HAL open science

Global Sensitivity Analysis: a novel generation of mighty estimators based on rank statistics

Fabrice Gamboa, Pierre Gremaud, Thierry Klein, Agnès Lagnoux

To cite this version:

Fabrice Gamboa, Pierre Gremaud, Thierry Klein, Agnès Lagnoux. Global Sensitivity Analysis: a novel generation of mighty estimators based on rank statistics. Bernoulli, 2022, 28 (4), pp.2345-2374. 10.3150/21-BEJ1421 . hal-02474902v6

HAL Id: hal-02474902

https://hal.science/hal-02474902v6

Submitted on 30 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Global Sensitivity Analysis: a novel generation of mighty estimators based on rank statistics

Fabrice Gamboa[1], Pierre Gremaud[2], Thierry Klein[3], and Agnès Lagnoux[4]

[1]Institut de Mathématiques de Toulouse and ANITI; UMR5219. Université de Toulouse; CNRS. UT3, F-31062 Toulouse, France.
[2]Department of Mathematics. NC State University. Raleigh, North Carolina 27695, USA.
[3]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; ENAC - Ecole Nationale de l'Aviation Civile , Université de Toulouse, France
[4]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse; CNRS. UT2J, F-31058 Toulouse, France.

June 30, 2023

Key words Global sensitivity analysis, Cramér-von-Mises distance, Pick-Freeze method, Chatterjee's coefficient of correlation, Sobol indices estimation.

AMS subject classification 62G05, 62G20, 62G30.

Abstract

We propose a new statistical estimation framework for a large family of global sensitivity analysis indices. Our approach is based on rank statistics and uses an empirical correlation coefficient recently introduced by Chatterjee [9]. We show how to apply this approach to compute not only the Cramér-von-Mises indices, directly related to Chatterjee's notion of correlation, but also first-order Sobol' indices, general metric space indices and higher-order moment indices. We establish consistency of the resulting estimators and demonstrate their numerical efficiency, especially for small sample sizes. In addition, we prove a central limit theorem for the estimators of the first-order Sobol' indices.

1 Introduction

The use of complex computer models for the analysis of applications from the sciences, engineering and other fields is by now routine. Often, the models are expensive to run in terms of computational time. It is thus crucial to understand, with just a few runs, the

global influence of one or several inputs on the output of the system under study [33]. When these inputs are regarded as random elements, this problem is generally referred to as Global Sensitivity Analysis (GSA). We refer to [12, 32, 35] for an overview of the practical aspects of GSA.

A popular and highly useful tool to quantify input influence is the Sobol' indices. These indices were first introduced in [36] and are well tailored to the case of scalar outputs (and even to the case of vectorial and functional outputs). Thanks to the Hoeffding decomposition [24], the Sobol' indices compare the conditional variance of the output knowing some of the input variables to the total variance of the output. Since Sobol' indices are variance based, they only quantify the second-order influence of the inputs. Many authors proposed other criteria to compare the conditional distribution of the output knowing some of the inputs to the distribution of the output (see, e.g., higher moments indices in [29, 31, 30], indices using divergences or distances between measures in [4, 5, 10], goal-oriented indices using contrast functions in [15], distribution-based indices as Cramér-von-Mises indices in [19]).

Many different estimation procedures of the Sobol' indices have been proposed and studied. Some estimation procedures are based on different designs of experiment using for example polynomial chaos (see [37] and the reference therein for more details). Some other natural procedures are based on Monte-Carlo or quasi Monte-Carlo design of experiments (see [26, 29] and references therein for more details). In particular, an efficient estimation of the Sobol' indices can be performed through the so-called Pick-Freeze method. See Section 2.1 below for its description. Observe that the Pick-Freeze estimation procedure allows the estimation of several sensitivity indices: the classical Sobol' indices for real-valued outputs, as well as their generalization for vectorial-valued codes, but also the indices based on higher moments [31] and the Cramér-von-Mises indices which take into account on the whole distribution (see [19, 16] and Section 2.2 below for more details on such indices). In addition, the Pick-Freeze estimators have desirable statistical properties such as consistency, central limit theorem (CLT) with a rate of convergence in $\sqrt{n}$, concentration inequalities and Berry-Esseen bounds, and asymptotic efficiency (see [25, 18] and Section 2.1 below for more details). However, the Pick-Freeze scheme has two major drawbacks. First, it relies on a particular experimental design that may be unavailable in practice. Second, its cost may be prohibitive when estimating several indices. Naturally, the cost of an estimator depends on the cost of each evaluation of the code and on the number of evaluations. The number of model calls to estimate all first-order Sobol' indices grows linearly with the number of input parameters. For example, if we consider $p = 99$ input parameters and only $n = 1000$ calls are allowed, then only a sample of size $n/(p + 1) = 10$ is available to estimate each single first-order Sobol' index. It is a poor amount of information to get a satisfying estimation of the Sobol' indices.

In a recent work [9], Chatterjee studies the dependence between two variables by introducing an empirical correlation coefficient based on rank statistics, see Section 3.1 below for the precise definition. Further, the quantification of the dependence has also been investigated in the bivariate case (namely, in the copula setting), see [38, 13, 3]. The striking point of [9] is that this empirical correlation coefficient converges almost surely (a.s.) to the Cramér-von-Mises index priorly introduced in [19] as the sample size goes to infinity.

In this paper, we show how to embed Chatterjee's method in the GSA framework, thereby eliminating the two drawbacks of the classical Pick-Freeze estimation mentioned above.

Thus no particular design of experiment is needed for the estimation that can be done with a unique n-sample. In addition, we generalize Chatterjee's approach to allow the estimation of a large class of GSA indices which includes the Sobol' indices and the higher-order moment indices proposed by Owen [29, 31, 30] (see Section 2.1 below). Using a single sample of size n, it is now possible to estimate at the same time all the first-order Sobol' indices, the Cramér-von-Mises indices, and other useful sensitivity indices. Furthermore, we show that this new procedure provides estimators also converging at rate $\sqrt{n}$ by proving a CLT in the estimation of the first-order Sobol' indices.

The paper is organized as follows. In Section 2, we recall the context of GSA, the definition of the Sobol' indices and Cramér-von-Mises indices, and their classical Pick-Freeze estimations. Section 3 focuses on Chatterjee's method, called rank-based method in this paper. More precisely, we show how the Cramér-von-Mises indices can be also estimated using the rank-based method (Section 3.1) and we present its generalization to estimate sensitivity indices together with the consistency of the estimation procedure (Section 3.2). Section 4 is dedicated to Sobol' indices. We prove the asymptotic normality of their estimators based on rank statistics. In addition, we propose a comparison of the different estimation procedures in Section 4.3 while Section 4.4 considers other classical sensitivity indices. Section 5 is dedicated to a numerical comparison between the Pick-Freeze estimation procedure and the rank-based method. We first compare the numerical performances of both estimators on a linear model. Finally, we consider a real life application. As expected, the rank-based estimation method outperforms the classical Pick-Freeze procedure, even for small sample sizes (which are common in practice). Conclusions and perspectives are offered in Section 6.

After a first submission of this paper, we have been aware of the very nice work of Broto *et al* [8] concerning the statistical estimation of Shapley effect where the use of closest neighbors is also put in action to built consistent estimates. We also notice that there is actually a strong scientific interest around asymptotic behavior for the statistical method introduced in [9]. Indeed, during the revision of this paper, we have a look on the very nice paper [2] where an asymptotic contiguity study is performed.

2 Global sensitivity analysis and Pick-Freeze estimation

2.1 Sobol' indices

Context and definition of the Sobol' indices The quantity of interest (QoI) Y is obtained from the numerical code and is regarded as a function f of the vector of the distributed input $(X_i)_{i=1,\dots,p}$

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

where f is defined on the state space $E_1 \times \dots \times E_p$, $X_i \in E_i$, $i = 1, \dots, p$. Classically, the X_i's are assumed to be independent random variables and a sensitivity analysis is performed using the Hoeffding decomposition [1, 39] leading to the standard Sobol' indices [35]. This assumption is made throughout the paper, unless explicitly stated otherwise. More precisely, assume f to be real-valued and square integrable and let $\mathbf{u}$ be a subset of $\{1, \dots, p\}$ and $\sim\mathbf{u}$ its complementary set in $\{1, \dots, p\}$. Setting $X_{\mathbf{u}} = (X_i, i \in \mathbf{u})$ and

$X_{\sim\mathbf{u}} = (X_i, i \in \sim \mathbf{u})$, the corresponding Sobol' indices take the form

$$S^{\mathbf{u}} = \frac{\mathrm{Var}\left(\mathbb{E}[Y|X_{\mathbf{u}}]\right)}{\mathrm{Var}(Y)} \quad \text{and} \quad S^{\sim\mathbf{u}} = \frac{\mathrm{Var}\left(\mathbb{E}[Y|X_{\sim\mathbf{u}}]\right)}{\mathrm{Var}(Y)}. \tag{2}$$

By definition, the Sobol' indices quantify the fluctuations of the output Y around its mean. When the practitioner is not interested in the mean behavior of Y but rather in its median, in its tail, or even in its quantiles, the Sobol' indices become less appropriate to quantify sensitivity. GSA must then be performed in a framework which takes into account more than one specific moment, such as the variance for Sobol' indices.

Pick-Freeze estimation procedure of the Sobol' indices A Monte-Carlo scheme can be used to estimate the Sobol' indices. The corresponding Pick-Freeze approach from [18, 19, 25] relies on expressing the variances of the conditional expectations in terms of covariances which are easily and well estimated by their empirical versions. To that end, we define, for any subset $\mathbf{u}$ of $\{1, \ldots, p\}$

$$Y^{\mathbf{u}} := f(X^{\mathbf{u}}). \tag{3}$$

where $X^{\mathbf{u}}$ is such that $X^{\mathbf{u}}_{\mathbf{u}} = X_{\mathbf{u}}$ and $X^{\mathbf{u}}_i = X'_i$ if $i \in \sim \mathbf{u}$, X'_i being an independent copy of X_i. The estimation procedure relies on the following result

$$\mathrm{Var}(\mathbb{E}[Y|X_{\mathbf{u}}]) = \mathrm{Cov}(Y, Y^{\mathbf{u}}). \tag{4}$$

The reader is referred to [25, Lemma 1.2] for its proof.
The natural estimator of $S^{\mathbf{u}}$ is then given by

$$S^{\mathbf{u}}_n = \frac{\frac{1}{n}\sum_{j=1}^n Y_j Y^{\mathbf{u}}_j - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)\left(\frac{1}{n}\sum_{j=1}^n Y^{\mathbf{u}}_j\right)}{\frac{1}{n}\sum_{j=1}^n (Y_j)^2 - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}. \tag{5}$$

A slightly different estimator that uses all the information available is introduced in [25]:

$$T^{\mathbf{u}}_n = \frac{\frac{1}{n}\sum_{j=1}^n Y_j Y^{\mathbf{u}}_j - \left(\frac{1}{n}\sum_{j=1}^n \frac{Y_j + Y^{\mathbf{u}}_j}{2}\right)^2}{\frac{1}{n}\sum_{j=1}^n \frac{(Y_j)^2 + (Y^{\mathbf{u}}_j)^2}{2} - \left(\frac{1}{n}\sum_{j=1}^n \frac{Y_j + Y^{\mathbf{u}}_j}{2}\right)^2}. \tag{6}$$

Asymptotic study Such estimation procedures have been proved to be consistent and asymptotically normal (i.e. the rate of convergence is $\sqrt{n}$) in [25, 18]. The limiting variances can be computed explicitly, allowing the practitioner to build confidence intervals. In addition, the sequence of estimators $(T^{\mathbf{u}}_n)_n$ is asymptotically efficient to estimate $S^{\mathbf{u}}$ from such a design of experiment (see, [39] for the definition of the asymptotic efficiency and [18] for the details of the result).

2.2 Cramér-von-Mises indices

Definition of the Cramér-von-Mises indices The Cramér-von-Mises indices introduced in [19] provide alternative indices based on the whole distribution rather than on the second moment of the output Y only. The main idea of Cramér-von-Mises indices is

to compare the conditional cumulative distribution function (c.d.f.) to the unconditional one via the L^2-norm. As for the Sobol' indices, they compare the conditional expectation of the output to the unconditional one. Notably, they are constructed following a similar scheme so that any procedure that estimates one index can be adapted to estimate the other.
More precisely, the Cramér-von-Mises indices are defined by

$$S^{\mathbf{u}}_{2,CVM} = \frac{\int_{\mathbb{R}} \mathbb{E}\left[(F(t) - F^{\mathbf{u}}(t))^2\right] dF(t)}{\int_{\mathbb{R}} F(t)(1 - F(t))dF(t)} \tag{7}$$

where F is the cumulative distribution function of Y

$$F(t) = \mathbb{P}\left(Y \leqslant t\right) = \mathbb{E}\left[\mathbb{1}_{\{Y \leqslant t\}}\right] \quad (t \in \mathbb{R})$$

and $F^{\mathbf{u}}$ is its Pick-Freeze version:

$$F^{\mathbf{u}}(t) = \mathbb{P}\left(Y \leqslant t | X_{\mathbf{u}}\right) = \mathbb{E}\Big[\mathbb{1}_{\{Y \leqslant t\}} | X_{\mathbf{u}}\Big] \quad (t \in \mathbb{R}).$$

This definition stems from the Hoeffding decomposition of the collection of r.v. $(\mathbb{1}_{\{Y \leqslant t\}})_{t \in \mathbb{R}}$.

