_{2(\mathfrak{u}\cup\mathfrak{w})^c} \tag{103}$$

is equal to $\mathbb{E}\big[f(\dot{\mathbf{X}}^1)f(\dot{\mathbf{X}}^1_{\mathfrak{u}\cup\mathfrak{w}}, \dot{\mathbf{X}}^2_{(\mathfrak{u}\cup\mathfrak{w})^c})\big]$, Lemma 6 gives

$$\mathbb{E}\big[\ddot{Y}^j\ddot{Y}^j_{\mathfrak{u}}\big] = \sum_{\mathfrak{w}\subseteq\mathfrak{u}^c} \Big(\frac{1}{n}\Big)^{|\mathfrak{w}|} \Big(\mathbb{E}[Y]^2 + \underline{\tau}^2_{\mathfrak{u}\cup\mathfrak{w}} + B_{\mathfrak{u}\cup\mathfrak{w},n}\Big) \,. \tag{104}$$

By Lemmas 5 and 7, and noting that $\mathbb{E}[Y]^2 + \underline{\tau}^2_{\mathfrak{u}\cup\mathfrak{w}} \leq \mathbb{E}[Y^2]$, we obtain

$$\mathbb{E}\big[\ddot{Y}^j\ddot{Y}^j_{\mathfrak{u}}\big] = \mathbb{E}[Y]^2 + \underline{\tau}^2_{\mathfrak{u}} + B_{|\mathfrak{u}|,n} \tag{105}$$

where

$$\big|B_{|\mathfrak{u}|,n}\big| \leq \Big(\frac{d-|\mathfrak{u}|+1}{n} + 2\Big)\Big(\frac{d-|\mathfrak{u}|+1}{n-1}\Big)\mathbb{E}[Y^2] \,. \tag{106}$$

Following the same proof, it is easy to show that for $j \neq l$, we have

$$\mathbb{E}\big[\ddot{Y}^j\ddot{Y}^l_{\mathfrak{u}}\big] = \mathbb{E}[Y]^2 + B_{n,1} \tag{107}$$

where

$$\big|B_{n,1}\big| \leq \Big(\frac{d+1}{n} + 2\Big)\Big(\frac{d+1}{n-1}\Big)\mathbb{E}[Y^2] \,. \tag{108}$$

Thus noting that

$$\mathbb{E}\big[\underline{\mathcal{T}}^{2,RLHS}_{\mathfrak{u},n}\big] = \frac{n-1}{n}\mathbb{E}[\ddot{Y}^1\ddot{Y}^1_{\mathfrak{u}}] - \frac{n-1}{n}\mathbb{E}[\ddot{Y}^1\ddot{Y}^2_{\mathfrak{u}}] \tag{109}$$

we conclude that

$$\mathbb{E}\big[\widetilde{\underline{\tau}}^{2,RLHS}_{\mathfrak{u},n}\big] = \underline{\tau}^2_{\mathfrak{u}} - \frac{1}{n}\underline{\tau}^2_{\mathfrak{u}} + \frac{n-1}{n}\big(B_{n,1} + B_{|\mathfrak{u}|,n}\big) \tag{110}$$

where the biases are $O(n^{-1})$ as specified above. Concerning $\widetilde{\sigma}_n^{2,RLHS}$, note that $\widetilde{\sigma}_n^{2,RLHS} = \widetilde{\sigma}_n^{2,LHS}$ and the conclusion follows from (iii) in Proposition 1. Concerning $\widehat{\underline{\tau}}_{\mathrm{u},n}^{2,RLHS}$ and $\widehat{\sigma}_n^{2,RLHS}$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\ddot{Y}^j+\ddot{Y}_{\mathrm{u}}^j}{2}\right)^2\right] &= \frac{1}{4n}\mathbb{E}\big[(\ddot{Y}^1+\ddot{Y}_{\mathrm{u}}^1)^2\big]+\frac{1}{4n^2}\sum_{j=1}^{n}\sum_{\substack{l=1\\l\neq j}}^{n}\mathbb{E}\big[(\ddot{Y}^j+\ddot{Y}_{\mathrm{u}}^j)(\ddot{Y}^l+\ddot{Y}_{\mathrm{u}}^l)\big]\\ &= \frac{1}{2n}\Big(\mathbb{E}\big[(\ddot{Y}^1)^2\big]+\mathbb{E}[\ddot{Y}^1\ddot{Y}_{\mathrm{u}}^1]\Big)+\frac{n-1}{2n}\Big(\mathbb{E}[\ddot{Y}^1\ddot{Y}^2]+\mathbb{E}[\ddot{Y}^1\ddot{Y}_{\mathrm{u}}^2]\Big). \qquad (111)\end{aligned}$$

