Test comparison for Sobol Indices over nested sets of variables

Thierry Klein, Paul Rochet

Test comparison for Sobol Indices over nested sets of variables

Thierry Klein[1] and Paul Rochet[2]

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

November 4, 2021

Abstract

Sensitivity indices are commonly used to quantify the relative influence of any specific group of input variables on the output of a computer code. One crucial question is then to decide whether a given set of variables has a significant impact on the output. Sobol indices are often used to measure this impact but their estimation can be difficult as they usually require a particular design of experiment. In this work, we take advantage of the monotonicity of Sobol indices with respect to set inclusion to test the influence of some of the input variables. The method does not rely on a direct estimation of the Sobol indices and can be performed under classical iid sampling designs.

Keywords: Global sensitivity indices, Sobol indices, significance test, Donsker's Theorem

AMS subject classification 60F05,62G05, 62G20, 62E20, 62F03, 62F05.

1 Introduction

The use of complex computer models for the analysis of applications from sciences, engineering and other fields is by now routine. For instance, in the area of marine submersion, complex computer codes have been developed to simulate submersion events (see e.g. [1, 9]) while sensitivity analysis and meta-modelling are intensively used to optimize the airplanes designs [14]. Meta-models usually depend on many input variables and are computationally expensive. Thus, it is crucial to understand which of the input parameters have an influence on the output. One classical approach to deal with this kind of problem is to consider the inputs as random elements, a point of view generally called (global) sensitivity analysis. We refer to [2, 15, 16] for an overview of the practical aspects.

Sobol indices, based on the Hoeffding decomposition [8] of the output's variance, are one of the most used tool to perform global sensitivity analysis. They were first introduced in [13] and later revisited in [17]. In the general framework, a square integrable output variable Y is assumed to obey a non-parametric relation of the form

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

where the X_j's are the observed input variables, W is an extra unobserved random input (for instance, W appears naturally in the context of stochastic computer codes) and f is an unknown function. In

practice, an analytical expression for f is usually not available and the only access we have to f is through experimentation or computer code. The Sobol index of Y with respect to a subset $u \subset \{1, \dots, p\}$ of input variables is then defined by

$$S^{(u)} := \frac{\mathrm{var}\left(\mathbb{E}[Y|X_i, i \in u]\right)}{\mathrm{var}(Y)}.$$

One of the main tasks the practitioner has to deal with is to decide whether a group of variables has any influence on the output Y. This objective can be pursued by noticing that the equality of Sobol indices for nested sets $u \subset v$ reduces to the (almost sure) equality of the conditional expectations:

$$\forall u, v \ , \ u \subset v \ : \quad S^{(u)} = S^{(v)} \iff \mathbb{E}[Y|X_i, i \in u] \overset{a.s.}{=} \mathbb{E}[Y|X_i, i \in v].$$

Thus, a non-parametric notion of significance can be established by comparing Sobol indices over nested sets of input variables.

Many different estimation procedures of the Sobol indices have been proposed in the literature. Some are based on Monte-Carlo or quasi Monte-Carlo designs of experiment, see [11, 12]. More recently, a method based on nested Monte-Carlo [7] has been developed. Other estimation procedures are based on different designs of experiment using for example polynomial chaos expansions [18]. An efficient estimation of the Sobol indices can be performed through the so-called "Pick-Freeze" method, whose theoretical properties (consistency, central limit theorem, concentration inequalities and Berry-Esseen bounds) have been studied in [4, 10]. In particular, the joint central limit theorem enables to build asymptotic comparison tests on Sobol indices. However, the Pick-Freeze method requires a specific design of experiment which makes it inapplicable in the classical iid framework and computationally expensive (for instance, the p order one Sobol indices estimators need $n(p+1)$ computations of the function f). This drawback was recently partially solved in [5], where the order one Sobol indices are estimated from rank statistics in a classical iid sample scheme. Nevertheless, the absence of a joint CLT in this case makes it impossible to test hypotheses involving more than one Sobol index at a time.

