Efficiency of the averaged rank-based estimator for first order Sobol index inference

Thierry Klein, Paul Rochet

Efficiency of the averaged rank-based estimator for first order Sobol index inference

Thierry Klein[1] and Paul Rochet[2]

[1]Institut de Mathématiques de Toulouse; UMR5219. Université de Toulouse
[1,2]ENAC - Ecole Nationale de l'Aviation Civile , Université de Toulouse, France.

Abstract

Among the many estimators of first order Sobol indices that have been proposed in the literature, the so-called rank-based estimator is arguably the simplest to implement. This estimator can be viewed as the empirical auto-correlation of the response variable sample obtained upon re-ordering the data by increasing values of the inputs. This simple idea can be extended to higher lags of auto-correlation, thus providing several competing estimators of the same parameter. We show that these estimators can be combined in a simple manner to achieve the theoretical variance efficiency bound asymptotically.

Keywords: Sensibility analysis, estimator averaging, asymptotic efficiency

1 Introduction

Sobol indices are by now a common tool for Global sensitivity methods which aim at detecting the most influential input variables/parameters in complex computer models. In this framework, the input variables are considered as random elements and the relative influence on the quantity of interest of each subset of its components is classically quantified by the Sobol indices, usually denoted by S (see the book by Saltelli [26] for an overview on global sensitivity analysis). These indices, based on the Hoeffding's decomposition of the variance [15], were first introduced in [23] and later revisited in the framework of sensitivity analysis in [27] (see also [28]). In a nutshell, a square integrable real-valued random variable Y, referred to as the *output*, is entirely or partially explained by a collection X of *inputs* variables. The relative influence of X on Y is quantified by the Sobol index :

$$S := \frac{\mathrm{var}(\mathbb{E}(Y|X))}{\mathrm{var}(Y)} \in [0,1].$$

In practice, an analytical expression of S is rarely available making statistical inference on Sobol indices an important question. In the last decades, several approaches were developed in the literature, each one falling into one of the four following categories.

- Those based on Monte Carlo, quasi Monte Carlo or nested Monte Carlo designs of experiments (see, e.g., [17, 22, 12]).

- Those based on spectral approaches (e.g. Fourier Amplitude Sensitivity Test (FAST) [3], Random Balance Design (RBD) [31], Effective Algorithm for computing global Sensitivity Indices (EASI). [24] and polynomial chaos expansions [30]
- Those based on the so-called Pick freeze estimator in [11, 16].
- Those based on a nearest neighbors approach [9, 19, 20, 7, 13, 8, 2, 10, 1]) or similar kernel-based methods [33, 21], studied in the particular case of first-order Sobol indices [6, 4, 25, 29, 14] and in [5] for general Sobol indices.

Theoretical properties for the last two categories are well documented, especially in the case of first order Sobol indices. Consistency and asymptotic normality have been proved for kernel estimators [4, 32], Pick Freeze [16], nearest neighbors estimators [8] as well as the rank based estimator [10]. All these methods allow to estimate simultaneously all first-order Sobol indices from a single independent and identically distributed sample (two in the case of [8]), with the exception of the Pick freeze approach which requires a specific design of experiment associated to each input.

The kernel based approach developed in [4] is shown to be asymptotically optimal in quadratic mean, with its variance approaching the efficiency bound for a regular estimator of the conditional second order moment $\eta = \mathbb{E}\big(\mathbb{E}(Y \mid X)^2\big)$. However, the method is particularly tedious to implement and the estimator not easily tractable in practice. On the contrary, the rank-based approach developed by [10] has by far the simplest implementation among all consistent methods but is sub-optimal in the sense that its variance does not reach the efficiency bound asymptotically. We show that the asymptotic variance of the rank estimator only differs from the efficiency bound by the additional term $\mathbb{E}\big(\operatorname{var}^2(Y|X)\big)$, which quantifies how far it is from optimality.

We introduce the family of lagged rank estimators $\widehat{\eta}^{(\ell)}, \ell \geq 1$ that generalizes the method of [10]. We show that each lagged rank estimator $\widehat{\eta}^{(\ell)}$ performs similarly in quadratic mean as the original, under some control over the growth of the lag ℓ relative to the sample size n. By calculating the first order asymptotic expansion of the covariance matrix of a collection of lag estimators up to some maximal lag k, we derive an asymptotically optimal combination in the spirit of estimator averaging [18]. More importantly, we show how the average estimator can be made to reach the efficiency bound of [4] by choosing k growing sufficiently slowly to infinity relative to n.

The article is organised as follows. We set the theoretical framework and the definition of the lagged rank estimators $\widehat{\eta}^{(\ell)}$ in Section 2. Their properties are investigated in Section 3, with a special focus on their joint second order moments and convergence in quadratic mean, paving the way to proving the efficiency of the averaging method. A numerical analysis to illustrate and validate the various results is presented in Section 4. The proofs and technical lemmas are postponed to the Appendix.

2 Rank estimators of Sobol indices

Let (Y, X) be a couple of random variables with Y real-valued and square-integrable. The Sobol index of Y with respect to X, which measures the part of the variance of the output Y that is "explained" by the input X, is given by

$$S := \frac{\operatorname{var}(\mathbb{E}(Y|X))}{\operatorname{var}(Y)} = \frac{\mathbb{E}\big(\mathbb{E}(Y \mid X)^2\big) - \mathbb{E}(Y)^2}{\operatorname{var}(Y)}.$$

For inference purposes, because the expectation and variance of Y do not depend on the input, the only real difficulty lies in estimating the second order conditional moment

$$\eta := \mathbb{E}\big(\mathbb{E}(Y \mid X)^2\big).$$

When X is real-valued, in which case S is generally referred to as a first-order Sobol index, a simple estimator of η can be obtained from an iid sample $(Y_1, X_1), ..., (Y_n, X_n)$ following the method developed in [10]. Let $(Y_{(i)}, X_{(i)})_{i=1,...,n}$ denote the data points sorted by increasing values of the X_i's, i.e. such that $X_{(1)} \leq ... \leq X_{(n)}$, the rank estimator of η is defined by

$$\widehat{\eta} = \frac{1}{n-1} \sum_{i=1}^{n-1} Y_{(i)} Y_{(i+1)}.$$

This estimator is known to be consistent and asymptotically Gaussian under mild conditions [10]. A natural generalization of the idea consists in defining the lagged rank estimator associated to a lag $\ell \geq 1$ as

$$\widehat{\eta}^{(\ell)} = \frac{1}{n-\ell} \sum_{i=1}^{n-\ell} Y_{(i)} Y_{(i+\ell)}.$$