Pick-Freeze estimation procedure of the Cramér-von-Mises indices The estimation procedure relies on (4) with $Y \leftarrow \mathbb{1}_{\{Y \leqslant t\}}$:

$$\mathrm{Var}(\mathbb{E}[\mathbb{1}_{\{Y \leqslant t\}} | X_{\mathbf{u}}]) = \mathrm{Cov}(\mathbb{1}_{\{Y \leqslant t\}}, \mathbb{1}_{\{Y^{\mathbf{u}} \leqslant t\}}). \tag{8}$$

Consequently, the Monte-Carlo estimation can be done as follows. In addition to the classical design of experiment required to estimate the Sobol' indices (an n-sample $(Y_1, \ldots, Y_n)$ of the output Y and an n-sample $(Y_1^{\mathbf{u}}, \ldots, Y_n^{\mathbf{u}})$ of its Pick-Freeze version $Y^{\mathbf{u}}$), a third independent n sample $(W_1, \ldots, W_n)$ of the output Y is necessary in order to deal with the integral with respect to $dF(t)$ in (7). Then the empirical estimator of $S^{\mathbf{u}}_{2,CVM}$ is

$$\frac{\frac{1}{n}\sum_{k=1}^n \left(\frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j \leqslant W_k\}} \mathbb{1}_{\{Y_j^{\mathbf{u}} \leqslant W_k\}} - \frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j \leqslant W_k\}} \frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j^{\mathbf{u}} \leqslant W_k\}}\right)}{\frac{1}{n}\sum_{k=1}^n \left(\frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j \leqslant W_k\}} - \left(\frac{1}{n}\sum_{j=1}^n \mathbb{1}_{\{Y_j \leqslant W_k\}}\right)^2\right)}. \tag{9}$$

Asymptotic study As showed in [19], this estimator is consistent and asymptotically Gaussian (i.e. the rate of convergence is $\sqrt{n}$). The limiting variance can be computed explicitly, allowing the practitioner to build confidence intervals.

3 A novel generation of estimators based on rank statistics

3.1 Chatterjee's correlation coefficient

In [9], Chatterjee considers a pair of real-valued random variables (V, Y) and an i.i.d. sample $(V_j, Y_j)_{1 \leqslant j \leqslant n}$. In order to simplify the presentation, we assume that the laws of

V and Y are both diffuse (ties are excluded). The pairs $(V_{(1)}, Y_{(1)}), \ldots, (V_{(n)}, Y_{(n)})$ are rearranged in such a way that

$$V_{(1)} < \ldots < V_{(n)}.$$

Then let $\pi(j)$ be the rank of V_j in the sample $(V_1, \ldots, V_n)$ of V and define

$$N'(j) = \begin{cases} \pi^{-1}(\pi(j)+1) & \text{if } \pi(j)+1 \leqslant n, \\ j & \text{if } \pi(j) = n. \end{cases} \tag{10}$$

The new correlation coefficient defined by Chatterjee in [9] is denoted $\xi_n(V, Y)$ and given by

$$\frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_{\{Y_k \leqslant Y_j\}}\mathbb{1}_{\{Y_k \leqslant Y_{N'(j)}\}} - \left(\frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_{\{Y_j \leqslant Y_k\}}\right)^2\right) \Big/ \frac{1}{n}\sum_{j=1}^{n} F_n(Y_j)(1 - F_n(Y_j)) \tag{11}$$

where F_n stands for the empirical distribution function of Y: $F_n(t) = \frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_{\{Y_k \leqslant t\}}$.
The author proves that $\xi_n(V, Y)$ converges a.s. to a deterministic limit $\xi(V, Y)$ which is equal to the Cramér-von-Mises sensitivity index $S^V_{2,CVM}$ with respect to V as soon as V is one of the random variables X_1, ..., X_p in the model (1) that are assumed to be real-valued. Further, he also proves a CLT when V and Y are independent.
Observe that the analogue of the Pick-Freeze version Y^V with respect to V of Y becomes Y_N and (8) is replaced by the formula

$$\mathbb{E}[\mathbb{1}_{\{Y_j \geqslant t\}}\mathbb{1}_{\{Y_{N'(j)} \geqslant t\}}|V_1, \ldots, V_n] = G_{V_j}(t)G_{V_{N'(j)}}(t) \tag{12}$$

for all $j = 1, \ldots, n$ that is mentioned in the proof of Lemma 7.10 in [9, p.24], with G_V the conditional survival function: $G_V(t) = \mathbb{P}(Y \geqslant t|V)$.
It is worth noticing that a unique n sample of input-output provides consistent estimations of the p first-order Cramér-von-Mises indices.

3.2 Generalization of Chatterjee's method

In this section, we propose a universal estimation procedure of expectations of the form

$$\mathbb{E}[\mathbb{E}[g(Y)|V]\mathbb{E}[h(Y)|V]],$$

for two integrable functions g and h. In fact, we consider a more general random element V (no longer assumed to be real) and a more general permutation denoted by τ_n. This result is a generalization of (12) and can be interpreted as an approximation of (4). To this end, we introduce the function Ψ_V defined by

$$\Psi_V(g) = \mathbb{E}[g(Y)|V] \tag{13}$$

for any integrable function g. Let $\mathcal{F}_n$ be the σ-algebra generated by $\{V_1, \ldots, V_n\}$. Note that in Section 3.1, we have considered $g(x) = g_t(x) = \mathbb{1}_{\{x \geqslant t\}}$ so that $\Psi_V(g) = \mathbb{P}(Y \geqslant t|V) = G_V(t)$.

Lemma 3.1. *Let g and h be two integrable functions such that gh is also integrable. Let $(V_j, Y_j)_{1 \leqslant j \leqslant n}$ be an n-sample of (V, Y). Consider a $\mathcal{F}_n$-measurable random permutation τ_n such that $\tau_n(j) \neq j$, for all $j = 1, \ldots, n$. Then*

$$\mathbb{E}\left[g(Y_j)h(Y_{\tau_n(j)})|V_1, \ldots, V_n\right] = \Psi_{V_j}(g)\Psi_{V_{\tau_n(j)}}(h). \tag{14}$$

The previous lemma (the proof of which has been postponed to Appendix A) leads to a generalization of the first part of the numerator of ξ_n defined in (11). Following the same lines as in [9], one may prove that such a quantity converges a.s. as $n \to \infty$ under some mild conditions. The reader is referred to Appendix A for the detailed proof of Proposition 3.2.

Proposition 3.2. *Let g and h be two bounded measurable functions. Consider a $\mathcal{F}_n$-measurable random permutation τ_n with no fix point (i.e. $\tau_n(j) \neq j$ for all $j = 1, \ldots, n$) and such that $V_{\tau_n(i)} \overset{\mathcal{L}}{=} V_{\tau_n(j)}$ for any i and $j = 1, \ldots, n$. In addition, we assume that for any $j = 1, \ldots, n$, $V_{\tau_n(j)} \to V_j$ as $n \to \infty$ a.s. Then $\chi_n(V, Y; g, h)$ defined by*

$$\chi_n(V, Y; g, h) = \frac{1}{n} \sum_{j=1}^{n} g(Y_j) h(Y_{\tau_n(j)}) \tag{15}$$

converges a.s. as $n \to \infty$ to $\chi(V, Y; g, h) = \mathbb{E}[\Psi_V(g)\Psi_V(h)]$, where Ψ_V has been defined in (13).

Notice that the permutation $\tau_n = N$ defined by

$$N(j) = \begin{cases} \pi^{-1}(\pi(j) + 1) & \text{if } \pi(j) + 1 \leqslant n, \\ \pi^{-1}(1) & \text{if } \pi(j) = n. \end{cases} \tag{16}$$

satisfies the assumptions of Lemma 3.1 and Proposition 3.2. Observe that N only differs from N' defined in (10) at j such that $\pi(j) = n$.

4 The rank estimator of the first-order Sobol' indices

4.1 Estimation procedure based on rank statistics

We can now leverage the above results and construct a new family of estimators for Sobol' indices. More precisely, let us consider the model (1) and assume we want to estimate the first-order Sobol' index S^1 defined in (2) with respect to $V = X_1$ assumed to be real-valued. We then define N as in (16) where π is the rank of X_1. Taking $g(x) = h(x) = x$ and $\tau_n = N$, (14) provides the analogue to ξ_n to estimate the classical Sobol' indices:

$$\xi_n^{\text{Sobol'}}(X_1, Y) := \frac{\frac{1}{n}\sum_{j=1}^{n} Y_j Y_{N(j)} - \left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)^2}{\frac{1}{n}\sum_{j=1}^{n} (Y_j)^2 - \left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)^2}, \tag{17}$$

where the denominator is reduced to the empirical variance of Y. As the functions g and h are here unbounded, Proposition 3.2 does not apply and thus offers no asymptotic information. However, the quantity of interest Y being generally bounded in practice, appropriately truncated versions of g and h could be considered.

4.2 A central limit theorem

We establish a CLT for the estimator $\xi_n^{\text{Sobol'}}(X_1, Y)$ of the first-order Sobol' index with respect to X_1 (assumed to be real-valued) under some mild assumptions on the model f and the random input X_1 in (1). The proof of the theorem is given in Appendix B.

Theorem 4.1. *Assume that X_1 is uniformly distributed on $[0,1]$ and f in (1) is a twice differentiable function with respect to its first coordinate. Further, we suppose that f and its two first derivatives (with respect to its first coordinate) are bounded. Then*

$$\sqrt{n}\left(\xi_n^{Sobol'}(X_1,Y) - S^1\right)$$

is asymptotically Gaussian with zero mean and explicit variance σ^2 given in Appendix B.4.

Remark 4.2. The boundedness of f implies that f has a fourth moment, that is the minimal assumption to get a CLT.
Moreover, let us observe that Theorem 4.1 only implies the convergence in probability. Nevertheless, under the assumptions of Theorem 4.1 (f bounded so is Y), Proposition 3.2 applies to derive the almost sure convergence of $\xi_n^{Sobol}(X1,Y)$.

The assumption on the distribution of X_1 can be relaxed as stated in the following corollary.

Corollary 4.3. *Let F_{X_1} be the cumulative distribution function of X_1. Assume that $f \circ F_{X_1}^{-1}$ is a twice differentiable function such that $f \circ F_{X_1}^{-1}$ and its two first derivatives are bounded. Then the conclusion of Theorem 4.1 still holds.*

Theorem 4.1 and Corollary 4.3 naturally allow to build statistical tests for testing $H_0 : S^1 = 0$ against $H_1 : S^1 \neq 0$. One can note that Chatterjee [9] result allows to test the independence of the input X_1 with respect to the output Y which is a stronger assumption than $S^1 = 0$, this was for example studied in [34]. In addition, our result allows to compute the power of the statistical test against any alternative of the kind $H_{1,0} : S^1 > s_0^1$ for any $s_0^1 > 0$.

Remark 4.4. A careful reading of the different steps of the proof shows that Theorem 4.1 can be slightly extended to more general situations involving more than two successive order statistics and with more general second variable $(X_2, \ldots, X_p)$. See the forthcoming paper [20].

The proof of our CLT is a bit long and technical and is postponed to the Appendix B. In a nutshell, this proof stands on three main ingredients. First, the regularity assumption on the function f allows to expand the statistic under study as a quadratic functional of the two independent sequences of random variables. The quadratic part for the first sequence involves order statistics of the uniform distribution and may be linearized. The second ingredient is the distribution representation of uniform order statistics by ratios of exponential convolution. The third ingredient is less classical and involves a conditional trick to show a central limit theorem for an empirical mean of a product. Let sketch the idea on a simple example. Let $(\xi_n)_n$ and $(\delta_n)_n$ be two independent sequences of centered square integrable random variables. We set $M_n = n^{-1/2}\sum_{j=1}^n \xi_j\delta_j$ and let $\mathcal{T}$ be the σ-field generated by the sequence (δ_n). Of course, the classical CLT gives that M_n converges in distribution towards a centered Gaussian distribution with variance $\mathrm{Var}(\xi_1)\mathrm{Var}(\delta_1)$. A less classical proof of this result consists in showing that, a.s., conditionally to $\mathcal{T}$ the same convergence in distribution holds. Indeed, this last result follows directly from the Lindeberg CLT and the strong law of large numbers for $n^{-1}\sum_{j=1}^n \delta_j^2$.

4.3 Comparison of the different estimation procedures

The estimator based on rank statistics $\xi_n^{\text{Sobol'}}(X_1, Y)$ defined in (17) can be compared to the classical Pick-Freeze estimators S_n^1 and T_n^1 given in (5) and (6) respectively (with $\mathbf{u} = \{1\}$) but also to a sequence of estimators involving the estimators $\widehat{T}_n$ introduced in [11].

Required sample sizes With the rank-based procedure, a unique n-sample of input-output provides consistent and asymptotically normal estimations of the p first-order Sobol' indices (together with consistent and asymptotically normal estimations of the p first-order Cramér-von-Mises indices with no extra cost). In contrast, using the Pick-Freeze estimation, if one wants to estimate all the p first-order Sobol' indices and the p Cramér-von-Mises indices, $(p+2)n$ calls of the computer code are required. The number of calls grows linearly with respect to the number of input parameters. This is a practical issue for large input dimension domains. A second drawback of the Pick-Freeze estimation scheme comes from the need of the particular Pick-Freeze design that is not always available.