The main goal of this work is to present an alternative way to build non-parametric significance tests. The originality of our approach stems from a reformulation of the null hypothesis in terms of the empirical process, thus bypassing the difficulty of having to estimate the Sobol indices. We introduce the framework and derive the theoretical tools that will be used to build the test in Section 2, while Section 3 is dedicated to the theoretical construction of level α statistical tests. In Section 4, we carry out a numerical study to compare the new statistical test procedure to the classical one introduced in [4, 10].

2 Theoretical framework

Les (Y, X, W) be random variables taking values in $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ for some $p \geq 1$ and some $q \geq 1$. We assume that there exists a square integrable function f such that

$$Y = f(X, W). \tag{2}$$

For any subset u of $\{1, ..., p\}$ and $x = (x_1, ..., x_p) \in \mathbb{R}^p$, we denote $x^{(u)} = (x_i)_{i \in u}$ with the convention $x^{(u)} = 0$ if $u = \emptyset$. We recall that the Sobol index of Y associated to $X^{(u)}$ is defined by

$$S^{(u)} = \frac{\mathrm{var}\left(\mathbb{E}[Y|X^{(u)}]\right)}{\mathrm{var}(Y)} \in [0, 1]. \tag{3}$$

Remark that Sobol indices are non-decreasing with respect to set inclusion: $u \subseteq v \implies S^{(u)} \leq S^{(v)}$. The Sobol index associated to the whole set $\{1, ..., p\}$ is denoted by S to simplify notations. Letting $\overline{u} = \{1, ..., p\} \setminus u$, we say that the variables $X^{(\overline{u})}$ are not significant to explain Y in the presence of $X^{(u)}$ if $\mathbb{E}[Y|X^{(u)}] = \mathbb{E}[Y|X]$, or equivalently, if $S^{(u)} = S$. The aim of this paper is to construct a non-parametric significance test based on the hypotheses

$$H_0 \ : S^{(u)} = S \qquad \text{against} \qquad H_1 : \ S^{(u)} < S.$$

Here, the Sobol index S over the whole collection X is not necessarily equal to one, due to the presence of the inaccessible input variable W. Therefore, any nested non-parametric significance test can be performed using an appropriate choice for X and u, leaving the remaining unused variables as part of W. For instance, the significance of a single variable X_1 via the null hypothesis $H_0 : \mathbb{E}[Y|X_1] = \mathbb{E}[Y]$ is achieved by setting $X = X_1$ and $u = \emptyset$.

A natural test for H_0 exists whenever one can construct estimators of the Sobol indices with known (or estimable) joint limit distribution. However, typical estimation methods such as Pick-Freeze usually requires a specific design of experiment. The test procedure proposed in this paper does not rely on a direct estimation of the Sobol indices and applies in the typical iid sampling design. The method makes use of an equivalent formulation of H_0 described in Lemma 2.1 below.

For two vectors $a = (a_1, ..., a_k), b = (b_1, ..., b_k)$, $a \wedge b := (a_1 \wedge b_1, ..., a_k \wedge b_k)$ denotes the component-wise minimum, while the inequality $a \leq b$ is meant as $a \wedge b = a$. The indicator function is denoted by $\mathbb{1}\{.\}$.

Lemma 2.1. *If the components of X are independent, then for all $u \subseteq \{1, ..., p\}$ the following assertions are equivalent,*

i) $S^{(u)} = S$.

ii) $\mathbb{E}[Y|X^{(u)}] = \mathbb{E}[Y|X]$.

iii) For all $x \in \mathbb{R}^p$ such that $\mathbb{P}(X \leq x) > 0$, $\mathbb{E}[Y|X^{(u)} \leq x^{(u)}] = \mathbb{E}[Y|X \leq x]$.

iv) For all $x \in \mathbb{R}^p$, $\mathbb{E}[Y\mathbb{1}\{X \leq x\}] = \mathbb{E}[Y\mathbb{1}\{X^{(u)} \leq x^{(u)}\}]\,\mathbb{P}(X^{(\overline{u})} \leq x^{(\overline{u})})$.