In order to investigate the properties of the lagged rank estimator $\widehat{\eta}^{(\ell)}$, let us introduce some technical assumptions related to the regularity of the relation between Y and X. Let Φ, V be measurable functions such that $\Phi(X) = \mathbb{E}(Y|X)$ and $V(X) = \mathrm{var}(Y|X)$. We assume that Φ and V are bounded

$$\forall x \in \mathbb{R}\,,\ |\Phi(x)| \leq M_\Phi \ \text{ and } |V(x)| \leq M_V \tag{H1}$$

and Lipschitz

$$\forall x, x' \in \mathbb{R}\,,\ |\Phi(x) - \Phi(x')| \leq L_\Phi |x - x'| \ \text{ and } |V(x) - V(x')| \leq L_V |x - x'| \tag{H2}$$

for some positive constants M_Φ, M_V, L_Φ, L_V. Remark that under these assumptions, Φ^2 is also bounded and Lipschitz, which will be useful in later proofs. These conditions are quite mild compared to the usual assumptions for first order Sobol index inference. For instance, it is extremely common in the literature to assume that the inputs are uniformly distributed on $[0, 1]$ or have compact support. In this case, it is typically sufficient to assume that the conditional expectation and variance are continuously differentiable for (H1) and (H2) to hold.

In the sequel, we shall denote by $k = k(n)$ the total number of lags considered for our purposes. This number is allowed to increase with n but must somehow be constrained by the distribution of the inputs, in particular by their range

$$\Delta_n := X_{(n)} - X_{(1)}.$$

Typically, we want to be able to consider as many lags ℓ as possible provided that the average distance between two data points $X_{(i)}$ and $X_{(i+\ell)}$ is sufficiently small (roughly speaking, we want $X_{(i)}$ to be close enough to $X_{(i+\ell)}$ so that $Y_{(i)}$ and $Y_{(i+\ell)}$ are almost identically distributed conditionally to $X_{(i)}$). This way, $\widehat{\eta}^{(\ell)}$ should provide an accurate depiction of the second conditional moment of Y. By a telescoping argument, the average distance can be bounded by

$$\frac{1}{n-\ell} \sum_{i=1}^{n-\ell} \big(X_{(i+\ell)} - X_{(i)}\big) \leq \frac{\ell}{n-\ell} \Delta_n \leq \frac{k}{n-k} \Delta_n. \tag{1}$$

Hence, we require this term to vanish fast enough as $n \to \infty$, via the following simple assumption

$$\mathbb{E}\big(k^2 \Delta_n^2\big) = o(n). \tag{H3}$$

This condition can be understood as both a regularity assumption on the tail of the input's distribution and a restriction on the maximal number $k = k(n)$ of lags considered. It is nonetheless quite mild and can always be met unless the distribution of the inputs is heavy tailed, leading to extreme behaviors of the inputs' range Δ_n. The minimal requirement, corresponding to the situation where the distribution of the X_i's has compact support (excluding the trivial case $\Delta_n \overset{a.s.}{=} 0$), is to take $k = o(\sqrt{n})$. If the distribution of the inputs decays exponentially fast, the asymptotic behavior $\Delta_n = O_P(\log n)$ imposes the slightly stronger condition $k = o\big(\sqrt{n}/\log n\big)$, far from prohibitive in practice. Finally, remark that we do not rule out data-driven values of k, such as e.g. $k \sim n^{1/3}/\Delta_n$ which automatically satisfies (H3) regardless of the distribution of the inputs. Nevertheless, the cautious and simple $k = \lfloor n^{1/3} \rfloor$, which we use in all numerical applications, fulfills all theoretical requirements while providing a good rule of thumb for practical purposes, as discussed in Section 4.

3 Theoretical results

We are now in position to investigate some properties of the lagged rank estimators. Because we are ultimately interested in their convergence in quadratic mean, we focus on controlling the bias and variance, both for a finite sample size n and asymptotically as n grows to infinity. Only the main results are presented in this section, the detailed proofs and technical steps can be found in the Appendix.

Proposition 3.1. *Under* (H1), (H2), *we have for all* $\ell = 1, ..., k$,

$$\big|\mathbb{E}\big(\widehat{\eta}^{(\ell)}\big) - \eta\big| \leq \frac{\ell}{n-\ell}\Big(L_\Phi M_\Phi + 2M_\Phi^2\,\mathbb{E}\big(\Delta_n\big)\Big).$$

In particular, if (H3) *is also met, then* $\mathbb{E}\big(\widehat{\eta}^{(\ell)}\big) = \eta + o\big(n^{-1/2}\big)$.

This result, which is a direct consequence of Lemma 6.1 in the Appendix, illustrates how the bias of $\widehat{\eta}^{(\ell)}$ may strongly depend on the lag ℓ. We observe this phenomenon in some examples of the numerical analysis in Section 4 where the bias term is shown to highly vary in function of the lag, especially for smaller sample sizes n. Nevertheless, the variance becomes the dominating term asymptotically, as shown in the next proposition.

Proposition 3.2. *Under* (H1), (H2) *and* (H3), *we have for all* $\ell = 1, ..., k$,

$$n\,\mathrm{var}\,\big(\widehat{\eta}^{(\ell)}\big) = 4\,\mathbb{E}\big(\Phi^2(X)V(X)\big) + \mathbb{E}\big(V^2(X)\big) + \mathrm{var}\,\big(\Phi^2(X)\big) + o(1).$$

Let us compare the limit variance to that of other existing estimators of single input Sobol indices. The main term (up to the convergence rate of $1/n$), given by

$$\sigma^2_{\mathrm{rank}} = 4\,\mathbb{E}\big(\Phi^2(X)V(X)\big) + \mathbb{E}\big(V^2(X)\big) + \mathrm{var}\,\big(\Phi^2(X)\big)$$

falls short to the theoretical optimal value

$$\sigma^2_{\mathrm{opt}} = 4\,\mathbb{E}\big(\Phi^2(X)V(X)\big) + \mathrm{var}\,\big(\Phi^2(X)\big)$$

shown in [4] to be the asymptotic lower bound for the variance of an estimator of η. In the same paper, the authors propose a method that achieves the theoretical lower bound for the asymptotic variance, but relies on a preliminary non-parametric estimation of the joint density of (X, Y) along with various tuning parameters, making its construction somewhat tedious. Note that the rank estimator $\widehat{\eta}^{(\ell)}$ is asymptotically optimal if, and only if, $V(X) \overset{a.s.}{=} 0$ in which case the Sobol index is equal to one.