Limiting variances Since the empirical mean and variance are already known to be asymptotically efficient in the statistical sense[1] to estimate the expectation and the variance of the output, we restrict our study to the comparison of the limiting variances obtained via the Pick-Freeze and the rank-based procedures in the estimation of $\mathbb{E}[\mathbb{E}[Y|X_1]^2]$ only.
In view of the proof of [25, Proposition 2.2], the Pick-Freeze limiting variance obtained using both S_n^1 and T_n^1 in estimating $\mathbb{E}[\mathbb{E}[Y|X_1]^2] = \mathbb{E}[YY^1]$ is simply given by $\text{Var}(YY^1)$, where $Y^1 = f(X_1, W^1)$ is the Pick-Freeze version of $Y = f(X_1, X_2, \ldots, X_p) = f(X_1, W)$. Using the above Lemmas B.1 and B.2 together with (41) leads to the rank-based limiting variance obtained using $\xi_n^{\text{Sobol'}}(X_1, Y)$:

$$\begin{aligned}\Sigma_B^{1,1} + \Sigma_C^{1,1} = \mathbb{E}\left[\text{Var}\left(YY^1|X_1\right)\right] + \mathbb{E}\left[\text{Cov}\left(YY^1, YY^{11}|X_1\right)\right] - \mathbb{E}[(Y+Y^1)f_x(X_1,W)X_1]^2 \\ + \mathbb{E}[(Y+Y^1)(\tilde{Y}+\tilde{Y}^1)f_x(X_1,W)f_x(\tilde{X}_1,\tilde{W})(X_1 \wedge \tilde{X}_1)],\end{aligned} \tag{18}$$

where $Y = f(X_1, X_2, \ldots, X_p) = f(X_1, W)$, $Y^1 = f(X_1, W^1)$, $Y^{11} = f(X_1, W^{11})$, $\tilde{Y} = f(\tilde{X}_1, \tilde{W})$, and $\tilde{Y}^1 = f(\tilde{X}_1, \tilde{W}^1)$ with X_1 and $\tilde{X}_1$ i.i.d., W, $\tilde{W}$, W^1, and W^{11} i.i.d. also independent of X_1 and $\tilde{X}_1$. Note that Y^1 and Y^{11} (respectively $\tilde{Y}^1$) are Pick-Freeze versions of Y (resp. $\tilde{Y}$). The paragraph's aim is to compare the limiting variances obtained by the two methods (Pick-Freeze and rank-based).
To do so, we recall that the Pick-Freeze experiment requires $n(p+1)$ observations (or computations of the black-box code) to estimate the p first-order Sobol' indices. In order to have a fair comparison of both estimation methods, we then consider that we have $n(p+1)$ i.i.d. observations of Y given by model (1) to estimate the p first-order Sobol' indices using the rank statistics. With $n(p+1)$ observations instead of n, the asymptotic variance obtained using the rank-based methodology is divided by $(p+1)$, so that we want to compare

$$V_{\text{PF}} := (p+1)(\text{Var}(YY^1), \ldots, \text{Var}(YY^p))^\top \text{ to } V_{\text{Rank}} := (\Sigma_B^{1,1} + \Sigma_C^{1,1}, \ldots, \Sigma_B^{p,p} + \Sigma_C^{p,p})^\top$$

[1]The reader is referred to [39, Section 25] for the definition of the asymptotic efficiency and related results.

where Y^i is the Pick-Freeze version of Y with respect to X_i (for $i = 2, \ldots, p$) and $\Sigma_B^{i,i} + \Sigma_C^{i,i}$ has the same expression as $\Sigma_B^{1,1} + \Sigma_C^{1,1}$ in (18) replacing the superscripts and the subscripts 1 by i (for $i = 2, \ldots, p$).

Example. We consider the following linear model

$$Y = f(X_1, \ldots, X_p) = \alpha X_1 + X_2 + \ldots + X_p, \tag{19}$$

where $\alpha > 0$ is a fixed constant, X_1, X_2, ..., and X_p are p independent and uniformly distributed random variables on $[0, 1]$.

We denote by $m_{1,p}$ and $m_{2,p}$ the two first moments of $Z_p := X_2 + \ldots + X_p$ and $m_{1,p,\alpha}$ and $m_{2,p,\alpha}$ the two first moments of $Z_{p,\alpha} := \alpha X_1 + X_3 + \ldots + X_p$. In addition, let v_p and $v_{p,\alpha}$ be the variances of Z_p of $Z_{p,\alpha}$. Hence $v_p = m_{2,p} - m_{1,p}^2$, $v_{p,\alpha} = m_{2,p,\alpha} - m_{1,p,\alpha}^2$,

$$m_{1,p} = \frac{1}{2}(p-1), \quad m_{2,p} = \frac{1}{12}(p-1)(3p-2), \quad m_{1,p,\alpha} = \frac{1}{2}(\alpha + m_{1,p-1}) = \frac{1}{2}(\alpha + p - 2),$$
$$m_{2,p,\alpha} = \frac{1}{3}\alpha^2 + \alpha m_{1,p-1} + m_{2,p-1} = \frac{1}{3}\alpha^2 + \frac{1}{2}(p-2)\alpha + \frac{1}{12}(p-2)(3p-5).$$

By symmetry, after obvious computations, one gets, for $i = 2, \ldots, p$,

$$\mathrm{Var}(YY^1) = \frac{4}{45}\alpha^4 + \frac{1}{3}m_{1,p}\alpha^3 + \frac{1}{3}\Big(2v_p + m_{1,p}^2\Big)\alpha^2 + 2m_{1,p}v_p\alpha + v_p(v_p + 2m_{1,p}^2),$$
$$\mathrm{Var}(YY^i) = \frac{4}{45} + \frac{1}{3}m_{1,p,\alpha} + \frac{1}{3}\Big(2v_{p,\alpha} + m_{1,p,\alpha}^2\Big) + 2m_{1,p,\alpha}v_{p,\alpha} + v_{p,\alpha}(v_{p,\alpha} + 2m_{1,p,\alpha}^2)$$

while

$$V_{\mathrm{Rank}}^1 = \frac{4}{45}\alpha^4 + \frac{1}{3}m_{1,p}\alpha^3 + \frac{1}{3}\Big(4v_p + m_{1,p}^2\Big)\alpha^2 + 4m_{1,p}v_p\alpha + v_p\Big(v_p + 4m_{1,p}^2\Big),$$
$$V_{\mathrm{Rank}}^i = \frac{4}{45} + \frac{1}{3}m_{1,p,\alpha} + \frac{1}{3}\Big(4v_{p,\alpha} + m_{1,p,\alpha}^2\Big) + 4m_{1,p,\alpha}v_{p,\alpha} + v_{p,\alpha}\Big(v_{p,\alpha} + 4m_{1,p,\alpha}^2\Big).$$

We compare these limiting variances in Figures 1 and 2. The results are clear and illustrate the fact that the rank-based methodology works much better for all value of $p \geqslant 2$. In addition, the more the value of p increases the greater the gain, as expected.

Remark 4.5. Observe that a more precise comparison should consists in comparing (via definite-positiveness) the limiting covariance-variance matrices involving both the limiting variances and the limiting covariances. If it is straightforward to compute the covariance terms for the Pick-Freeze methodology: for $i = 2, \ldots, p$,

$$\mathrm{Cov}(YY^1, YY^i) = \frac{1}{24}\alpha^4 + \frac{1}{12}m_{1,p-1}\alpha^3 + \Big(\frac{7}{144} + \frac{1}{4}v_{p-1} + \frac{1}{6}\Big(m_{1,p-1} + \frac{1}{2}\Big)^2\Big)\alpha^2$$
$$+ \Big(\frac{1}{8} + \frac{1}{12}m_{1,p-1} + \frac{1}{2}v_{p-1} + v_{p-1}m_{1,p-1}\Big)\alpha + v_{p-1}\Big(m_{1,p-1} + \frac{1}{2}\Big)^2,$$

it is much more tricky to deal with the rank-based procedure. Indeed, to do so a joint CLT is required for the vector of all p first-order Sobol' indices whose proof is not a direct generalization of the proof of Theorem 4.1. Such an extension will be done in a forthcoming paper.

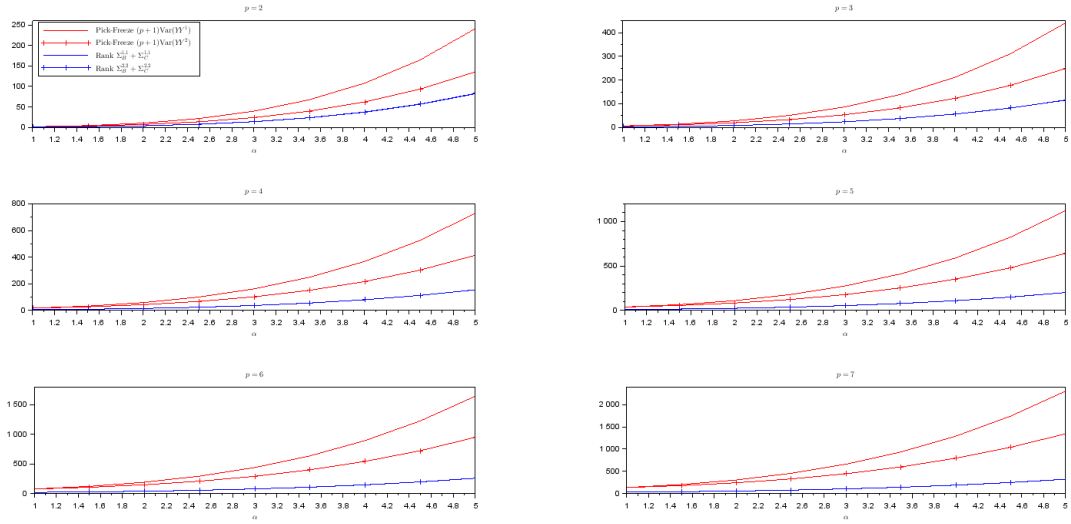

Figure 1: Linear model defined in (19). The limiting variances with respect to X_1 (plain lines) and to X_2 (plain lines with $+$) are plotted. The rank-based estimation procedure is represented in blue while the Pick-Freeze estimation procedure is represented in red. As explained, the Pick-Freeze estimation procedure has been weighted by $(p+1)$ to have a fair comparison. The number of variables involved in the model varies from $p = 2$ to $p = 7$.

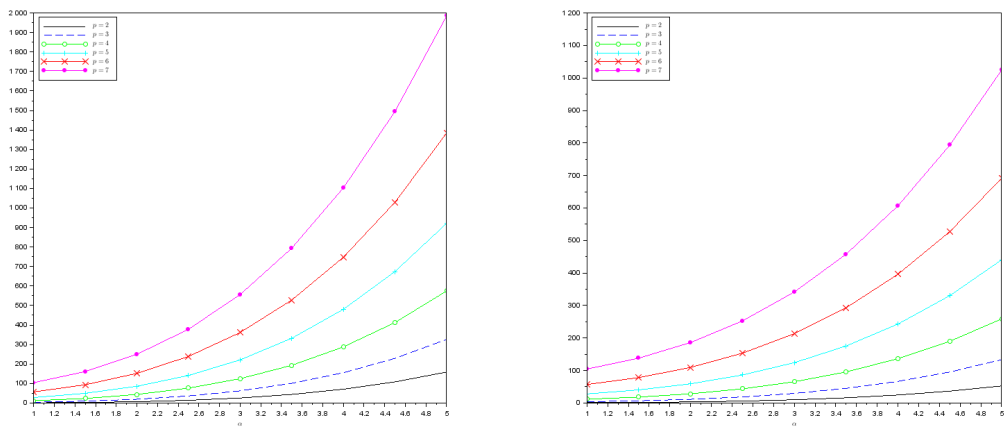

Figure 2: Linear model defined in (19). The difference between the limiting variances with respect to X_1 (left panel) and to X_2 (right panel) are plotted. As explained, the Pick-Freeze estimation procedure has been weighted by $(p+1)$ to have a fair comparison. The number of variables involved in the model varies from $p = 2$ to $p = 7$.

Asymptotic efficiency The two previous procedures do not rely on the same design of experiment so that it is not possible to determine which one is the more efficient in the sense of [39, Section 25].

By [18, Proposition 2.5], the sequence of estimators $(T_n^1)_n$ is asymptotically efficient to estimate S^1 when the distribution P of (Y, Y^1) belongs to $\mathcal{P}$, the set of all c.d.f. of

exchangeable random vectors in $L^2(\mathbb{R}^2)$.
Using a unique n-sample, one may compare the rank-based estimators introduced in this paper and the procedure involving the estimators $\widehat{T}_n$ defined in [11, page 11]. Such estimator is particularly tricky to compute and not easily tractable in practice. More precisely, the initial n-sample is split into two samples of sizes n_1 and $n_2 = n - n_1$. The first sample is dedicated to the estimation of the joint density of (X, Y) while the second one is used to compute a Monte-Carlo estimation of the integral involved in the quantity of interest. In a work under progress [22], another estimator based on kernels and the same design of experiment is proposed. This estimator is more tractable in practice.
By [11, Theorems 3.4 and 3.5], the sequence of estimators $(\widehat{T}_n)_n$ is asymptotically efficient to estimate $\mathbb{E}[\mathbb{E}[Y|X]^2]$ leading to an asymptotically efficient sequence of estimators of S^1. The proof of the following proposition has been postponed in Appendix C.

Proposition 4.6. *Consider the sequence of estimators $\widehat{T}_n$ introduced in [11, page 11]. Assume that the joint distribution P of (X, Y) is absolutely continuous with respect to the product probability $P_X \otimes P_Y$, namely $P(dx, dy) = f(x, y)P_X(dx)P_Y(dy)$. Then the sequence $(R_n^1)_n$*

$$R_n^1 = \frac{\widehat{T}_n - \left(\frac{1}{n}\sum_{i=1}^n Y_i\right)^2}{\frac{1}{n}\sum_{i=1}^n Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^n Y_i\right)^2}$$

is asymptotically efficient in estimating S^1. In addition, its (minimal) variance $\sigma^2_{\min}$ is

$$\sigma^2_{\min} := \frac{1}{Var(Y)^2} Var\Big(2\mathbb{E}[Y](1 - S^1)Y + S^1Y^2 + \mathbb{E}[Y|X](\mathbb{E}[Y|X] - 2Y)\Big).$$

Thus we are interested in the comparison of $\sigma^2_{\min}$ and σ^2 given in Theorem 4.1. Let us consider again the example of the linear model (19) introduced in the previous paragraph.
Example (continued). We consider the model defined in (19). As done in the previous paragraph, we only compare $V^1_{\text{Eff}} := \text{Var}(\mathbb{E}[Y|X_1](2Y - \mathbb{E}[Y|X_1]))$ to $\Sigma_B^{1,1} + \Sigma_C^{1,1}$ and $V^i_{\text{Eff}} := \text{Var}(\mathbb{E}[Y|X_i](2Y - \mathbb{E}[Y|X_i]))$ to $\Sigma_B^{i,i} + \Sigma_C^{i,i}$ for $i = 2, \ldots, p$. After some trivial computations, one gets

$$V^1_{\text{Eff}} = \frac{4}{45}\alpha^4 + \frac{1}{3}m_{1,p}\alpha^3 + \frac{1}{3}\Big(4v_p + m_{1,p}^2\Big)\alpha^2 + 4m_{1,p}v_p\alpha + 4v_p m_{1,p}^2,$$
$$V^i_{\text{Eff}} = \frac{4}{45} + \frac{1}{3}m_{1,p,\alpha} + \frac{1}{3}\Big(4v_{p,\alpha} + m_{1,p,\alpha}^2\Big) + 4m_{1,p,\alpha}v_{p,\alpha} + 4v_{p,\alpha}m_{1,p,\alpha}^2.$$

We compare these limiting variances in Figure 3. We observe that the limiting variances obtained with the rank methodology do not differ much from the efficient variances.