Proof. Let $Z = \mathbb{E}[Y|X]$ and remark that $\mathbb{E}[Y|X^{(u)}] = \mathbb{E}[Z|X^{(u)}]$. The equivalence between *i)* and *ii)* follows from the well known decomposition

$$\operatorname{var}(Z) = \operatorname{var}\big(\mathbb{E}\big[Z|X^{(u)}\big]\big) + \mathbb{E}\big[\operatorname{var}\big(Z|X^{(u)}\big)\big],$$

where the non-negative term $\mathbb{E}\big[\operatorname{var}\big(Z|X^{(u)}\big)\big]$ is zero if, and only if, $\mathbb{E}[Z|X^{(u)}] = Z$. By definition of the conditional expectation

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X^{(u)}] \iff \forall x \in \mathbb{R}^p \ , \ \mathbb{E}\big[Y\mathbb{1}\{X \leq x\}\big] = \mathbb{E}\big[\mathbb{E}[Y|X^{(u)}]\mathbb{1}\{X \leq x\}\big].$$

By the independence of the X_j's and using that $\mathbb{1}\{X \leq x\} = \mathbb{1}\{X^{(u)} \leq x^{(u)}\}\mathbb{1}\{X^{(\overline{u})} \leq x^{(\overline{u})}\}$,

$$\mathbb{E}\big[\mathbb{E}(Y|X^{(u)})\mathbb{1}\{X \leq x\}\big] = \mathbb{E}\big[Y\mathbb{1}\{X^{(u)} \leq x^{(u)}\}\big]\,\mathbb{P}\big(X^{(\overline{u})} \leq x^{(\overline{u})}\big)$$

which shows that *ii)* $\iff$ *iv)*. Finally, the equivalence *iii)* $\iff$ *iv)* follows by dividing both sides of the equality *iii)* by $\mathbb{P}(X \leq x)$. □

Assume we observe an iid sample $(Y_1, X_1), ..., (Y_n, X_n)$ drawn from the same distribution as (Y, X). For all $k \in \mathbb{N}$ and $u \subseteq \{1, ..., p\}$, let $m_k^{(u)} : x \mapsto \mathbb{E}[Y^k \mathbb{1}\{X^{(u)} \leq x^{(u)}\}]$ and denote by $\widehat{m}_k^{(u)}(.)$ its empirical counterpart:

$$\widehat{m}_k^{(u)}(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i^k \mathbb{1}\{X_i^{(u)} \leq x^{(u)}\} \ , \ x \in \mathbb{R}^p.$$

For ease of notation, we shall simply write m_k and $\widehat{m}_k$ for the case $u = \{1, ..., p\}$. By Lemma 2.1, we know that the null hypothesis $H_0 : S^{(u)} = S$ can be stated as $\xi := m_1 - m_1^{(u)} m_0^{(\overline{u})}$ being identically zero. In this logic, we use the empirical version $\widehat{\xi}$ to build a test statistics for H_0.

In the next lemma, $x^{(u)} \oplus x'^{(\overline{u})}$ denotes the vector of $\mathbb{R}^p$ with components x_i if $i \in u$ and x'_i if $i \notin u$.

Proposition 2.2. *Let $\eta = (m_1, m_1^{(u)}, m_0^{(\overline{u})})^\top$, the process $\widehat{\eta} := (\widehat{m}_1, \widehat{m}_1^{(u)}, \widehat{m}_0^{(\overline{u})})^\top$ is asymptotically Gaussian:*

$$\sqrt{n}\big(\widehat{\eta} - \eta\big) \xrightarrow[n\to\infty]{d} G,$$

where G is a 3-dimensional centered Gaussian field indexed by $\mathbb{R}^p$ with auto-covariance function