For the sake of comparison, the alternative estimator of η proposed in [8] and based on a nearest neighbors estimation of the conditional expectation, achieves an asymptotic theoretical variance of

$$\sigma^2_{\rm nn} = 5\,\mathbb{E}\big(\Phi^2(X)V(X)\big) + 2\mathbb{E}\big(V^2(X)\big) + 2\,\mathrm{var}\,\big(\Phi^2(X)\big).$$

While the three variances are always comparable,

$$\sigma^2_{\rm opt} \leq \sigma^2_{\rm rank} \leq \sigma^2_{\rm nn},$$

an important advantage of the nearest neighbors approach over the rank method is that it can handle the estimation of multiple inputs Sobol indices, a problem that notoriously suffers from the curse of dimensionality. More recently, a kernel approach inspired from [10] was proposed in [32] with an asymptotic theoretical variance of

$$\sigma^2_{\rm ker} = 4\,\mathbb{E}\big(\Phi^2(X)V(X)\big) + 4\,\mathrm{var}\,\big(\Phi^2(X)\big).$$

This variance is, of course, higher than the theoretical lower bound $\sigma^2_{\rm opt}$ but is not comparable to the other two.

From an implementation point of view, the rank estimator $\widehat{\eta}^{(\ell)}$ is by far the easiest to construct, with the ordering of the inputs as its main computational hurdle. Besides its simplicity, a notable advantage of the method is to provide a new estimator for each lag ℓ, with similar properties asymptotically. This feature can be exploited by combining an appropriate number of rank estimators obtained with different lags, in order to improve the estimation. The next result shows how the rank estimators $\widehat{\eta}^{(1)}, ..., \widehat{\eta}^{(k)}$ form a collection of competing estimators with symmetric behaviors asymptotically.

Proposition 3.3. *Under* (H1), (H2) *and* (H3), *we have for* $1 \leq \ell < m \leq k$,

$$n\,\mathrm{cov}\,\big(\widehat{\eta}^{(\ell)}, \widehat{\eta}^{(m)}\big) = 4\,\mathbb{E}\big(\Phi^2(X)V(X)\big) + \mathrm{var}\,\big(\Phi^2(X)\big) + o(1).$$

For a fixed k, Propositions 3.2 and 3.3 give the following first order term in the asymptotic expansion of the covariance matrix $\Sigma := \big(\Sigma_{\ell m}\big)_{\ell,m=1,\ldots,k}$ of $\widehat{\eta}^{(1)}, ..., \widehat{\eta}^{(k)}$:

$$\Sigma_{\ell m} = \lim_{n\to\infty} n\ \mathrm{cov}\,\big(\widehat{\eta}^{(\ell)}, \widehat{\eta}^{(m)}\big) = \left\{ \begin{array}{cc} \sigma^2_{\rm opt} + \mathbb{E}\big(V^2(X)\big) & \text{if } \ell = m \\ \sigma^2_{\rm opt} & \text{if } \ell \neq m \end{array} \right. \tag{2}$$

Remark that Σ is of full rank provided that $V(X)$ is not almost surely zero, which indicates that the $\widehat{\eta}^{(\ell)}$'s are linearly independent asymptotically. Therefore, it is possible to reduce the asymptotic variance of an estimator of η by considering a linear combination

$$\widehat{\eta}^{(k)}_{\rm av} = \sum_{\ell=1}^{k} \lambda_\ell\, \widehat{\eta}^{(\ell)},$$

where the weights $\lambda_\ell, \ell = 1, ..., k$ are constrained to sum up to one. This heuristics is investigated in [18] to determine the weights minimizing the asymptotic variance as a function of Σ. Although Σ is unknown in practice, the symmetrical form of Σ in this case, having the same diagonal values as well as off-diagonal values, suffices to deduce that the solution corresponds to the equal weights $\lambda_\ell = 1/k$. This simple way of combining the rank estimators actually achieves the theoretical efficiency bound of [4] under mild assumptions, as shown in the next theorem.

Theorem 3.4. *If* $k = k(n)$ *tends to infinity as* $n \to \infty$ *and the conditions* (H1), (H2) *and* (H3) *are met, the average estimator* $\widehat{\eta}^{(k)}_{\rm av}$ *obtained with equal weights* $\lambda_\ell = 1/k$ *satisfies*

$$\lim_{n\to\infty} n\,\mathrm{var}\,\big(\widehat{\eta}^{(k)}_{\rm av}\big) = 4\,\mathbb{E}\big(\Phi^2(X)V(X)\big) + \mathrm{var}\,\big(\Phi^2(X)\big) = \sigma^2_{\rm opt}.$$

The fact that the averaged rank estimator achieves the variance efficiency bound σ^2_{opt} as $n \to \infty$ is certainly encouraging, although the result concerns the actual mean square error (MSE) of the estimator and not the variance of the Gaussian limit for a regular estimator, as introduced in [4]. While the regularity and asymptotic normality of $\widehat{\eta}^{(k)}_{\text{av}}$ are to be expected under the appropriate assumptions, it has not been investigated in this paper as it deviates from the original objective of variance reduction.

4 Numerical analysis

We investigate the performances of the rank estimators and their averages in different models of the form

$$Y = \Phi(X) + \sqrt{V(X)}\,\epsilon$$

where ϵ is a standard Gaussian random variable independent from X. Each model is simulated $N = 10000$ times to give a faithful representation of the distributions of the different estimators. We show the boxplots of the rank estimators obtained for all lags from $\ell = 1$ to $\ell = 50$ for four samples sizes from $n = 100$ to $n = 2000$, which we compare to the boxplots of the averages obtained for $k = 5$ to $k = 50$ with 5 estimators added at each step.

Due to the similarities in the interpretations of the results produced from various models, we choose to discuss only two values of the conditional expectation function, namely $\Phi(X) = \sin(5X)$ and $\Phi(X) = X^2 - 3X$. For the conditional variance, all examples are generated with $V(X) = 4X^2$ as other values of V hardly had any noticeable impact on the results. For the distribution of the inputs X_i, we considered the uniform distribution on $[0, 1]$ and the standard exponential distribution.