4.4 Recovering other classical indices

In [16], the authors considered computer codes of the form (1) valued on a compact Riemannian manifold. In this framework, they proposed a sensitivity index in the flavour of the Cramé-von-Mises index and they used the Pick-Freeze scheme to provide a consistent estimator. The authors of [21] extend the previous indices to the context of general metric spaces and propose U-statistics-based estimators improving the classical Pick-Freeze

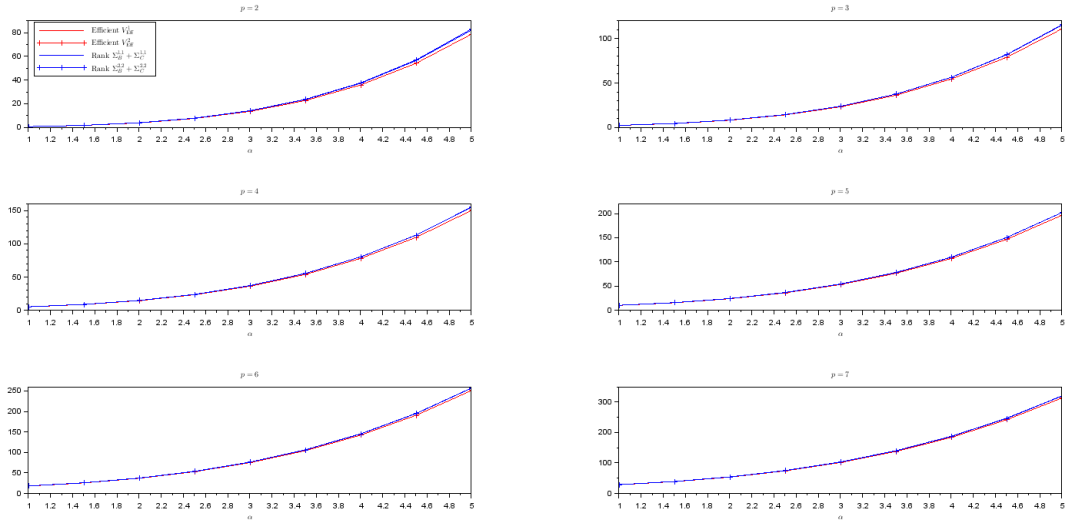

Figure 3: Linear model defined in (19). The limiting variances with respect to X_1 (plain lines) and to X_2 (plain lines with +) are plotted. The rank-based estimation procedure is represented in blue while the efficient variances are represented in red. The number of variables involved in the model varies from $p = 2$ to $p = 7$.

procedure. In light of Section 3.2, one may introduce a novel estimation of the indices introduced in [21] requiring a unique n-sample. The reader is referred to [14] for more details on the procedure.
Following [30, 31], extensions to Sobol' indices are obtained by replacing their numerator by higher-order moments. In [19], the authors construct a Pick-Freeze estimator for such extensions. One again, we are now able to propose another estimation scheme based on a unique n-sample. The reader is referred to [20] for the generalization of Lemma 3.1 and the corresponding asymptotic study.

5 Numerical experiments

5.1 Numerical comparison on the Sobol' g-function: conventional Pick-Freeze estimators vs rank estimators

In this section, we compare the performances of both estimation procedures on an analytic function: the so-called Sobol' g-function, that is defined by

$$g(X_1, \ldots, X_p) = \prod_{i=1}^{p} \frac{|4X_i - 2| + a_i}{1 + a_i}, \tag{20}$$

where $(a_i)_{i\in\mathbb{N}}$ is a sequence of real numbers and the X_i's are i.i.d. random variables uniformly distributed on $[0, 1]$. In this setting, one may easily compute the exact expression of the first-order Sobol' indices:

$$S^i = \frac{(1 + a_i^2)^{-1}/3}{3^{-p} \prod_{i=1}^{p} (1 + a_i^2)^{-1} - 1}.$$

As expected, the lower the coefficient a_i, the more significant the variable X_i. In the sequel, we simply fix $a_i = i$. Due to its complexity (non-linear and non-monotonic correlations) and the analytical expression of the Sobol' indices, the Sobol' g-function is a classical test example commonly used in GSA (see e.g. [32]).

Convergence as the sample size increases In Figure 4, we compare the estimations of the six first-order Sobol' indices given by both methods ($p = 6$). In the Pick-Freeze estimations given by (6), several sizes of sample N have been considered: $N = 100$, 500, 1000, 5000, 10000, 50000, 100000, and 500000. The Pick-Freeze procedure requires $(p+1) = 7$ samples of size N. To have a fair comparison, the sample sizes considered in the estimation of $\xi_n^{\text{Sobol'}}$ are $n = (p+1)N = 7N$. Both methods converge and give precise results for large sample sizes.

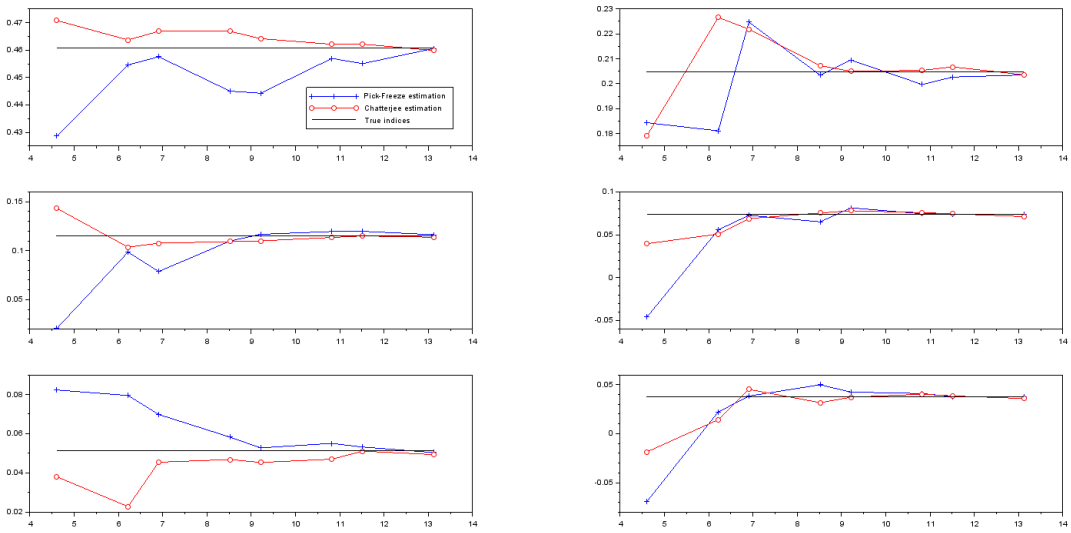

Figure 4: The Sobol' g-function model (20). Convergence of both methods when N increases. The sixth first-order Sobol' indices have been represented from left to right and up to bottom. Several sample sizes have been considered: $N = 100$, 500, 1000, 5000, 10000, 50000, 100000, and 500000 for the Pick-Freeze estimation procedure (in blue) and correspondingly $(p+1)N$ for the rank estimation procedure (in red). The true indices are displayed in black plain line. The x-axis is in log. scale.

Comparison of the mean square errors We now compare the efficiency of both methods at a fixed sample size. In that view, we assume that only $n = 700$ calls of the computer code f are allowed to estimate the six first-order Sobol' indices. We repeat the estimation procedure 500 times. The boxplot of the mean square errors for the estimation of the first-order Sobol' index S^1 with respect to X_1 has been represented in Figure 5. We observe that, for a fixed sample size $n = 700$ (corresponding to a Pick-Freeze sample size $N = 100$), the rank estimation procedure performs much better than the Pick-Freeze method with significantly lower mean errors. The same behavior can be observed for all the first Sobol' indices as can be seen in Table 1 that provides some characteristics of the mean squares errors.

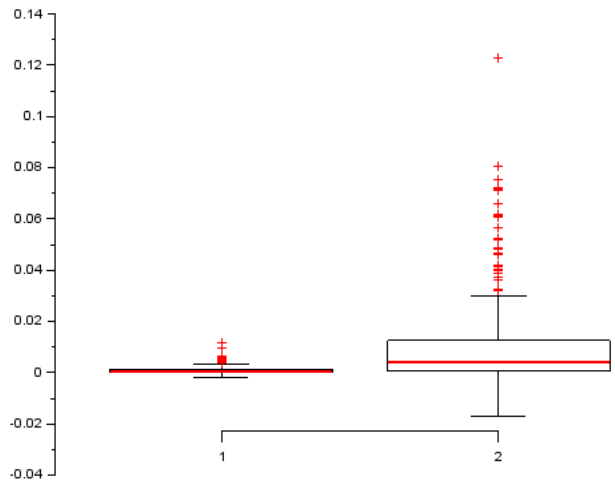

Figure 5: The Sobol' g-function model (20). Boxplot of the mean square errors of the estimation of S^1 with a fixed sample size and 500 replications. The results of the rank methodology with $n = 700$ are provided in the left panel. The results of the Pick-Freeze estimation procedure with $N = 100$ are provided in the right panel.

Performances for small sample sizes or for large number of input variables As expected, we can observe in Table 2 that the rank estimation procedure proceeds much better than the Pick-Freeze methodology for small sample sizes. Similarly, if the number of input variables increases drastically, we can observe the same behavior as can be seen in Figure 6. In that case, we consider the model (20) for several values of p: 6, 10, 15, 20, 30, 40, and 50.

5.2 An application in biology

Here, we illustrate the nature and the performance of the Cramér-von-Mises indices and their corresponding rank estimators as a screening mechanism for high-dimensional problems. To do so, we consider the neurovascular coupling model from [23]. Mathematically, this corresponds to the following differential-algebraic equation (DAE) system

$$\frac{dW}{dt} = G(W, Z, X), \quad 0 = H(W, Z, X), \tag{21}$$

where $W = (W_1, \ldots, W_N)$ and $Z = (Z_1, \ldots, Z_M)$ correspond respectively to the differential and algebraic state variables of the models. The variables $X = (X_1, \ldots, X_p)$ correspond to the uncertain parameters of the model. Our quantity of interest corresponds to the time average over $[0, T]$ of W^* (which is one of the differential state variables W_1, ..., W_N), i.e.

$$Y = \frac{1}{T} \int_0^{\top} W^*(t)\, dt. \tag{22}$$

As above, we regard Y as a function of the unknown parameters, i.e., $Y = f(X_1, \ldots, X_p)$. In our implementation, the values of W^* are obtained by solving the above DAE system

	Pick-Freeze			Rank		
	Mean	Median	Stdev	Mean	Median	Stdev
mse S^1	0.0095548	0.0039458	0.0145033	0.0010218	0.0004498	0.0013999
mse S^2	0.0105727	0.0046104	0.0148873	0.0017314	0.0006870	0.0027436
mse S^3	0.0101785	0.0041789	0.0143846	0.0016667	0.0006409	0.0024392
mse S^4	0.0105463	0.0047284	0.0178064	0.0018522	0.0008126	0.0025296
mse S^5	0.0097979	0.0042995	0.0135533	0.0016285	0.0006855	0.0024264
mse S^6	0.0096109	0.0046822	0.0134822	0.0015590	0.0007080	0.0021333

Table 1: The Sobol' g-function model (20). Characteristics of the mean square errors for the estimation of the six first-order Sobol' indices with a fixed sample size and 500 replications. In the rank methodology, the sample size is $n = 700$ while in the Pick-Freeze estimation procedure, it is $N = 100$.

	Pick-Freeze			Rank		
	$N = 10$	$N = 50$	$N = 100$	$n = 70$	$n = 350$	$n = 700$
mse S^1	0.1128686	0.0172275	0.0095548	0.0116790	0.0022941	0.0010218
mse S^2	0.1509575	0.0223196	0.0105727	0.0177522	0.0033719	0.0017314
mse S^3	0.1469124	0.0220015	0.0101785	0.0175517	0.0032474	0.0016667
mse S^4	0.1591130	0.0196357	0.0105463	0.0159360	0.0033948	0.0018522
mse S^5	0.1646339	0.0240353	0.0097979	0.0158563	0.0032230	0.0016285
mse S^6	0.1466408	0.0217638	0.0096109	0.0166701	0.0029653	0.0015590

Table 2: The Sobol' g-function model (20). Mean squares errors of the estimation of the six first-order Sobol' indices with small sample sizes and with both methods.

(Equation (21)) by the MATLAB routine ode15s (it can be checked that (21) form an index one system). Further, in the current example, $N = 67$ and $p = 160$ and the distributions of most of the X_i's are uniform and allowed to vary $\pm 10\%$ from nominal values (see [23] for additional details).
We compare the results from the rank estimators as described above to those resulting from the linear regression

$$f(X_1, \ldots, X_{160}) \approx \lambda_0 + \sum_{j=1}^{160} \lambda_j X_j.$$

As shown in [23], the above approximation performs well for the considered QoI. We assign to each variable $X_1, \ldots, X_{160}$ a relative importance L_j where

$$L_j = \frac{|\lambda_j|}{\sum_{\ell=1}^{160} |\lambda_\ell|}, \qquad j = 1, \ldots, 160.$$

Figure 7 displays the results. Both screening approaches identify the same to three influential parameters. More parameters are identified as being non-influential through the linear regression approach than using the Cramér-von-Mises indices.