Then using notation at the beginning of Section 3.1., note that

$$\mathbb{E}\big[\ddot{Y}^1\ddot{Y}^2\big]=\mathbb{E}\big[\dot{Y}^1\dot{Y}^2\big]=\mathrm{Cov}\big(\dot{Y}^1,\dot{Y}^2\big)+\mathbb{E}[Y]^2=\mathrm{Cov}\big(\dot{Y}^1,\dot{Y}^2_{\{1,\ldots,d\}}\big)+\mathbb{E}[Y]^2 \qquad (112)$$

and by (106), (108) and Lemma 4, we deduce

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}\frac{\ddot{Y}^j+\ddot{Y}_{\mathrm{u}}^j}{2}\right)^2\right]=\frac{1}{2n}\underline{\tau}_{\mathrm{u}}^2+\mathbb{E}[Y]^2+B_{n,1}+B_{n,2}\ . \qquad (113)$$

where

$$|B_{n,2}|\leq\frac{\sigma^2}{2n} \qquad (114)$$

and $B_{n,1}$ is specified in (108). Then it is easy to conclude that

$$\begin{aligned}\mathbb{E}\big[\widehat{\underline{\tau}}_{\mathrm{u},n}^{2,RLHS}\big] &= \underline{\tau}_{\mathrm{u}}^2-\frac{1}{2n}\underline{\tau}_{\mathrm{u}}^2+B_{n,1}+B_{n,2}+\frac{n-1}{n}B_{|\mathrm{u}|,n} \qquad (115)\\ \mathbb{E}\big[\widehat{\sigma}_n^{2,RLHS}\big] &= \sigma^2-\frac{1}{2n}\underline{\tau}_{\mathrm{u}}^2+B_{n,1}+B_{n,2} \qquad (116)\end{aligned}$$

where the biases are $O(n^{-1})$ as specified above. □

C Phytoplankton growth model

The phytoplankton growth is given by the five following equations, where A stands for `pp`, `np` or `mp`.

$$\begin{aligned}\mu_A &= \mu_{max\mathrm{A}}\,\mathrm{lim_{IA}}\,\mathrm{lim_{TA}}\,(\mathrm{lim_{NO_3A}}+\mathrm{lim_{NH_4A}})\\ \mathrm{lim_{NO_3A}} &= \left(\frac{\mathrm{NO_3}}{\mathrm{NO_3}+\mathrm{K_{NO_3A}}}\right)\exp(-\Psi\mathrm{NH_4})\\ \mathrm{lim_{NH_4A}} &= \left(\frac{\mathrm{NH_4}}{\mathrm{NH_4}+\mathrm{K_{NH_4A}}}\right)\\ \mathrm{lim_{IA}} &= \frac{2(1+\beta_{\mathrm{IA}})\frac{\mathrm{PAR}}{I_{opt\mathrm{A}}}}{\left(\frac{\mathrm{PAR}}{I_{opt\mathrm{A}}}\right)^2+2\beta_{\mathrm{IA}}\frac{\mathrm{PAR}}{I_{opt\mathrm{A}}}+1}\\ \mathrm{lim_{TA}} &= \max\left(\frac{2(1+\beta_{\mathrm{TA}})\frac{T-T_{let\mathrm{A}}}{T_{opt\mathrm{A}}-T_{let\mathrm{A}}}}{\left(\frac{T-T_{let\mathrm{A}}}{T_{opt\mathrm{A}}-T_{let\mathrm{A}}}\right)^2+2\beta_{\mathrm{IA}}\frac{T-T_{let\mathrm{A}}}{T_{opt\mathrm{A}}-T_{let\mathrm{A}}}+1},0\right)\end{aligned}$$

where the parameters are defined in the following table

parameter	definition
μ_A	growth rate of A
$\mu_{max\mathrm{A}}$	maximum growth rate of A
$\mathrm{lim}_{\mathrm{NO_3A}}$	limitation by NO_3 for A
$\mathrm{lim}_{\mathrm{NH_4A}}$	limitation by NH_4 for A
$\mathrm{K}_{\mathrm{NO_3A}}$	half-saturation coefficient of NO_3 for A
$\mathrm{K}_{\mathrm{NH_4A}}$	half-saturation coefficient of NH_4 for A
Ψ	inhibition coefficient by NH_4
NO_3	NO_3 concentration
NH_4	NH_4 concentration
$\mathrm{lim}_{\mathrm{IA}}$	limitation by light for A
β_{IA}	shape factor for photoinhibition curve
$I_{opt\mathrm{A}}$	optimum insolation for A
PAR	photosynthetic active radiation
$\mathrm{lim}_{\mathrm{TA}}$	limitation by temperature for A
β_{TA}	shape factor for thermoinhibition curve
$T_{opt\mathrm{A}}$	optimum temperature for A
$T_{let\mathrm{A}}$	lower lethal temperature for A
T	temperature

Table 7: Parameters of the phytoplankton growth model

References

Blatman, G. and Sudret, B. (2010). Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. *Reliability Engineering and System Safety*, 95:1216–1229.