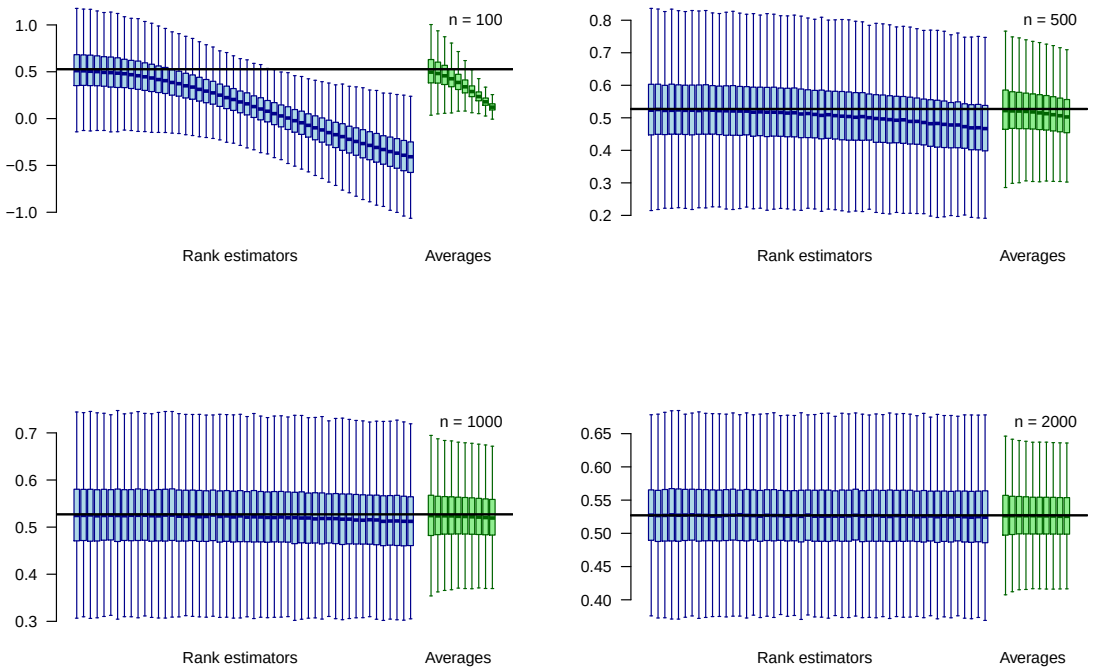

Figure 1: The model $Y = \sin(5X) + 2X\epsilon$ where X is uniformly distributed on $[0, 1]$ and ϵ is a standard Gaussian random variable independent from X. In blue, the boxplots of the rank estimators of η from lag $\ell = 1$ to $\ell = 50$. In green, the average estimators obtained with $k = 5$ to $k = 50$ by steps of 5. The sample sizes vary from $n = 100$ (top-left) to $n = 2000$ (bottom-right). The boxplots are constructed by Monte-Carlo using $N = 10^4$ repetitions.

In Figure 1, we observe that the bias of the rank estimators is important and varies strongly with the lag for the smaller sample sizes n, but does vanish asymptotically as predicted by the theory. The averaging procedure appears to improve significantly the performances of the rank estimators, as can be expected in this model with a maximal theoretical improvement of around $(\sigma^2_{\text{rank}} - \sigma^2_{\text{opt}})/\sigma^2_{\text{rank}} \approx 49\%$. The positive effect of the averaging is mostly visible on the variance (smaller inter-quartile intervals) but can not compensate for the biases, all of the same sign.

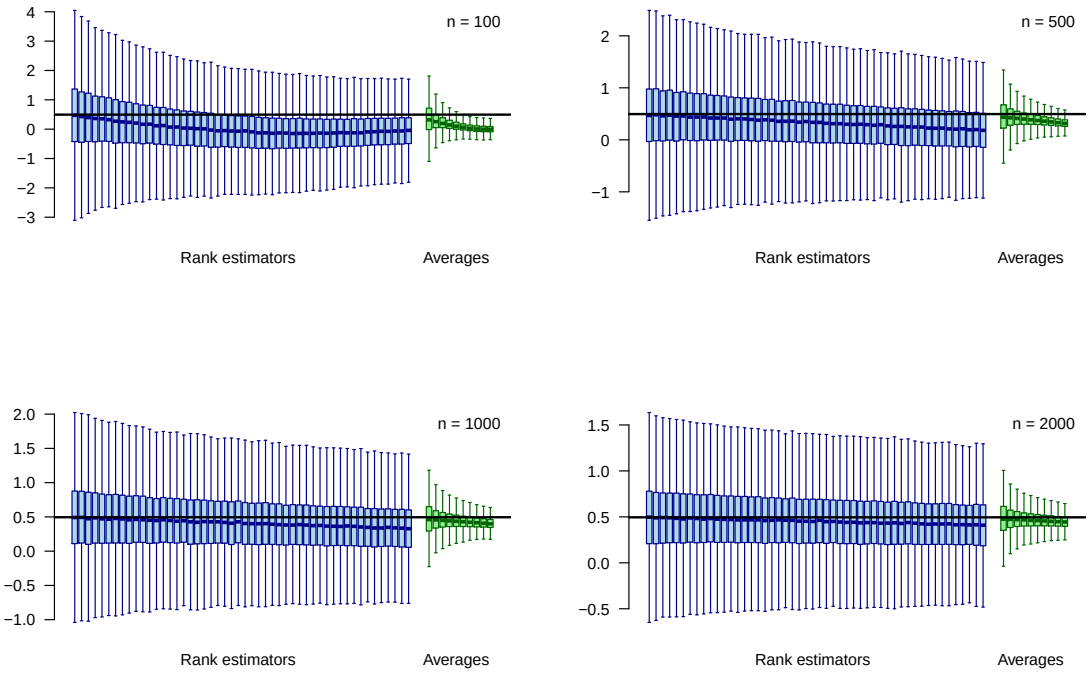

Figure 2: The model $Y = \sin(5X) + 2X\epsilon$ where X has standard exponential distribution and ϵ is a standard Gaussian random variable independent from X. In blue, the boxplots of the rank estimators of η from lag $\ell = 1$ to $\ell = 50$. In green, the average estimators obtained with $k = 5$ to $k = 50$ by steps of 5. The sample sizes vary from $n = 100$ (top-left) to $n = 2000$ (bottom-right). The boxplots are constructed by Monte-Carlo using $N = 10^4$ repetitions.

A similar behaviour can be observed in Figure 2 despite the distribution of the input X not having compact support. The maximal theoretical improvement from the averaging procedure is even higher in this case, being around $(\sigma^2_{\text{rank}} - \sigma^2_{\text{opt}})/\sigma^2_{\text{rank}} \approx 96\%$.