6 Conclusion

In this paper, we explain how to use the estimator proposed by Chatterjee in [9] to provide a very nice and mighty procedure to estimate both all the first-order Sobol' indices and

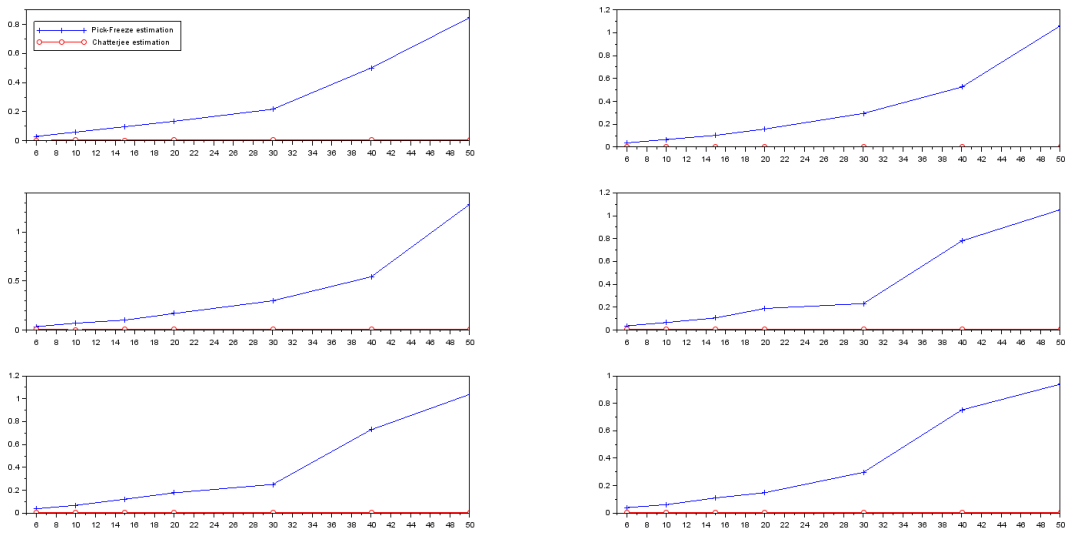

Figure 6: The Sobol' g-function model (20). Mean square errors of the estimation of the six first-order Sobol' indices with respect to the number of input variables with a fixed sample size and 500 replications. We consider the sample sizes $n = 200$ in the rank methodology (in red) and $N = n/(p+1)$ in the Pick-Freeze procedure (in blue). The number of input variables considered are $p = 6, 10, 15, 20, 30, 40$, and 50.

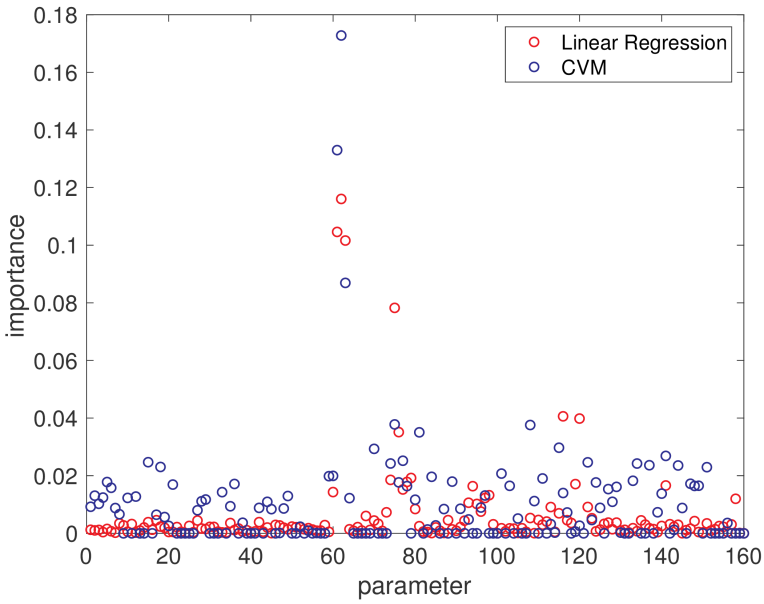

Figure 7: Rank estimators corresponding to the Cramér-von-Mises indices as a screening mechanics for the DAE system given by (21) and (21).

the so-called Cramér-von-Mises indices [19] at a small cost (only n calls of the computer code). We emphasize on the fact that this estimation procedure requires a unique sample contrary to the Pick-Freeze procedure based on a particular design of experiment, the size of which is 2n when estimating a single index and increases with the number of indices to estimate. We also extend Chatterjee's method to estimate more general quantities.

Furthermore, we show a CLT for our estimations of Sobol' indices. As examples, we consider two indices already introduced in sensitivity analysis: the indices adapted to output valued in general metric spaces defined in [21] and the higher-moment indices [30, 31]. A general CLT will be established soon in [20].

Acknowledgment. We warmly thank Robin Morillo for the numerical study provided in Section 5.2. Moreover, we deeply thank the anonymous referee of the early version of our paper who pushed us to prove the CLT. We also gratefully thank the anonymous reviewer of the current version of this paper for his comments, critics and advises, which greatly helped us to improve the manuscript.
Support from the ANR-3IA Artificial and Natural Intelligence Toulouse Institute is gratefully acknowledged. This work was also supported by the National Science Foundation under grant DMS-1745654.

A Proof of the consistency

Proof of Lemma 3.1. Since τ_n has no fix point, and using the measurability of τ_n and the independence, we have

$$\begin{aligned}
\mathbb{E}\left[g(Y_j)h(Y_{\tau_n(j)})|\mathcal{F}_n\right] &= \mathbb{E}\Big[g(Y_j)\sum_{\substack{l=1,\\ l\neq j}}^{n} h(Y_l)\mathbb{1}_{\{\tau_n(j)=l\}}|\mathcal{F}_n\Big] = \sum_{\substack{l=1,\\ l\neq j}}^{n} \mathbb{1}_{\{\tau_n(j)=l\}}\mathbb{E}\Big[g(Y_j)h(Y_l)|\mathcal{F}_n\Big] \\
&= \sum_{\substack{l=1,\\ l\neq j}}^{n} \mathbb{1}_{\{\tau_n(j)=l\}}\mathbb{E}\Big[g(Y_j)|\mathcal{F}_n\Big]\mathbb{E}\Big[h(Y_l)|\mathcal{F}_n\Big] = \mathbb{E}\Big[g(Y_j)|V_j\Big]\sum_{\substack{l=1,\\ l\neq j}}^{n} \mathbb{1}_{\{\tau_n(j)=l\}}\mathbb{E}\Big[h(Y_l)|V_l\Big] \\
&= \Psi_{V_j}(g)\sum_{\substack{l=1,\\ l\neq j}}^{n} \mathbb{1}_{\{\tau_n(j)=l\}}\Psi_{V_l}(h) = \Psi_{V_j}(g)\Psi_{V_{\tau_n(j)}}(h).
\end{aligned}$$

□

Proof of Proposition 3.2. We follow the steps of the proof of Corollary 7.12 in [9]. Our proof is significantly simpler since τ_n is assumed to have no fix points and V is continuous so that there are no ties in the sample. To simplify the notation, we denote $\chi_n(V, Y; g, h)$ and $\chi(V, Y; g, h)$ by χ_n and χ respectively.
We first prove that, for any measurable function φ,

$$\varphi(V_1) - \varphi(V_{\tau_n(1)}) \to 0 \tag{23}$$

in probability as $n \to \infty$. Let $\varepsilon > 0$. By the special case of Lusin's theorem (see [9, Lemma 7.5]), there exists a compactly supported continuous function $\tilde{\varphi}\colon \mathbb{R} \to \mathbb{R}$ such that $\mathbb{P}(\{x;\, \varphi(x) \neq \tilde{\varphi}(x)\}) < \varepsilon$, where $\mathbb{P}$ stands for the distribution of V. Then for any $\delta > 0$,

$$\begin{aligned}
\mathbb{P}\Big(\left|\varphi(V_1) - \varphi(V_{\tau_n(1)})\right| > \delta\Big) &\leqslant \mathbb{P}\left(\left|\tilde{\varphi}(V_1) - \tilde{\varphi}(V_{\tau_n(1)})\right| > \delta\right) \\
&\quad + \mathbb{P}\left(\varphi(V_1) \neq \tilde{\varphi}(V_1)\right) + \mathbb{P}(\varphi(V_{\tau_n(1)}) \neq \tilde{\varphi}(V_{\tau_n(1)}))\Big).
\end{aligned} \tag{24}$$

By continuity of $\tilde{\varphi}$ and since $V_{\tau_n(1)} \to V_1$ as $n \to \infty$ with probability one, the first term in the right hand side of (24) converges to 0 as $n \to \infty$. By construction of $\tilde{\varphi}$, the second

term is lower than ε. Turning to the third one, we have thus

$$\mathbb{E}[\varphi(V_{\tau_n(1)})] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\varphi(V_{\tau_n(j)})] = \frac{1}{n}\sum_{j=1}^{n}\sum_{\substack{l=1\\ l\neq j}}^{n}\mathbb{E}[\varphi(V_l)\mathbb{1}_{\{\tau_n(j)=l\}}]$$
$$= \frac{1}{n}\sum_{l=1}^{n}\sum_{\substack{j=1\\ j\neq l}}^{n}\mathbb{E}[\varphi(V_l)\mathbb{1}_{\{\tau_n(j)=l\}}] = \frac{1}{n}\sum_{l=1}^{n}\mathbb{E}[\varphi(V_l)\sum_{\substack{j=1\\ j\neq l}}^{n}\mathbb{1}_{\{\tau_n(j)=l\}}] = \frac{1}{n}\sum_{l=1}^{n}\mathbb{E}[\varphi(V_l)] = \mathbb{E}[\varphi(V_1)]$$

where we have used the fact that τ_n has no fix point, $V_{\tau_n(i)} \overset{\mathcal{L}}{=} V_{\tau_n(j)}$ for any i and $j = 1,\dots,n$, and the V_i's have no ties. This yields

$$\mathbb{P}(\varphi(V_{\tau_n(1)}) \neq \tilde{\varphi}(V_{\tau_n(1)})) = \mathbb{P}(\varphi(V_1) \neq \tilde{\varphi}(V_1)) < \varepsilon,$$

and, since ε and δ are arbitrary, (23) is therefore proved. Now, since $x \mapsto \Psi_x$ is a measurable and bounded function and applying (23), we have

$$\begin{cases} \Psi_{V_1}(g) - \Psi_{V_{\tau_n(1)}}(g) & \to 0, \\ \Psi_{V_1}(h) - \Psi_{V_{\tau_n(1)}}(h) & \to 0, \end{cases} \quad \text{in probability as } n \to \infty. \tag{25}$$

Lemma 3.1 and the dominated convergence theorem lead to

$$\mathbb{E}[\chi_n] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[g(Y_j)h(Y_{\tau_n(j)})] = \mathbb{E}[g(Y_1)h(Y_{\tau_n(1)})] = \mathbb{E}[\Psi_{V_1}(g)\Psi_{V_{\tau_n(1)}}(h)] \to \mathbb{E}[\Psi_V(g)\Psi_V(h)] = \chi \tag{26}$$

where we have taken into account the fact that $\Psi_V(g)$ and $\Psi_V(h)$ are bounded (due to the boundedness of g and h) and used (25).
The last step of the proof consists in comparing χ_n with $\mathbb{E}[\chi_n]$ using Mc Diarmid's concentration inequality [27]. Sharper constants can be obtained in Mc Diarmid's inequality by using the inequalities from [6, 7]. As we are interested in asymptotic results the accuracy of the constant has no impact on the result. Following the same lines as in the proof of [9, Lemma 7.11], Mc Diarmid's concentration inequality in [27] then implies

$$\mathbb{P}(|\chi_n - \mathbb{E}[\chi_n]| \geqslant t) \leqslant 2\exp\{-2n^2t^2/C^2\}, \tag{27}$$

where C is a universal constant and we conclude the proof by combining (26) and (27). □

B Proof of the asymtotic normality

Framework and goal We consider the model defined in (1) that can be rewritten as $Y = f(X, W)$ where $X = X_1$ and $W = (X_2, \dots, X_p)$ are two independent inputs of the numerical code f that is assumed to be bounded.
The random variables X and W are defined on a product space $\Omega = \Omega_X \times \Omega_W$; so that for any $\omega \in \Omega$, there exists $\omega_X \in \Omega_X$ and $\omega_W \in \Omega_W$ and we have $(X, W)(\omega) = (X(\omega_X), W(\omega_W))$. Further, we consider π_W the projection on Ω_W and the product measure $\mathbb{P} = \mathbb{P}_X \otimes \mathbb{P}_W = \mathcal{L}_X \otimes \mathcal{L}_W$, where $\mathcal{L}_X$ is the distribution of X and $\mathcal{L}_W$ is the distribution of W. Naturally, $\mathbb{P}_W = \mathbb{P} \circ \pi_W^{-1}$.