Bose, R. (1938). On the application of the theory of galois fields to the problem of construction of hyper-graeco-latin squares. *Sankhya*, 3:323–338.

Cukier, R. I., Fortuin, C. M., Shuler, K. E., Petschek, A. G., and Schaibly, J. H. (1973). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients: Theory. *Journal of Chemical Physics*, 59:3873–3878.

Cukier, R. I., Levine, H. B., and Shuler, K. E. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 26:1–42.

Devroye, L. (1986). *Non-uniform random variate generation*. Springer-Verlag, New York.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, 9(3):586–596.

Hoeffding, W. F. (1948). A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19:293–325.

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

Ishigami, T. and Homma, T. (1990). An importance quantification technique in uncertainty analysis for computers models. *First International Symposium on Uncertainty Modeling and Analysis Proceedings*, pages 398–403.

Janon, A., Klein, T., Lagnoux, A., Nodet, M., and Prieur, C. (2012+). Asymptotic normality and efficiency of two sobol index estimators. *Preprint available at http://hal.inria.fr/docs/00/66/50/48/PDF/ArtAsymptSobol.pdf*.

Lacroix, G. and Nival, P. (1998). Influence of meteorogical variability on primary production dynamics in the ligurian sea (nw mediterranean sea) with 1d hydrodynamic/biological model. *Journal of Marine Systems*, 37:229–258.

Loh, W. L. (1996). On latin hypercube sampling. *The Annals of Statistics*, 24(5):2058–2080.

Loh, W. L. (2008). A multivariate central limit theorem for randomized orthogonal array sampling designs in computer experiments. *The Annals of Statistics*, 36:1983–2023.

McKay, M. D. (1995). Evaluating prediction uncertainty. *Technical Report NUREG/CR-6311, US Nuclear Regulatory Commission and Los Alamos National Laboratory*, pages 1–79.

McKay, M. D., Conover, W. J., and Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245.

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

Niederreiter, H. (1992). Random number generation and quasi-monte carlo methods. *SIAM, Philadelphia*.

Owen, A. B. (1992). Orthogonal arrays for computer experiments, integration and visualization. *Statistica Sinica*, 2:439–452.

Owen, A. B. (1995). Randomly permuted (t,m,s)-nets and (t,s)-sequences. In Niederreiter, H. and Shiue, P. J.-S., editors, *Monte Carlo and Quasi-Monte CArlo Methods in Scientific Computing*, pages 299–317. Springer-Verlag, New York.

Owen, A. B. (1997a). Monte carlo variance of scrambled equidistribution quadrature. *SIAM J. Numer. Anal.*, 34(5):1884–1910.

Owen, A. B. (1997b). Scrambled net variance for integrals of smooth functions. *Ann. Statist.*, 25(4):1541–1562.

Owen, A. B. (2012a). Better estimation of small sobol' sensitivity indices. *ACM TOMACS*, 23(2).

Owen, A. B. (2012+b). Variance components and generalized sobol' indices. *Preprint available at http://statistics.stanford.edu/∼ckirby/techreports/GEN/2012/2012-07.pdf*.

Qian, P. Z. G. (2009). Nested latin hypercube sampling. *Biometrika*, 96(4):957–970.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145:280–297.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. (2010). Variance-based sensitivity analysis of model output. design and estimator for the total index. *Computer Physics Communications*, 181:259–270.

Saltelli, A., Chan, K., and Scott, M. (2000). *Sensitivity Analysis*. John Wiley & Sons.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Saisana, D. G. M., and Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*. John Wiley & Sons.

Saltelli, A. and Sobol', I. M. (1995). About the use of rank transformation in sensitivity analysis of a model. *Reliability Engineering and System Safety*, 50:225–239.

Saltelli, A., Tarantola, S., and Chan, K. P. S. (1999). A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics*, 41:39–56.

Sobol', I. M. (1993). Sensitivity analysis for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414.

Stanley, R. P. (2012). *Enumerative Combinatorics, Volume 1 (2nd edition)*. Cambridge University Press.

Stein, M. (1987). Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151.

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

Tang, B. (1993). Orthogonal array-based latin hypercubes. *J. Amer. Statist. Assoc.*, 88:1392–1397.

Tarantola, S., Gatelli, D., and Mara, T. A. (2006). Random balance designs for the estimation of first-order global sensitivity indices. *Reliability Engineering and System Safety*, 91:717–727.

Tissot, J. Y. and Prieur, C. (2012+). Variance-based sensitivity analysis using harmonic analysis. *Preprint avalaible at http://hal.archives-ouvertes.fr/docs/00/68/07/25/PDF/ FAST_RBD_revisited.pdf*.

Van der Vaart, A. W. (1998). *Asymptotics Statistics*. Cambridge University Press.