The convergence in quadratic mean of the rank and averaged estimators are sensible to the regularity conditions of the model, as can be seen in Figure 3. In the model $Y = \sin(5X) + 2X\epsilon$, with uniformly distributed inputs, where the regularity conditions (H1) and (H2) are satisfied, the MSEs of the various estimators do appear to behave accordingly to the theory in function of the sample size, rapidly reaching the asymptotic regime. The numerical results are not as convincing in the same model with exponentially distributed inputs, where the various estimators are slower to reach their asymptotic regime. This is especially true for the lagged rank estimator $\widehat{\eta}^{(k)}$ with k growing to infinity, although it surprisingly performs better than expected by the theory. Remark that in this case, none of the conditions (H1) and (H2) hold for the conditional variance V, which is neither bounded nor Lipschitz on the support of the inputs distribution. Nevertheless, the evolution of the MSE of the averaged estimator seems to validate in both cases the theoretical first order expansion

$$n \text{ var}\left(\widehat{\eta}^{(k)}_{\text{av}}\right) \approx \sigma^2_{\text{opt}} + \frac{1}{k}\mathbb{E}\big(V^2(X)\big)$$

derived from Equation (3) in the proof of Theorem 3.4. In all these scenarios, the squared bias account for less than 1% of the MSE, making it indistinguishable from the variance in the graphical representations.

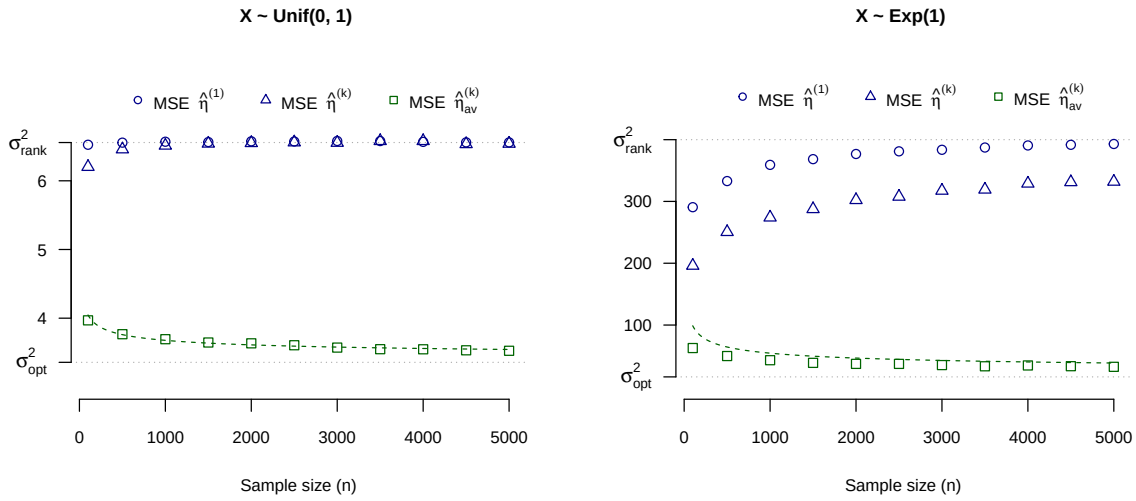

Figure 3: Evolution of the MSE's of the estimators $\widehat{\eta}^{(1)}$, $\widehat{\eta}^{(k)}$ and the average $\widehat{\eta}^{(k)}_{\text{av}}$ with $k = \lfloor n^{1/3} \rfloor$, as a function of n in the model $Y = \sin(5X) + 2X\epsilon$, under uniform (left) and exponential (right) distribution of the inputs. The MSEs are multiplied by the sample size n to highlight the convergence as $n \to \infty$. The values are obtained from Monte-Carlo estimations with $N = 10^5$ repetitions. The green dashed line represents the MSE first order asymptotic expansion for the averaged estimator derived from Equation (3).

Figure 4 illustrates how things can fall apart when the regularity conditions in (H1) and (H2) are not met for the conditional expectation function Φ. Here, the bias of the rank estimators remains high even for small lags ℓ and large sample sizes n. This is due to the large differences between consecutive extreme values in the inputs X_i, amplified by the behavior of $\Phi : x \mapsto x^2 - 3x$ (which is neither bounded nor Lipschitz in this case), causing the bias to remain high as $n \to \infty$. This example highlights the importance of the regularity conditions for the rank-based method to work and its potentially high bias, even when dealing with a single input.

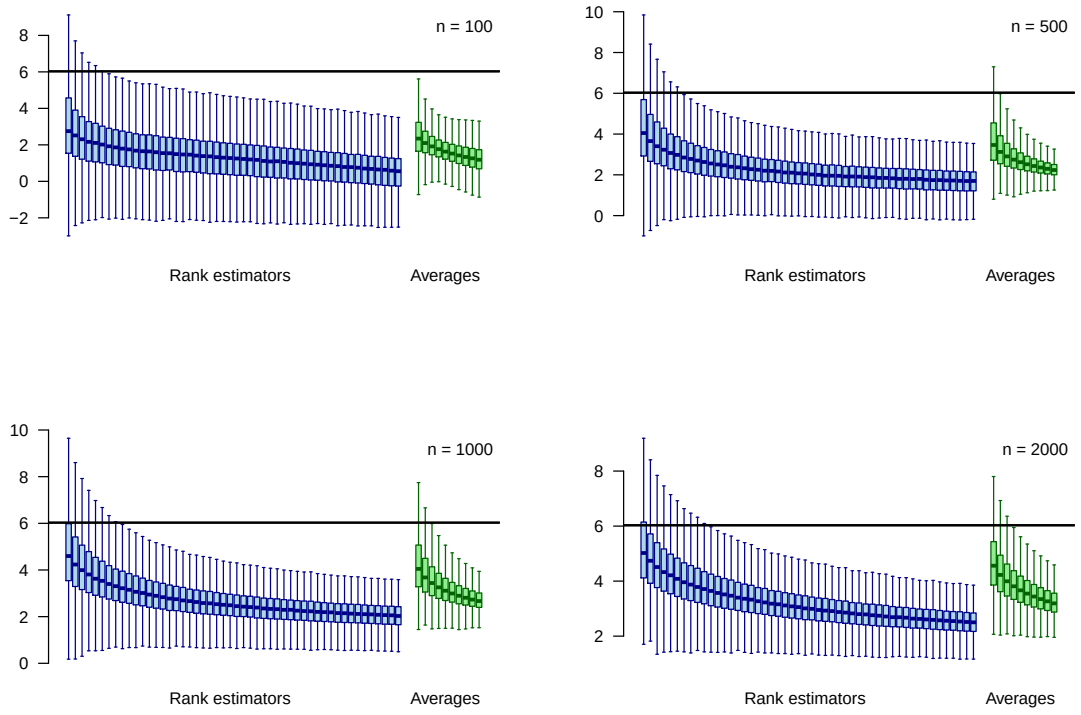

Figure 4: The model $Y = X^2 - 3X + 2X\epsilon$ where X has exponential distribution and ϵ is a standard Gaussian random variable independent from X. In blue, the boxplots of the rank estimators of η from lag $\ell = 1$ to $\ell = 50$. In green, the average estimators obtained with $k = 5$ to $k = 50$ by steps of 5. The sample sizes vary from $n = 100$ (top-left) to $n = 2000$ (bottom-right). The boxplots are constructed by Monte-Carlo using $N = 10^4$ repetitions.