We aim to prove a CLT for the estimator $\xi_n^{\text{Sobol'}}(X,Y)$ of the classical first-order Sobol' index with respect to X given by (2), the estimator of which defined in (17) is given by

$$\xi_n^{\text{Sobol'}}(X_1,Y) = \frac{\frac{1}{n}\sum_{j=1}^n Y_j Y_{N(j)} - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}{\frac{1}{n}\sum_{j=1}^n Y_j^2 - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}$$

where N is defined in (16). Notice that the denominator is reduced to the empirical variance of Y. As explained in Section 3.1, we denote by $Y_{(j)}$ the output associated to $X_{(j)}$ where $X_{(j)}$ stands for the j-th order statistics of $(X_1,\dots,X_n)$. Then observing that

$$\sum_{j=1}^n Y_j Y_{N(j)} = \sum_{j=1}^n Y_{(j)} Y_{(j+1)} =: \sum_{j=1}^n Y_{\sigma_n(j)} Y_{\sigma_n(j+1)}$$

where, to avoid any confusion, σ_n stands for the permutation that rearranges the sample $(X_1,\dots,X_n)$, the estimator $\xi_n^{\text{Sobol'}}(X_1,Y)$ can be written as

$$\xi_n^{\text{Sobol'}}(X_1,Y) = \frac{\frac{1}{n}\sum_{j=1}^{n-1} Y_{\sigma_n(j)} Y_{\sigma_n(j+1)} - \left(\frac{1}{n}\sum_{j=1}^n Y_{\sigma_n(j)}\right)^2}{\frac{1}{n}\sum_{j=1}^n Y_{\sigma_n(j)}^2 - \left(\frac{1}{n}\sum_{j=1}^n Y_{\sigma_n(j)}\right)^2}. \tag{28}$$

B.1 Proof of Theorem 4.1

The proof will proceed as follows. First, in view of (28), we prove a CLT for

$$\left(\frac{1}{n}\sum_{j=1}^{n-1} Y_{\sigma_n(j)} Y_{\sigma_n(j+1)}, \frac{1}{n}\sum_{j=1}^{n} Y_{\sigma_n(j)}, \frac{1}{n}\sum_{j=1}^{n} Y_{\sigma_n(j)}^2\right).$$

that amounts to prove a CLT for

$$\left(\frac{1}{n}\sum_{j=1}^{n-1} Y_{\sigma_n(j)} Y_{\sigma_n(j+1)}, \frac{1}{n}\sum_{j=1}^{n-1} Y_{\sigma_n(j)}, \frac{1}{n}\sum_{j=1}^{n-1} Y_{\sigma_n(j)}^2\right),$$

since f is bounded. Secondly, we use the so-called delta method [39, Theorem 3.1] to conclude to Theorem 4.1.
It is worth noticing that the permutation on the W's do not affect the result as seen in the sequel. For $j = 1,\dots n-1$, introducing

$$\Delta_{n,j} := f\left(X_{\sigma_n(j)}, W_j\right) - f\left(\frac{j}{n+1}, W_j\right), \quad W_{n,j} := \left(\frac{j}{n+1}, W_j\right) \tag{29}$$

leads to $Y_{\sigma_n(j)} = f\left(X_{\sigma_n(j)}, W_{\sigma_n(j)}\right) \overset{\mathcal{L}}{=} f\left(X_{\sigma_n(j)}, W_j\right) = \Delta_{n,j} + f\left(W_{n,j}\right)$ and

$$\begin{aligned}
Y_{\sigma_n(j)} Y_{\sigma_n(j+1)} &= f\left(X_{\sigma_n(j)}, W_{\sigma_n(j)}\right) f\left(X_{\sigma_n(j+1)}, W_{\sigma_n(j+1)}\right) \\
&\overset{\mathcal{L}}{=} f\left(X_{\sigma_n(j)}, W_j\right) f\left(X_{\sigma_n(j+1)}, W_{j+1}\right) \\
&= \Big(f\left(W_{n,j}\right) + \Delta_{n,j}\Big)\Big(f\left(W_{n,j+1}\right) + \Delta_{n,j+1}\Big) \\
&= f\left(W_{n,j}\right) f\left(W_{n,j+1}\right) + \Delta_{n,j} f\left(W_{n,j+1}\right) + \Delta_{n,j+1} f\left(W_{n,j}\right) + \Delta_{n,j}\Delta_{n,j+1}.
\end{aligned}$$

Thus we are led to establish a CLT for

$$Z_n = \frac{1}{n}\sum_{j=1}^{n-1}\begin{pmatrix} f(W_{n,j})f(W_{n,j+1}) + \Delta_{n,j} f\left(W_{n,j+1}\right) + \Delta_{n,j+1} f\left(W_{n,j}\right) + \Delta_{n,j}\Delta_{n,j+1} \\ f(W_{n,j}) + \Delta_{n,j} \\ \left(f(W_{n,j}) + \Delta_{n,j}\right)^2 \end{pmatrix}. \quad (30)$$

Let us discard the negligible terms in the CLT for Z_n. In that view, noticing that

$$\mathbb{E}\left[X_{\sigma_n(j)}\right] = \frac{j}{n+1} \quad \text{and} \quad \mathrm{Var}(X_{\sigma_n(j)}) = \frac{j(n-j+1)}{(n+1)^2(n+2)} = \mathbb{E}\left[\left(X_{\sigma_n(j)} - \frac{j}{n+1}\right)^2\right] \leqslant \frac{4}{n+2},$$

we first establish

$$X_{\sigma_n(j)} - \frac{j}{n+1} = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right). \quad (31)$$

As explained below, (31) will imply

$$\frac{1}{n}\sum_{j=1}^{n-1}\Delta_{n,j}^2 = O_{\mathbb{P}}\left(\frac{1}{n}\right) \quad \text{and} \quad \frac{1}{n}\sum_{j=1}^{n-1}\Delta_{n,j}\Delta_{n,j+1} = O_{\mathbb{P}}\left(\frac{1}{n}\right). \quad (32)$$

First of all, we expand $\Delta_{n,j}$ (resp. $\Delta_{n,j+1}$) using the Taylor-Lagrange formula, for any $j = 1, \dots n-1$ and we obtain

$$\Delta_{n,j} = \left(X_{\sigma_n(j)} - \frac{j}{n+1}\right) f_x\left(W_{n,j}\right) + \frac{1}{2}\left(X_{\sigma_n(j)} - \frac{j}{n+1}\right)^2 f_{xx}\left(\delta_{n,j}, W_{\sigma_n(j)}\right), \quad (33)$$

where $\delta_{n,j}$ (resp. $\delta_{n,j+1}$) lies in the unordered segment $(X_{\sigma_n(j)}, j/(n+1))$ (resp. $(X_{\sigma_n(j+1)}, (j+1)/(n+1))$) and where f_x and f_{xx} are the first and second derivatives of f with respect to the first coordinate. This leads to expansions for $\Delta_{n,j}^2$ and $\Delta_{n,j}\Delta_{n,j+1}$:

$$\begin{aligned}
\Delta_{n,j}^2 &= \left(X_{\sigma_n(j)} - \frac{j}{n+1}\right)^2\left(f_x\left(W_{n,j}\right) + \frac{1}{2}\left(X_{\sigma_n(j)} - \frac{j}{n+1}\right) f_{xx}\left(\delta_{n,j}, W_{\sigma_n(j)}\right)\right)^2 \\
\Delta_{n,j}\Delta_{n,j+1} &= \left(X_{\sigma_n(j)} - \frac{j}{n+1}\right)\left(X_{\sigma_n(j+1)} - \frac{j+1}{n+1}\right) \\
&\quad \times \left(f_x\left(W_{n,j}\right) + \frac{1}{2}\left(X_{\sigma_n(j)} - \frac{j}{n+1}\right) f_{xx}\left(\delta_{n,j}, W_{\sigma_n(j)}\right)\right) \\
&\quad \times \left(f_x\left(W_{n,j+1}\right) + \frac{1}{2}\left(X_{\sigma_n(j+1)} - \frac{j+1}{n+1}\right) f_{xx}\left(\delta_{n,j+1}, W_{\sigma_n(j+1)}\right)\right).
\end{aligned}$$

Finally, using the boundedness of f, f_x, and f_{xx}, together with (31), (32) follows. Remark that the proof of (32) yields also

$$\frac{1}{n}\sum_{j=1}^{n-1}\Delta_{n,j} = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right), \quad (34)$$

from which it is clear that this term will contribute in the CLT on Z_n. Then (32) entails that the asymptotic study reduces to that of the empirical mean of $Z_{n,j} = B_{n,j} + C_{n,j}$ where

$$B_{n,j} := \begin{pmatrix} f\left(W_{n,j}\right) f\left(W_{n,j+1}\right) \\ f(W_{n,j}) \\ f(W_{n,j})^2 \end{pmatrix} \text{ and } C_{n,j} := \begin{pmatrix} \Delta_{n,j} f\left(W_{n,j+1}\right) + \Delta_{n,j+1} f\left(W_{n,j}\right) \\ \Delta_{n,j} \\ 2\Delta_{n,j} f(W_{n,j}) \end{pmatrix}. \quad (35)$$

First, we consider $B_{n,j}$ in (35) and we establish the following result, the proof of which has been postponed to Appendix B.2.

Lemma B.1. *As $n \to \infty$, the random vector B_n given by*

$$\frac{1}{n}\sum_{j=1}^{n-1} B_{n,j} = \frac{1}{n}\sum_{j=1}^{n-1}\left(f\left(W_{n,j}\right) f\left(W_{n,j+1}\right), f\left(W_{n,j}\right), f\left(W_{n,j}\right)^2\right)^\top$$

satisfies a CLT. More precisely, $\sqrt{n}\Big(B_n - m_B\Big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_3(0, \Sigma_B)$, where

$$m_B := \Big(\mathbb{E}[YY'], \mathbb{E}[Y], \mathbb{E}[Y^2]\Big)^\top, \tag{36}$$

$Y' = f(X, W')$, W' is an independent copy of W, and Σ_B has an explicit expression given in Appendix B.2.

Remark that Y' is the so-called Pick-Freeze version of Y with respect to X. Secondly, we establish a conditional CLT for the empirical mean of the $C_{n,j}$'s defined in (35). The reader is referred to Appendix B.3 for the proof of this result.

Lemma B.2. *There exists a measurable set $\Pi \in \Omega_W$ having $\mathbb{P}_W$-probability one such that, for any $\omega_W \in \Pi$, we have*

$$\sqrt{n}C_n(\cdot, \omega_W) \xrightarrow[n\to\infty]{\mathcal{L}_X} \mathcal{N}_3(0, \Sigma_C).$$

Moreover, Σ_C does not depend on ω_W and has an explicit expression given Appendix B.3.

Considering the characteristic function of the vector $\sqrt{n}(B_n - \mathbb{E}[B_n], C_n)$, one may write

$$\mathbb{E}\left[e^{i(\sqrt{n}\langle s,(B_n-\mathbb{E}[B_n])\rangle+\sqrt{n}\langle t,C_n\rangle)}\right] = \mathbb{E}\left[e^{i\sqrt{n}\langle s,(B_n-\mathbb{E}[B_n])\rangle}\mathbb{E}\left[e^{i\sqrt{n}\langle t,C_n\rangle}\Big|\mathcal{F}_W\right]\right]$$

for any s and $t \in \mathbb{R}^3$. On the one hand, $\mathbb{E}\left[e^{i\sqrt{n}\langle t,C_n\rangle}\Big|\mathcal{F}_W\right]$ converges a.s. to $\exp\{-t^\top\Sigma_C t/2\}$ which is not random. On the other hand, $\sqrt{n}\langle s, (B_n - \mathbb{E}[B_n])\rangle$ converges in distribution to a Gaussian random variable denoted by B_s. By Slutsky's lemma,

$$\left(\sqrt{n}\langle s, (B_n - \mathbb{E}[B_n])\rangle, \mathbb{E}\left[e^{i\sqrt{n}\langle t,C_n\rangle}\Big|\mathcal{F}_W\right]\right)$$

converges in distribution to $(B_s, \exp\{-t^\top\Sigma_C t/2\})$. We consider the application $h\colon (u, v) \in \mathbb{R} \times D(0, 1) \mapsto e^{iu}v \in \mathbb{C}$ where $D(0, 1)$ is the unit disc in $\mathbb{C}$. The continuity and the boundedness of h lead to the convergence in distribution of $e^{i\sqrt{n}\langle s,(B_n-\mathbb{E}[B_n])\rangle}\left[e^{i\sqrt{n}\langle t,C_n\rangle}\Big|\mathcal{F}_W\right]$ and we conclude to the asymptotic normality of $\sqrt{n}(B_n - \mathbb{E}[B_n], C_n)$ to a six-dimensional Gaussian random vector with zero mean and variance-covariance matrix $\begin{pmatrix}\Sigma_B & 0\\ 0 & \Sigma_C\end{pmatrix}$. It remains to apply the so-called delta method [39, Theorem 3.1] and Slutsky's lemma to get the required result. The details of the computation of the asymptotic variance σ^2 can be found in Appendix B.4.

B.2 Proof of Lemma B.1

One has

$$\mathbb{E}[B_n] = \frac{1}{n}\sum_{j=1}^{n-1}\left(\mathbb{E}\left[f\left(W_{n,j}\right) f\left(W_{n,j+1}\right)\right], \mathbb{E}\left[f\left(W_{n,j}\right)\right], \mathbb{E}\left[f\left(W_{n,j}\right)^2\right]\right)^\top,$$

the first coordinate of which converges as $n \to \infty$ to

$$\begin{aligned}\int \mathbb{E}\left[f\left(x,W\right)f\left(x',W'\right)\right]\mathrm{d}\mathcal{L}_{(X,X)}(x,x') &= \int_0^1 \mathbb{E}\left[f\left(x,W\right)f\left(x,W'\right)\right]\mathrm{d}x \\ &= \mathbb{E}\left[\mathbb{E}\left[f\left(X,W\right)f\left(X,W'\right)|X\right]\right] \\ &= \mathbb{E}\left[f\left(X,W\right)f\left(X,W'\right)\right] = \mathbb{E}\left[YY'\right].\end{aligned}$$

The two other coordinates can be handled similarly leading to

$$\mathbb{E}[B_n] \underset{n\to\infty}{\to} \left(\mathbb{E}[YY'],\mathbb{E}[Y],\mathbb{E}[Y^2]\right)^\top = m_B.$$

We apply the CLT for dependent variables proved in [28] to $\widetilde{B}^1_{n,j}$, the centered version of the random variables $f\left(W_{n,j}\right)f\left(W_{n,j+1}\right)/\sqrt{n}$ with $m=1$, $\alpha=0$, and because f is bounded (so is $\widetilde{B}^1_{n,j}$). Assumptions (1) and (2) in [28] obviously hold, the assumption (3) is naturally fulfilled and assumption (4) is a mere consequence of Chebyshev's inequality and the boundedness of f. Now, it remains to check that assumption (5) holds. We have

$$\begin{aligned}\sum_{i,j=1}^{n-1}\mathrm{Cov}(\widetilde{B}^1_{n,i},\widetilde{B}^1_{n,j}) &= \frac{1}{n}\sum_{i,j=1}^{n-1}\mathrm{Cov}\left(f\left(W_{n,i}\right)f\left(W_{n,i+1}\right),f\left(W_{n,j}\right)f\left(W_{n,j+1}\right)\right) \\ &= \frac{1}{n}\sum_{j=1}^{n-1}\mathrm{Var}\left(f\left(W_{n,j}\right)f\left(W_{n,j+1}\right)\right) + \frac{2}{n}\sum_{j=1}^{n-2}\mathrm{Cov}\left(f\left(W_{n,j}\right)f\left(W_{n,j+1}\right),f\left(W_{n,j+1}\right)f\left(W_{n,j+2}\right)\right).\end{aligned}$$

On the one hand, by [17, Lemma 1.1],

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{n-1}\mathrm{Var}\left(f\left(W_{n,j}\right)f\left(W_{n,j+1}\right)\right) \underset{n\to\infty}{\to} \int \mathrm{Var}\left(f\left(x,W\right)f\left(x',W'\right)\right)\mathrm{d}\mathcal{L}_{(X,X)}(x,x') \\ &\quad = \int_0^1 \mathrm{Var}\left(f\left(x,W\right)f\left(x,W'\right)\right)\mathrm{d}x = \mathbb{E}\left[\mathrm{Var}\left(f\left(X,W\right)f\left(X,W'\right)|X\right)\right] = \mathbb{E}\left[\mathrm{Var}\left(YY'|X\right)\right],\end{aligned}$$

where W' is an independent copies of W, $Y=f(X,W)$, and $Y'=f(X,W')$. On the other hand, by [17, Lemma 1.1],

$$\begin{aligned}&\frac{1}{n}\sum_{j=1}^{n-2}\mathrm{Cov}\left(f\left(W_{n,j}\right)f\left(W_{n,j+1}\right),f\left(W_{n,j+1}\right)f\left(W_{n,j+2}\right)\right) \\ &\underset{n\to\infty}{\to} \mathbb{E}\left[\mathrm{Cov}\left(f\left(X,W\right)f\left(X,W'\right),f\left(X,W'\right)f\left(X,W''\right)|X\right)\right] = \mathbb{E}\left[\mathrm{Cov}\left(YY',YY''|X\right)\right],\end{aligned}$$

where W' and W'' are two independent copies of W. Further, $Y=f(X,W)$, $Y'=f(X,W')$, and $Y''=f(X,W'')$. Actually, notice that all linear combination of the coordinates of

$$\left(f(W_{n,j})f(W_{n,j+1}),f(W_{n,j}),f(W_{n,j})^2\right)^\top \tag{37}$$

is a one-dependent random variable. In addition, following the same lines as above, one may check that any linear combination still satisfies the assumptions of [28]. Hence, any linear combination of the coordinates of B_n satisfies a CLT so that Lemma B.1 is proved, up to the computation of the asymptotic variance-covariance matrix Σ_B done in what follows.