5 Conclusion

The rank-based method proposed in [10] provides an easily implementable estimator for first order Sobol indices. Specifically, given real-valued output Y and input X, the second conditional moment $\eta = \mathbb{E}\big(\mathbb{E}(Y|X)^2\big)$ is estimated by the lag-one cross-product of the outputs Y_i ordered by increasing values of the inputs :

$$\widehat{\eta}^{(1)} = \frac{1}{n-1}\sum_{i=1}^{n-1} Y_{(i)}Y_{(i+1)}.$$

Under regularity conditions on the expectation and variance of the response conditionally to the input, the estimator is known to be consistent and asymptotically Gaussian. In this paper, we discuss a natural extension of the method which consists in considering rank estimators obtained from higher

order lags $\ell \geq 1$:

$$\widehat{\eta}^{(\ell)} = \frac{1}{n-\ell} \sum_{i=1}^{n-\ell} Y_{(i)} Y_{(i+\ell)} \, , \, \ell = 1, ..., k.$$

We show that these estimators share the same asymptotic properties under technical regularity conditions, provided that the maximal lag k grows sufficiently slowly relative to n. We derive a closed form expression for the asymptotic covariance matrix of the collection $(\widehat{\eta}^{(1)}, ..., \widehat{\eta}^{(k)})$, which allows to study the asymptotic behavior of the average estimator

$$\widehat{\eta}_{\text{av}}^{(k)} = \sum_{\ell=1}^{k} \lambda_\ell \, \widehat{\eta}^{(\ell)},$$

for suitable weights $\lambda_\ell, \ell = 1, ..., k$. Base on the symmetry of the covariance matrix, the averaging procedure of [18] justifies the equal weights $\lambda_\ell = 1/k$ as an asymptotically optimal choice. This is confirmed theoretically with the variance of average estimator $\widehat{\eta}_{\text{av}}^{(k)}$ reaching the efficiency bound of [4] for a regular estimator of η, whenever k grows to infinity sufficiently slowly. In practice, the rule of thumb $k = \lfloor n^{1/3} \rfloor$ provides an entirely satisfactory choice in the various simulated examples, while verifying all the technical conditions for asymptotic efficiency. The theoretical results, as well as the importance of the regularity assumptions, are well validated by the numerical analysis.

6 Appendix

Let $\mathcal{F}_n$ denote the σ-algebra generated by $X_1, ..., X_n$. The proofs of the results rely essentially on firstly investigating the distribution of the various estimators conditionally to $\mathcal{F}_n$. In particular, we exploit the fact that the $Y_{(i)}$'s remain independent conditionally to $X_1, ..., X_n$ despite the sample re-shuffling, since the permutation that orders the inputs increasingly is $\mathcal{F}_n$-measurable.

To ease notation, we shall write $\phi_i = \Phi(X_i) = \mathbb{E}(Y_i|X_i)$ and $v_i = V(X_i) = \text{var}(Y_i|X_i)$ for all $i = 1, ..., n$, and similarly for the ordered sample, e.g. $\phi_{(i)} = \Phi(X_{(i)})$, $v_{(i)} = V(X_{(i)})$.

Technical lemmas

Lemma 6.1. *If* (H1) *and* (H2) *hold, then for* $\ell = 1, ..., k$,

$$\left| \mathbb{E}\big(\widehat{\eta}^{(\ell)} \,|\, \mathcal{F}_n\big) - \frac{1}{n} \sum_{i=1}^{n} \phi_i^2 \right| \leq \frac{\ell}{n-\ell} \Big(L_\Phi M_\Phi \Delta_n + 2 M_\Phi^2 \Big).$$

Proof. Remark that $\mathbb{E}\big(Y_{(i)} Y_{(i+\ell)} \,|\, \mathcal{F}_n\big) = \phi_{(i)} \phi_{(i+\ell)}$ due to $Y_{(i)}$ and $Y_{(i+\ell)}$ being independent conditionally to $X_1, ..., X_n$. It follows

$$\left| \mathbb{E}\big(Y_{(i)} Y_{(i+\ell)} \,|\, \mathcal{F}_n\big) - \phi_{(i)}^2 \right| = \left| \phi_{(i)} \big(\phi_{(i+\ell)} - \phi_{(i)} \big) \right| \leq L_\Phi M_\Phi \big(X_{(i+\ell)} - X_{(i)} \big),$$

leading to

$$\left| \mathbb{E}\big(\widehat{\eta}^{(\ell)} \,|\, \mathcal{F}_n\big) - \frac{1}{n-\ell} \sum_{i=1}^{n-\ell} \phi_{(i)}^2 \right| \leq L_\Phi M_\Phi \frac{1}{n-\ell} \sum_{i=1}^{n-\ell} \big(X_{(i+\ell)} - X_{(i)} \big) \leq L_\Phi M_\Phi \frac{\ell}{n-\ell} \Delta_n,$$

using Equation (1). Finally, summing over $n - \ell$ terms instead on n deviates of at most

$$\left| \frac{1}{n-\ell} \sum_{i=1}^{n-\ell} \phi_{(i)}^2 - \frac{1}{n} \sum_{i=1}^{n} \phi_i^2 \right| \leq \frac{2\ell}{n-\ell} M_\Phi^2$$

by (H1), ending the proof. □

Lemma 6.2. *Under Assumptions* (H1) *and* (H2)*, we have for* $\ell < n$*,*

$$\left|(n-\ell)\operatorname{var}\left(\widehat{\eta}^{(\ell)} \,|\, \mathcal{F}_n\right) - \frac{1}{n}\sum_{i=1}^{n}\left(4\phi_i^2 v_i + v_i^2\right)\right| \le \frac{\ell}{n-\ell}\big(C_1\Delta_n + C_2\big),$$

where C_1, C_2 *are positive constants that depend only on* Φ *and* V*.*

Proof. Let $Z_{i,j,\ell} = \operatorname{cov}\left(Y_{(i)}Y_{(i+\ell)}, Y_{(j)}Y_{(j+\ell)} \,|\, \mathcal{F}_n\right)$ and for a given $i \in \{1, ..., n-\ell\}$, consider the set $S_{i,\ell} \subset \{1, ..., n\}$ of indices j such that $i, i+\ell, j, j+\ell$ are not all distinct :

$$S_{i,\ell} = \{i, i-\ell, i+\ell\} \cap \{1, ..., n\}.$$

Since $Z_{i,j,\ell} = 0$ if $j \notin S_{i,\ell}$ by independence conditionally to $\mathcal{F}_n$, we have

$$\operatorname{var}\left(\widehat{\eta}^{(\ell)} \,|\, \mathcal{F}_n\right) = \frac{1}{(n-\ell)^2}\sum_{i=1}^{n-\ell}\sum_{j\in S_{i,\ell}} Z_{i,j,\ell}.$$

For $i = \ell+1, ..., n-2\ell$, the set $S_{i,\ell}$ contains exactly three elements that can be dealt with separately:

- If $j = i$, $Z_{i,j,\ell} = \phi_{(i)}^2 v_{(i+\ell)} + \phi_{(i+\ell)}^2 v_{(i)} + v_{(i)}v_{(i+\ell)}$, and by (H1) and (H2),
$$\left|Z_{i,j,\ell} - v_{(i)}\big(2\phi_{(i)}^2 + v_{(i)}\big)\right| \le (M_\Phi L_V + M_V L_\Phi + M_V L_V)\big(X_{(i+\ell)} - X_{(i)}\big).$$
- If $j = i-\ell$ (and $i > \ell$), $Z_{i,j,\ell} = \phi_{(i-\ell)}\phi_{(i+\ell)}v_i$ and
$$\left|Z_{i,j,\ell} - \phi_{(i)}^2 v_{(i)}\right| \le M_\Phi M_V L_\Phi\big(X_{(i+\ell)} - X_{(i-\ell)}\big).$$
- If $j = i+\ell$ (and $i \le n-2\ell$), $Z_{i,j,\ell} = \phi_{(i)}\phi_{(i+2\ell)}v_{(i+\ell)}$ and
$$\left|Z_{i,j,\ell} - \phi_{(i)}^2 v_{(i)}\right| \le M_\Phi(M_V L_\Phi + M_\Phi L_V)\big(X_{(i+2\ell)} - X_{(i)}\big).$$

Gathering all three terms, we obtain for all $i = \ell+1, ..., n-2\ell$,

$$\left|\sum_{j\in S_{i,\ell}} Z_{i,j,\ell} - \left(4\phi_{(i)}^2 v_{(i)} + v_{(i)}^2\right)\right| \le \frac{C_1}{3}\big(X_{(i+2\ell)} - X_{(i-\ell)}\big)$$

for some $C_1 > 0$. Moreover, for the 2ℓ terms corresponding to $i \le \ell$ and $n-2\ell < i \le n-\ell$, we can use the crude bound

$$\left|\sum_{j\in S_{i,\ell}} Z_{i,j,\ell} - \left(4\phi_{(i)}^2 v_{(i)} + v_{(i)}^2\right)\right| \le 2(4M_\Phi^2 M_V + M_V^2).$$

Using the telescoping argument of Equation (1), we deduce

$$\left|\sum_{i=1}^{n-\ell}\sum_{j\in S_{i,\ell}} Z_{i,j,\ell} - \sum_{i=1}^{n-\ell}\left(4\phi_{(i)}^2 v_{(i)} + v_{(i)}^2\right)\right| \le \big(C_1\Delta_n + 4(4M_\Phi^2 M_V + M_V^2)\big)\ell.$$

The missing terms for $i > n-\ell$ can be bounded similarly by

$$\left|\sum_{i=1}^{n-\ell}\left(4\phi_{(i)}^2 v_{(i)} + v_{(i)}^2\right) - \sum_{i=1}^{n}\left(4\phi_{(i)}^2 v_{(i)} + v_{(i)}^2\right)\right| \le \big(4M_\phi^2 M_V + M_V^2\big)\ell$$

and the result follows easily from here, using the triangular inequality. □

Proof of Proposition 3.2

The result follows from Lemmas 6.1 and 6.2, using the variance decomposition

$$\operatorname{var}\big(\widehat{\eta}^{(\ell)}\big) = \mathbb{E}\Big(\operatorname{var}\big(\widehat{\eta}^{(\ell)} \,|\, \mathcal{F}_n\big)\Big) + \operatorname{var}\Big(\mathbb{E}\big(\widehat{\eta}^{(\ell)}|\mathcal{F}_n\big)\Big).$$

Using Lemma 6.2, the first term in the variance decomposition is easily shown to verify

$$(n-\ell)\,\mathbb{E}\Big(\operatorname{var}\big(\widehat{\eta}^{(\ell)}|\mathcal{F}_n\big)\Big) = 4\,\mathbb{E}\big(\Phi^2(X)\sigma^2(X)\big) + \mathbb{E}\big(\sigma^4(X)\big) + o(1).$$

using that $\mathbb{E}(\ell\Delta_n) \le \mathbb{E}(k\Delta_n) \le \sqrt{\mathbb{E}(k^2\Delta_n^2)} = o\big(1/\sqrt{n}\big) = o(1)$ by (H3). For the second term, we have

$$\operatorname{var}\Big(\frac{1}{n}\sum_{i=1}^{n}\phi_i^2\Big) = \frac{1}{n}\operatorname{var}\big(\Phi^2(X)\big)$$

so that Lemma 6.1 combined with (H3) give us directly

$$\operatorname{var}\Big(\mathbb{E}\big(\widehat{\eta}^{(\ell)}|\mathcal{F}_n\big)\Big) = \frac{\operatorname{var}\big(\Phi^2(X)\big)}{n} + o\Big(\frac{1}{n}\Big).$$