Computation of the asymptotic covariance matrix Σ_B

We consider a linear combination of the random vector in (37) given by

$$uf(W_{n,j})f(W_{n,j+1}) + vf(W_{n,j}) + wf(W_{n,j})^2,$$

where $(u, v, w) \in \mathbb{R}^3$. This one-dimensional random vector is one-dependent and its centered version normalized by $\sqrt{n}$, denoted by $\widetilde{B}_{n,j}$, satisfies the assumptions of [28]. To calculate the asymptotic variance-covariance matrix Σ_B, we compute explicitly the limit of

$$\sum_{i,j=1}^{n-1} \mathrm{Cov}(\widetilde{B}_{n,i}, \widetilde{B}_{n,j}),$$

as $n \to \infty$ using [17, Lemma 1.1]. It remains to take $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ to get the diagonal terms of the asymptotic variance-covariance matrix and to solve a three-dimensional system of equations to get the remaining terms. Finally, as computed previously and using notation of [17, Lemma 1.1], the first diagonal term of Σ_B is :

$$\begin{aligned}
\Sigma_B^{1,1} &= \int \mathrm{Var}\left(f\left(x, W\right) f\left(x', W'\right)\right) \mathrm{d}\mathcal{L}_{(X,X)}(x, x') \\
&\quad + 2 \int \mathrm{Cov}\left(f\left(x, W\right) f\left(x', W'\right), f\left(x', W'\right) f\left(x'', W''\right)\right) \mathrm{d}\mathcal{L}_{(X,X,X)}(x, x', x'') \\
&= \int_0^1 \mathrm{Var}\left(f\left(x, W\right) f\left(x, W'\right)\right) \mathrm{d}x + 2 \int_0^1 \mathrm{Cov}\left(f\left(x, W\right) f\left(x, W'\right), f\left(x, W'\right) f\left(x, W''\right)\right) \mathrm{d}x \\
&= \mathbb{E}\left[\mathrm{Var}\left(f\left(X, W\right) f\left(X, W'\right) | X\right)\right] + 2\mathbb{E}\left[\mathrm{Cov}\left(f\left(X, W\right) f\left(X, W'\right), f\left(X, W'\right) f\left(X, W''\right) | X\right)\right] \\
&= \mathbb{E}\left[\mathrm{Var}\left(YY' | X\right)\right] + 2\mathbb{E}\left[\mathrm{Cov}\left(YY', YY'' | X\right)\right],
\end{aligned}$$

where we remind that $Y = f(X, W)$, $Y' = f(X, W')$, and $Y'' = f(X, W'')$ with W' and W'' independent copies of W. The other terms are

$$\begin{aligned}
\Sigma_B^{2,2} &= \int_0^1 \mathrm{Var}\left(f\left(x, W\right)\right) \mathrm{d}x = \mathbb{E}\left[\mathrm{Var}\left(f\left(X, W\right) | X\right)\right] = \mathbb{E}\left[\mathrm{Var}(Y | X)\right], \\
\Sigma_B^{3,3} &= \int_0^1 \mathrm{Var}\left(f\left(x, W\right)^2\right) \mathrm{d}x = \mathbb{E}\left[\mathrm{Var}\left(Y^2 | X\right)\right], \\
\Sigma_B^{1,2} &= \Sigma_B^{2,1} = 2 \int_0^1 \mathrm{Cov}\left(f\left(x, W\right) f\left(x, W'\right), f\left(x, W\right)\right) \mathrm{d}x = 2\mathbb{E}\left[\mathrm{Cov}\left(YY', Y | X\right)\right], \\
\Sigma_B^{1,3} &= \Sigma_B^{3,1} = 2 \int_0^1 \mathrm{Cov}\left(f\left(x, W\right) f\left(x, W'\right), f\left(x, W\right)^2\right) \mathrm{d}x = 2\mathbb{E}\left[\mathrm{Cov}\left(YY', Y^2 | X\right)\right], \\
\Sigma_B^{2,3} &= \Sigma_B^{3,2} = \int_0^1 \mathrm{Cov}\left(f\left(x, W\right), f\left(x, W\right)^2\right) \mathrm{d}x = \mathbb{E}\left[\mathrm{Cov}(Y, Y^2 | X)\right].
\end{aligned}$$

B.3 Proof of Lemma B.2

Let $\omega_W \in \Pi$ as defined in [17, Lemma 1.1]. The aim is to establish a CLT for $\sqrt{n} C_{n,j}(\cdot, \omega_W)$. To ease the reading, we omit the notation $(\cdot, \omega_W)$ as classically done in probability. First, dealing with the first coordinate $f\left(W_{n,j+1}\right) \Delta_{n,j} + f\left(W_{n,j}\right) \Delta_{n,j+1}$ of $C_{n,j}$ defined in (35), one has

$$\begin{aligned}
f\left(W_{n,j+1}\right) \Delta_{n,j} &= \left(X_{\sigma_n(j)} - \frac{j}{n+1}\right) f\left(W_{n,j+1}\right) f_x\left(W_{n,j}\right) \\
&\quad + \frac{1}{2}\left(X_{\sigma_n(j)} - \frac{j}{n+1}\right)^2 f\left(W_{n,j+1}\right) f_{xx}\left(\delta_{n,j}, W_j\right)
\end{aligned}$$

using the expansion of $\Delta_{n,j}$ given in (33). By (31) and using the boundedness of f and f_{xx}, we get that

$$\frac{1}{n}\sum_{j=1}^{n-1}\left(X_{\sigma_n(j)}-\frac{j}{n+1}\right)^2 f\left(W_{n,j+1}\right) f_{xx}\left(\delta_{n,j},W_j\right)$$

is $O_{\mathbb{P}}\left(1/n\right)$. We follow the same lines to treat the term $f\left(W_{n,j}\right)\Delta_{n,j+1}$ and thus

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n-1} f\left(W_{n,j+1}\right)\Delta_{n,j}+f\left(W_{n,j}\right)\Delta_{n,j+1} &= \frac{1}{n}\sum_{j=1}^{n-1}\left(X_{\sigma_n(j)}-\frac{j}{n+1}\right) f\left(W_{n,j+1}\right) f_x\left(W_{n,j}\right)\\ &+\frac{1}{n}\sum_{j=1}^{n-1}\left(X_{\sigma_n(j+1)}-\frac{j+1}{n+1}\right) f\left(W_{n,j}\right) f_x\left(W_{n,j+1}\right)+O_{\mathbb{P}}\left(\frac{1}{n}\right)\\ &=\frac{1}{n}\sum_{j=1}^{n-1}\left(X_{\sigma_n(j)}-\frac{j}{n+1}\right) f_x\left(W_{n,j}\right)\left(f\left(W_{n,j-1}\right)+f\left(W_{n,j+1}\right)\right)+O_{\mathbb{P}}\left(\frac{1}{n}\right).\end{aligned}$$

So that, using again the expansion of $\Delta_{n,j}$ given in (33), (31), and the boundedness of f and f_{xx} to handle the second and third coordinate of $C_{n,j}$, the study of C_n reduces to that of the random vector

$$\frac{1}{n}\sum_{j=1}^{n-1}\left(X_{\sigma_n(j)}-\frac{j}{n+1}\right) f_x\left(W_{n,j}\right)\begin{pmatrix} f\left(W_{n,j-1}\right)+f\left(W_{n,j+1}\right)\\ 1\\ 2f\left(W_{n,j+1}\right)\end{pmatrix} \tag{38}$$

by the independence between σ_n and $W_1,\ldots,W_n$. In that view, let us consider the following linear combination $u(f(W_{n,j-1})+f(W_{n,j+1}))+v+2wf(W_{n,j+1})$, where $(u,v,w)\in\mathbb{R}^3$ and the empirical mean

$$\frac{1}{n}\sum_{j=1}^{n-1}\left(X_{\sigma_n(j)}-\frac{j}{n+1}\right) f_x\left(W_{n,j}\right)\times\left(u(f(W_{n,j-1})+f(W_{n,j+1}))+v+2wf(W_{n,j+1})\right). \tag{39}$$

Now it remains to apply [17, Lemma 1.4] [2] with $\chi_j=(W_{j-1},W_j,W_{j+1})$ and $\psi=\psi_{uvw}$ with

$$\psi_{uvw}\left(\frac{j-1}{n+1},\frac{j}{n+1},\frac{j+1}{n+1},\chi_j\right)=f_x\left(W_{n,j}\right)\left(u(f(W_{n,j-1})+f(W_{n,j+1}))+v+2wf(W_{n,j+1})\right), \tag{40}$$

noticing that, as $n\to\infty$, $(1/n)\sum_{j=1}^{n-1}\delta_{(j-1)/(n+1),j/(n+1),(j+1)/(n+1),\chi_j}$ converges in distribution to $Q=\mathcal{L}_{(X,X,X)}\otimes\mathcal{L}_W\otimes\mathcal{L}_W\otimes\mathcal{L}_W$ by [17, Lemma 1.1]. Thus we deduce that the empirical mean in (39) converges in distribution for any 3-uplet (u,v,w). Since any linear combination of the components of the random vector defined in (38) satisfies a CLT, so does the random vector itself. The proof of Lemma B.2 is now complete, up to the computation of the asymptotic variance-covariance matrix Σ_C done in the paragraph that follows.

[2] A slightly generalization of this lemma is required to handle the pair $(j/(n+1),(j+1)/(n+1))$ rather than the quantity j/n. Its proof comes directly following the same lines as in the proof of this lemma

Computation of the asymptotic covariance matrix Σ_C

We use the explicit expression (4) in the proof of [17, Lemma 1.4] of the asymptotic variance σ_ψ^2 (actually a slightly generalized version of the lemma) with $Q = \mathcal{L}_{(X,X,X)} \otimes \mathcal{L}_W \otimes \mathcal{L}_W \otimes \mathcal{L}_W$ and with ψ given by (40). Then taking the values $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ leads to the diagonal terms of the asymptotic variance-covariance matrix Σ_C while solving a three-dimensional system of equations provides the remaining terms. For instance, reminding that $\chi_j = (W_{j-1}, W_j, W_{j+1})$ and $W_{n,j} = (j/(n+1), W_j)$ and

$$\psi_{100}\left(\frac{j-1}{n+1}, \frac{j}{n+1}, \frac{j+1}{n+1}, \chi_j\right) = f_x\left(W_{n,j}\right)\left(f(W_{n,j-1}) + f(W_{n,j+1})\right)$$

(namely, ψ_{uvw} with $(u,v,w) = (1,0,0)$), we have

$$\begin{aligned}
\Sigma_C^{1,1} =& \int \psi_{100}(x_1, x_1', x_1'', \chi_1)\psi_{100}(x_2, x_2', x_2'', \chi_2) x_1 \wedge x_2 \wedge x_1' \wedge x_2' \wedge x_1'' \wedge x_2'' \\
&\times \mathrm{d}Q(x_1, x_1', x_1'', \chi_1)\mathrm{d}Q(x_2, x_2', x_2'', \chi_2) - \left(\int \psi_{100}(x, x', x'', \chi) x \wedge x' \wedge x'' \mathrm{d}Q(x, x', x'', \chi)\right)^2 \\
=& \mathbb{E}[(Y_1 + Y_1')(Y_2 + Y_2') f_x(X_1, W_1) f_x(X_2, W_2)(X_1 \wedge X_2)] - \mathbb{E}[(Y + Y') f_x(X, W) X]^2,
\end{aligned}$$

where we remind that $Y = f(X, W)$ and $Y' = f(X, W')$ with W' an independent copy of W (and analogously for Y_1 and Y_2). Finally, the remaining terms of Σ_C are:

$$\begin{aligned}
\Sigma_C^{2,2} &= \mathbb{E}[f_x(X_1, W_1) f_x(X_2, W_2)(X_1 \wedge X_2)] - \mathbb{E}[f_x(X, W) X]^2 \\
\Sigma_C^{3,3} &= 4\mathbb{E}[Y_1' Y_2' f_x(X_1, W_1) f_x(X_2, W_2)(X_1 \wedge X_2)] - 4\mathbb{E}[Y' f_x(X, W) X]^2 \\
\Sigma_C^{1,2} &= \Sigma_C^{2,1} = \mathbb{E}[(Y_1 + Y_1') f_x(X_1, W_1) f_x(X_2, W_2)(X_1 \wedge X_2)] - \mathbb{E}[(Y + Y') f_x(X, W) X]\mathbb{E}[f_x(X, W) X] \\
\Sigma_C^{1,3} &= \Sigma_C^{3,1} = 2\mathbb{E}[(Y_1 + Y_1') f_x(X_1, W_1) Y_2' f_x(X_2, W_2)(X_1 \wedge X_2)] - 2\mathbb{E}[(Y + Y') f_x(X, W) X]\mathbb{E}[Y' f_x(X, \\
\Sigma_C^{2,3} &= \Sigma_C^{3,2} = 2\mathbb{E}[f_x(X_1, W_1) Y_2' f_x(X_2, W_2)(X_1 \wedge X_2)] - 2\mathbb{E}[f_x(X, W) X]\mathbb{E}[Y' f_x(X, W) X].
\end{aligned}$$