Hence,

$$(n-\ell)\operatorname{var}\Big(\mathbb{E}\big(\widehat{\eta}^{(\ell)}|\mathcal{F}_n\big)\Big) = \frac{n-\ell}{n}\operatorname{var}\big(\Phi^2(X)\big) + o(1) = \operatorname{var}\big(\Phi^2(X)\big) + o(1)$$

and the result follows. □

Proof of Proposition 3.3

We follow the same steps as in the proofs of Lemma 6.2 and Proposition 3.2, starting with

$$\operatorname{cov}\big(\widehat{\eta}^{(\ell)}, \widehat{\eta}^{(m)} \,|\, \mathcal{F}_n\big) = \frac{1}{(n-\ell)(n-m)}\sum_{i=1}^{n-\ell}\sum_{j\in S_{i,\ell,m}} Z_{i,j,\ell,m}$$

where $Z_{i,j,\ell,m} = \mathbb{E}\big(Y_{(i)}Y_{(i+\ell)}Y_{(j)}Y_{(j+m)} \,|\, \mathcal{F}_n\big) - \phi_{(i)}\phi_{(i+\ell)}\phi_{(j)}\phi_{(j+m)}$ and for all $i = 1, ..., n-\ell$,

$$S_{i,\ell,m} = \{i, i+\ell, i-m, i+\ell-m\} \cap \{1, ..., n\}.$$

In the (at most) four cases for $j \in S_{i,\ell,m}$ and $i \in \{m+1, ..., n-\ell-m\}$,

$$\Big|Z_{i,j,\ell,m} - \phi_{(i)}^2 v_{(i)}\Big| \le C\big(X_{(i+\ell+m)} - X_{(i-m)}\big)$$

for some constant $C > 0$. Using the same arguments, we arrive at

$$\Big|(n-\ell)\operatorname{cov}\big(\widehat{\eta}^{(\ell)}, \widehat{\eta}^{(m)} \,|\, \mathcal{F}_n\big) - \frac{4}{n}\sum_{i=1}^{n}\phi_i^2 v_i\Big| \le \frac{\ell}{n-\ell}\big(C_1'\Delta_n + C_2'\big),$$

for some constants $C_1', C_2' > 0$. On the other hand

$$\operatorname{cov}\Big(\mathbb{E}\big(\widehat{\eta}^{(\ell)} \,|\, \mathcal{F}_n\big), \mathbb{E}\big(\widehat{\eta}^{(m)} \,|\, \mathcal{F}_n\big)\Big) = \frac{\operatorname{var}\big(\Phi^2(X)\big)}{n} + o\Big(\frac{1}{n}\Big)$$

by Lemma 6.1. We conclude using the decomposition

$$\operatorname{cov}\big(\widehat{\eta}^{(\ell)}, \widehat{\eta}^{(m)}\big) = \mathbb{E}\Big(\operatorname{cov}\big(\widehat{\eta}^{(\ell)}, \widehat{\eta}^{(m)} \,|\, \mathcal{F}_n\big)\Big) + \operatorname{cov}\Big(\mathbb{E}\big(\widehat{\eta}^{(\ell)} \,|\, \mathcal{F}_n\big), \mathbb{E}\big(\widehat{\eta}^{(m)} \,|\, \mathcal{F}_n\big)\Big).$$ □

Proof of Theorem 3.4

Using Equation (2), we obtain by straightforward calculation

$$n \text{ var } \left(\widehat{\eta}_{\text{av}}^{(k)}\right) = \frac{1}{k^2} \sum_{\ell,m=1}^{k} n \text{ cov } \left(\widehat{\eta}^{(\ell)}, \widehat{\eta}^{(m)}\right) = \sigma_{\text{opt}}^2 + \frac{1}{k}\mathbb{E}\big(V^2(X)\big) + o(1), \tag{3}$$

after verification in the proofs that he residual terms $o(1)$ of Propositions 3.2 and 3.3 are negligible uniformly for all $\ell, m \leq k$. The asymptotic efficiency follows for k growing to infinity. □

References

[1] 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.

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

[3] R. Cukier, H. Levine, and K. Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of computational physics*, 26(1):1–42, 1978.

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

[5] S. Da Veiga, F. Gamboa, A. Lagnoux, T. Klein, and C. Prieur. New estimation of sobol'indices using kernels. *arXiv preprint arXiv:2303.17832*, 2023.

[6] S. Da Veiga, F. Wahl, and F. Gamboa. Local polynomial estimation for sensitivity analysis on models with correlated inputs. *Technometrics*, 51(4):452–463, 2009.

[7] L. Devroye, P. G. Ferrario, L. Györfi, and H. Walk. Strong universal consistent estimate of the minimum mean squared error. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 143–160, 2013.

[8] L. Devroye, L. Györfi, G. Lugosi, and H. Walk. A nearest neighbor estimate of the residual variance. *Electronic Journal of Statistics*, 12(1):1752–1778, 2018.

[9] L. Devroye, D. Schäfer, L. Györfi, and H. Walk. The estimation problem of minimum mean squared error. *Statistics & Decisions*, 21(1):15–28, 2003.

[10] F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux. Global sensitivity analysis: A novel generation of mighty estimators based on rank statistics. *Bernoulli*, 28(4):2345–2374, 2022.

[11] 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.

[12] T. Goda. Computing the variance of a conditional expectation via non-nested Monte Carlo. *Operations Research Letters*, 45(1):63 – 67, 2017.

[13] L. Györfi and H. Walk. On the asymptotic normality of an estimate of a regression functional. *J. Mach. Learn. Res.*, 16:1863–1877, 2015.

[14] M. B. Heredia, C. Prieur, and N. Eckert. Nonparametric estimation of aggregated sobol indices: application to a depth averaged snow avalanche model. *Reliability Engineering & System Safety*, 212:107422, 2021.

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

[16] 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.

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

[18] F. Lavancier and P. Rochet. A general procedure to combine estimators. *Computational Statistics & Data Analysis*, 94:175–192, 2016.

[19] E. Liitiäinen, F. Corona, and A. Lendasse. On nonparametric residual variance estimation. *Neural Processing Letters*, 28:155–167, 2008.

[20] E. Liitiäinen, F. Corona, and A. Lendasse. Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101(4):811–823, 2010.

[21] J.-M. Loubes, C. Marteau, and M. Solís. Rates of convergence in conditional covariance matrix with nonparametric entries estimation. *Communications in Statistics-Theory and Methods*, 49(18):4536–4558, 2020.

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

[23] K. Pearson. On the partial correlation ratio. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 91(632):492–498, 1915.

[24] E. Plischke. An effective algorithm for computing global sensitivity indices (easi). *Reliability Engineering & System Safety*, 95(4):354–360, 2010.

[25] E. Plischke and E. Borgonovo. Fighting the curse of sparsity: Probabilistic sensitivity measures from cumulative distribution functions. *Risk Analysis*, 40(12):2639–2660, 2020.

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

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

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

[29] M. Solís. Non-parametric estimation of the first-order sobol indices with bootstrap bandwidth. *Communications in Statistics-Simulation and Computation*, 50(9):2497–2512, 2021.

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

[31] S. Tarantola, D. Gatelli, and T. Mara. Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering & System Safety*, 91(6):717–727, 2006.

[32] S. D. Veiga, F. Gamboa, A. Lagnoux, T. Klein, and C. Prieur. New estimation of sobol' indices using kernels, 2023.

[33] L.-X. Zhu and K.-T. Fang. Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068, 1996.