B.4 Asymptotic variance σ^2 of Theorem 4.1

We have proved yet that

$$\sqrt{n}\left(\begin{pmatrix} B_n \\ C_n \end{pmatrix} - \begin{pmatrix} m_B \\ 0 \end{pmatrix}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_6\left(0, \begin{pmatrix} \Sigma_B & 0 \\ 0 & \Sigma_C \end{pmatrix}\right),$$

where the explicit expressions of m_B, Σ_B and Σ_C are given in (36) of Lemma B.1, Appendices B.2 and B.3 respectively. Applying the so-called delta method [39, Theorem 3.1] to the linear function $f(x,y) = x + y$, we conclude that

$$\sqrt{n}(Z_n - m_B) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_3\left(0, \Sigma_B + \Sigma_C\right) \tag{41}$$

Further, we notice that $\xi_n^{\text{Sobol'}}(X,Y) \overset{\mathcal{L}}{=} \Psi(Z_n)$ with $\Psi(x,y,z) = (x - y^2)/(z - y^2)$. The so-called delta method [39, Theorem 3.1] then gives

$$\sqrt{N}\left(\xi_n^{\text{Sobol'}}(X,Y) - S^X\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_1(0, \sigma^2)$$

where $S^X = \mathrm{Var}(\mathbb{E}[Y|X])/\mathrm{Var}(Y)$ is the first-order Sobol' index with respect to X and $\sigma^2 = g^\top(\Sigma_B + \Sigma_C)g$ with $g = \nabla\Psi(m_B)$. By assumption $\mathrm{Var}(Y) \neq 0$, Ψ is differentiable

at m_B and we will see in the sequel that $g^\top(\Sigma_B + \Sigma_C)g \neq 0$, so that the application of the delta method is justified. By differentiation, we get that, for any x, y, and z so that $z \neq y^2$:

$$\nabla\Psi(x,y,z) = \left(\frac{1}{z-y^2}, -2y\frac{z-x}{(z-y^2)^2}, -\frac{x-y^2}{(z-y^2)^2}\right)^\top \tag{42}$$

so that

$$g = \nabla\Psi(m_B) = \left(\frac{1}{\mathrm{Var}(Y)}, 2\mathbb{E}[Y]\frac{\mathbb{E}[YY'] - \mathbb{E}[Y^2]}{\mathrm{Var}(Y)^2}, -\frac{S^X}{\mathrm{Var}(Y)}\right)^\top = \frac{1}{\mathrm{Var}(Y)}\left(1, 2\mathbb{E}[Y](S^X - 1), -S^X\right)^\top$$

Hence the asymptotic variance σ^2 in Theorem 4.1 is finally given by $\sigma^2 = g^\top(\Sigma_B + \Sigma_C)g$ where Σ_B and Σ_C have been defined in Appendices B.2 and B.3 respectively. The matrix Σ_B rewrites as

$$\Sigma_B = \begin{pmatrix} v_{01} + 2c_{01,02} & 2c_{01,03} & 2c_{01,00} \\ 2c_{01,03} & \mathrm{Var}(Y)(1 - S^X) & 2c_{03,00} \\ 2c_{01,00} & 2c_{03,00} & v_{00} \end{pmatrix}$$

where $v_{ij} = \mathbb{E}[\mathrm{Var}(A_iA_j|X)]$, $c_{ij,kl} = \mathbb{E}[\mathrm{Cov}(A_iA_j, A_kA_l|X)]$, $A_0 = Y$, $A_1 = Y'$, $A_2 = Y''$, and $A_3 = 1$ (Y and Y'' have been defined just before (37)). The matrix Σ_C rewrites as

$$\Sigma_C = \begin{pmatrix} s^2_{\psi_{100}} & s^2_{\psi_{110}} & s^2_{\psi_{101}} \\ s^2_{\psi_{110}} & s^2_{\psi_{010}} & s^2_{\psi_{011}} \\ s^2_{\psi_{101}} & s^2_{\psi_{011}} & s^2_{\psi_{001}} \end{pmatrix}$$

where s^2_ψ and ψ_{uvw} have been defined in [17, Equation (4)] and (40) respectively.

C Proof of the asymtotic efficiency of R^1_n

Proof of Proposition 4.6. By [11, Theorems 3.4 and 3.5] and classical results on efficiency, observe that

$$U_n = \left(\widehat{T}_n, \frac{1}{n}\sum_{i=1}^n Y_i, \frac{1}{n}\sum_{i=1}^n Y_i^2\right)^\top$$

is asymptotically efficient, componentwise, for estimating $U = (\mathbb{E}[\mathbb{E}[Y|X]^2], \mathbb{E}[Y], \mathbb{E}[Y^2])^\top$. The efficiency in product space [39, Theorem 25.50] yields the joint efficiency from this componentwise efficiency. Now, we consider once again the function Ψ introduced in the proof of Theorem 4.1. Since Ψ is differentiable on $\mathbb{R}^3 \setminus \left\{(x,y,z) \,\middle|\, z \neq y^2\right\}$, the efficiency and delta method result [39, Theorem 25.47] implies that $(\Psi(U_n))_n$ is asymptotically efficient for estimating $\Psi(U)$. The conclusion follows as $\Psi(U) = S^X$.
Let us compute the minimal variance. To do so, assume that the joint distribution P of (X,Y) is absolutely continuous with respect to the Cartesian product $P_X \otimes P_Y$, namely $P(dx,dy) = f(x,y)P_X(dx)P_Y(dy)$. Then

$$\mathbb{E}[Y|X=x] = \int y f_{Y|X=x}(y)P_Y(dy) = \int y\frac{f(x,y)}{\int f(x,y)P_Y(dy)}P_Y(dy).$$

For any $t \in (0,1)$, let us introduce $f_t(x,y) := (1+th(x,y))f(x,y)$ and

$$P_t(dx,dy) := (1+th(x,y))f(x,y)P_X(dx)P_Y(dy)$$

where $h(x,y) > -1$ and $\int h(x,y)f(x,y)P_x(dx)P_Y(dy) = 0$. Now we consider the function

$$F(t) := \iint_{x,y'} \left(\frac{\int y f_t(x,y)P_Y(dy)}{\int f_t(x,y)P_Y(dy)} \right)^2 P_t(dx,dy').$$

Denoting by $G(x,t) := \int y f_t(x,y)P_Y(dy)/\int f_t(x,y)P_Y(dy)$, one gets

$$F'(t) = \iint_{x,y'} \left[2G(x,t)\frac{\partial}{\partial t}G(x,t)f_t(x,y') + G(x,t)^2 h(x,y')f(x,y') \right] P_X(dx)P_Y(dy')$$

so that $F'(0) = \langle \mathbb{E}[Y|X=x](2y - \mathbb{E}[Y|X=x]), h\rangle_P$. The interest function $I := \mathbb{E}[Y|X](2Y - \mathbb{E}[Y|X])$ has $\mathbb{E}[\mathbb{E}[Y|X]^2]$ and variance $\mathrm{Var}(\mathbb{E}[Y|X](2Y - \mathbb{E}[Y|X]))$. Hence it remains to apply the delta method to get the final (minimal) variance

$$g^\top \begin{pmatrix} \mathrm{Var}(I) & \mathrm{Cov}(I,Y) & \mathrm{Cov}(I,Y^2) \\ \mathrm{Cov}(I,Y) & \mathrm{Var}(Y) & \mathrm{Cov}(Y,Y^2) \\ \mathrm{Cov}(I,Y^2) & \mathrm{Cov}(Y,Y^2) & \mathrm{Var}(Y^2) \end{pmatrix} g$$

where $g := \nabla\Psi(U)$, and by (42),

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

Finally, one gets the minimal variance mentioned in Proposition 4.6. □

Remark C.1. This result can be also obtained making a LAN perturbation of the functional derivative on the tangent space. In this setting and following the notation of [39, Chapitre 25], let us consider the functional Φ defined by

$$\Phi(P) := \frac{\mathbb{E}_P[\mathbb{E}_P[Y|X]] - \mathbb{E}_P[Y]^2}{\mathbb{E}_P[Y^2] - \mathbb{E}_P[Y]^2}.$$

Then, with the notation P_t for $t \in (0,1)$ introduced in the above proof, one gets

$$\frac{d}{dt}\Phi(P_t)|_{t=0} = \frac{1}{\mathrm{Var}(Y)}\langle \mathbb{E}[Y|X](2Y - \mathbb{E}[Y|X]) - 2\mathbb{E}[Y]Y - S^X(Y^2 - 2\mathbb{E}[Y]Y), h\rangle_P$$

leading to $\tilde{\Phi} := \frac{1}{\mathrm{Var}(Y)}\left(2\mathbb{E}[Y]Y(1-S^X) + S^X Y^2 - \mathbb{E}[Y|X](\mathbb{E}[Y|X] - 2Y)\right)$ and the minimal variance is given by $\sigma^2_{\min} = \mathrm{Var}(\tilde{\Phi}) = \frac{1}{\mathrm{Var}(Y)^2}\mathrm{Var}\left(2\mathbb{E}[Y](1-S^X)Y + S^X Y^2 + \mathbb{E}[Y|X](\mathbb{E}[Y|X] \right.$ that coincides with the expression obtained via the delta method in Proposition 4.6.

Supplement: Technical results

We present and prove technical results that will be used in the proofs of the main results.

References

[1] A. Antoniadis. Analysis of variance on function spaces. *Statistics: A Journal of Theoretical and Applied Statistics*, 15(1):59–71, 1984.

[2] A. Auddy, N. Deb, and S. Nandy. Exact detection thresholds for chatterjee's correlation. *arXiv preprint arXiv:2104.15140*, 2021.

[3] M. Azadkia and S. Chatterjee. A simple measure of conditional dependence. *arXiv preprint arXiv:1910.12327*, 2019.

[4] E. Borgonovo. A new uncertainty importance measure. *Reliability Engineering & System Safety*, 92(6):771–784, 2007.

[5] E. Borgonovo, W. Castaings, and S. Tarantola. Moment independent importance measures: New results and analytical test cases. *Risk Analysis*, 31(3):404–428, 2011.

[6] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[7] S. Boucheron, G. Lugosi, P. Massart, et al. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14:1884–1899, 2009.

[8] B. Broto, F. Bachoc, and M. Depecker. Variance reduction for estimation of shapley effects and adaptation to unknown input distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):693–716, 2020.

[9] S. Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, pages 1–26, 2020.

[10] S. Da Veiga. Global sensitivity analysis with dependence measures. *J. Stat. Comput. Simul.*, 85(7):1283–1305, 2015.

[11] S. Da Veiga and F. Gamboa. Efficient estimation of sensitivity indices. *Journal of Nonparametric Statistics*, 25(3):573–595, 2013.

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

[13] H. Dette, K. F. Siburg, and P. A. Stoimenov. A copula-based non-parametric measure of regression dependence. *Scandinavian Journal of Statistics*, 40(1):21–41, 2013.

[14] J.-C. Fort, T. Klein, and A. Lagnoux. Global sensitivity analysis and wasserstein spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):880–921, 2021.

[15] J.-C. Fort, T. Klein, and N. Rachdi. New sensitivity analysis subordinated to a contrast. *ArXiv e-prints*, May 2013.

[16] R. Fraiman, F. Gamboa, and L. Moreno. Sensitivity indices for output on a Riemannian manifold. *arXiv e-prints*, page arXiv:1810.11591, Oct 2018.

[17] F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux. Supplementary material to "global sensitivity analysis: A novel generation of mighty estimators based on rank statistics". *Bernoulli*, 28(4), 2022.

[18] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for Sobol Pick-Freeze Monte Carlo method. *Statistics*, 50(4):881–902, 2016.

[19] F. Gamboa, T. Klein, and A. Lagnoux. Sensitivity analysis based on Cramér von Mises distance. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):522–548, Apr. 2018.

[20] F. Gamboa, T. Klein, and A. Lagnoux. A central limit theorem for generalized L-statistics. Preprint, 2021.

[21] F. Gamboa, T. Klein, A. Lagnoux, and L. Moreno. Sensitivity analysis in general metric spaces. *Reliability Engineering & System Safety*, 212:107611, 2021.

[22] F. Gamboa, T. Klein, A. Lagnoux, C. Prieur, and S. da Veiga. New estimations of sensitivity indices using kernels. Work in progress, 2021.

[23] J. Hart, P. Gremaud, and T. David. Global sensitivity analysis of high dimensional neuroscience models: an example of neurovascular coupling. *Bull Math Biol*, 2019.

[24] W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statistics*, 19:293–325, 1948.

[25] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM: Probability and Statistics*, 18:342–364, 1 2014.

[26] S. Kucherenko and S. Song. Different numerical estimators for main effect global sensitivity indices. *Reliability Engineering & System Safety*, 165:222–238, 2017.

[27] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[28] S. Orey et al. A central limit theorem for m-dependent random variables. *Duke Mathematical Journal*, 25(4):543–546, 1958.

[29] A. B. Owen. Better estimation of small sobol' sensitivity indices. *ACM Trans. Model. Comput. Simul.*, 23(2):11:1–11:17, May 2013.

[30] A. B. Owen. Variance components and generalized Sobol' indices. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):19–41, 2013.

[31] A. B. Owen, J. Dick, and S. Chen. Higher order Sobol' indices. *Information and Inference*, 3(1):59–81, 2014.

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

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

[34] H. Shi, M. Drton, and F. Han. On the power of chatterjee rank correlation, 2020.

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

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

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

[38] W. Trutschnig. On a strong metric on the space of copulas and its induced dependence measure. *Journal of mathematical analysis and applications*, 384(2):690–705, 2011.

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