$$\Omega(x, x') := \begin{bmatrix} m_2(x \wedge x') & m_2\big((x \wedge x')^{(u)} \oplus x^{(\overline{u})}\big) & m_1\big(x^{(u)} \oplus (x \wedge x')^{(\overline{u})}\big) \\ m_2\big((x \wedge x')^{(u)} \oplus x'^{(\overline{u})}\big) & m_2^{(u)}(x \wedge x') & m_1\big(x^{(u)} \oplus x'^{(\overline{u})}\big) \\ m_1\big(x'^{(u)} \oplus (x \wedge x')^{(\overline{u})}\big) & m_1\big(x'^{(u)} \oplus x^{(\overline{u})}\big) & m_0^{(\overline{u})}(x \wedge x') \end{bmatrix} - \eta(x)\eta(x')^\top$$

for all $x, x' \in \mathbb{R}^p$.

Proof. The convergence of the finite-dimensional distributions is a straightforward consequence of the central-limit theorem. Moreover, for $t \in \mathbb{R}^p$, let

$$f_t(y, x) = \big(y\mathbb{1}\{x \leq t\}, y\mathbb{1}\{x^{(u)} \leq t^{(u)}\}, \mathbb{1}\{x^{(\overline{u})} \leq t^{(\overline{u})}\}\big) \ , \ (y, x) \in \mathbb{R} \times \mathbb{R}^p.$$

Since for two distinct $x, x' \in \mathbb{R}^p$ such that $x \leq x'$ and $x_i = x'_i$ for some $i = 1, ..., p$, we have

$$\forall t \in \mathbb{R}^p \ , \ f_t(y, x) = (y, y, 1) \implies f_t(y, x') = (y, y, 1),$$

it follows easily that $\mathcal{F} := \{f_t : t \in \mathbb{R}^p\}$ is a Vapnik-Chervonenkis class of dimension 2. Moreover, $|f_t(y, x)|$ is dominated by the square integrable function $(y, x) \mapsto \big(|y|, |y|, 1\big)$ for all $t \in \mathbb{R}^p$. By Theorem 2.5.2 in [19], $\mathcal{F}$ is a Donsker class and the result follows. □

3 The test procedure

From Proposition 2.2, the asymptotic finite-dimensional distributions of $\widehat{\xi} = \widehat{m}_1 - \widehat{m}_1^{(u)} \widehat{m}_0^{(\overline{u})}$ follows from the delta method applied to the smooth function $\phi : (s, t, u) \mapsto s - tu$ from $\mathbb{R}^3$ to $\mathbb{R}$. More precisely, given a fixed collection $\mathbf{x} = (x_1, ..., x_K)$ of points in $\mathbb{R}^p$ chosen independently from the sample, we know that the random vector $\widehat{\xi}(\mathbf{x}) = \big(\widehat{\xi}(x_1), ..., \widehat{\xi}(x_K)\big)^\top$ is asymptotically Gaussian

$$\sqrt{n}\big(\widehat{\xi}(\mathbf{x}) - \xi(\mathbf{x})\big) \xrightarrow[n\to\infty]{d} \mathcal{N}\big(0, \Sigma(\mathbf{x})\big)$$

with covariance matrix

$$\Sigma(\mathbf{x}) := \left(\nabla\phi(x_k)^\top \Omega(x_k, x_{k'})\nabla\phi(x_{k'})\right)_{k,k'=1,\dots,K}.$$

The empirical version $\widehat{\Sigma}(\mathbf{x})$ obtained by replacing the functions $m_k^{(u)}$ by their empirical counterparts $\widehat{m}_k^{(u)}$ is the expression of $\Sigma(\mathbf{x})$, is clearly a consistent estimator of $\Sigma(\mathbf{x})$ in virtue of the law of large numbers. Since the hypothesis $H_0 : S^{(u)} = S$ can be stated equivalently as $H_0 : \xi = 0$, a test can be performed by comparing the observed value of $\|\widehat{\xi}(\mathbf{x})\|$ (for a well chosen norm $\|.\|$ on $\mathbb{R}^K$) to the appropriate quantile of the asymptotic distribution under H_0. Two natural approaches are then possible:

1. If $\|.\|$ is the natural Euclidean norm on $\mathbb{R}^K$, then under H_0,

$$S := n\|\widehat{\xi}(\mathbf{x})\|^2 = n\sum_{k=1}^{K}\widehat{\xi}(x_k)^2$$

converges in distribution towards a weighted χ^2 distribution with weights given by the eigenvalues $\lambda_1, ..., \lambda_K \geq 0$ of $\Sigma(\mathbf{x})$. In other words, $n\|\widehat{\xi}(\mathbf{x})\|^2$ has the same distribution asymptotically (under H_0) as $\epsilon^\top\Sigma(\mathbf{x})\epsilon$ where ϵ is a standard Gaussian vector in $\mathbb{R}^K$. This distribution can be approximated by Monte-Carlo using the empirical version $\widehat{\Sigma}(\mathbf{x})$ instead of the unknown $\Sigma(\mathbf{x})$, in order to determine the threshold τ_α over which the hypothesis is rejected, at any given significance level $\alpha \in (0, 1)$. The Monte-Carlo part can be time consuming as a large number of replications may be needed to approximate the asymptotic distribution and corresponding quantile sufficiently well.

2. A different approach consists in normalizing the vector $\widehat{\xi}(\mathbf{x})$ is order to achieve a true (non-weighted) χ^2 asymptotic distribution under H_0. If $\Sigma(\mathbf{x})$ is invertible, with inverse $\Gamma(\mathbf{x})$, a test statistics

$$T := n\widehat{\xi}(\mathbf{x})^\top\widehat{\Gamma}(\mathbf{x})\widehat{\xi}(\mathbf{x})$$

for $\widehat{\Gamma}(\mathbf{x})$ a consistent estimator of $\Gamma(\mathbf{x})$, has asymptotic distribution $\chi^2(K)$ under H_0, as $n \to \infty$. In practice, the naive estimator

$$\widehat{\Gamma}(\mathbf{x}) = \widehat{\Sigma}(\mathbf{x})^{-1}$$

is rarely a good choice, especially if $\widehat{\Sigma}(\mathbf{x})$ is close to singular. In this case, a regularized version of the inverse leads to a better approximation of the asymptotic distribution. Typically, $\widehat{\Gamma}(\mathbf{x})$ can be obtained by truncated singular value decomposition where the eigenvalues of $\widehat{\Sigma}(\mathbf{x})$ below a certain threshold t are ignored (see for instance [3] for further details on inverse matrix regularization). The observed value of the test statistics is then compared to the quantile of the χ^2 distribution with $r = \text{rank}(\widehat{\Gamma}(\mathbf{x}))$ degrees of freedom. In the numerical study, we use the regularization threshold $t = 0.1n^{-1/3}\rho(\widehat{\Sigma}(\mathbf{x}))\lambda_1$ where λ_1 is the largest eigenvalue of $\widehat{\Sigma}(\mathbf{x})$, which ensures in particular that $r \geq 1$. Further details are discussed in Section 4.

For both these approaches, the number K of points over which the empirical process $\widehat{\xi}$ is evaluated is only constrained by the computation time. A larger experimental design $\mathbf{x}$ may improve the power of the test with no negative impact on the significance level, as we discuss in Section 4.

In practice, the x_k's may be drawn uniformly on the domain of X if it is bounded, or from an arbitrary distribution μ on $\mathbb{R}^p$. In this case, the normalized test statistics can be viewed as a Monte-Carlo approximation

$$\frac{S}{K} \approx \int n\widehat{\xi}^2 d\mu.$$

Although possible in practice, we do not recommend using the available sample $(X_1, ..., X_n)$ as the design due to the poor resulting performance of the test. If the distribution of the X_i's is known to the practitioner, we may use the same distribution to draw the x_k's. Under the alternative H_1, the power of the test highly depends on the design $\mathbf{x}$ (or the underlying distribution μ) which should ideally favor regions of the space for which ξ is far from zero, enabling the test statistics to grow more rapidly to infinity.

4 Numerical application

Let $(Y_1, X_1), ..., (Y_n, X_n)$ be an iid sample on $\mathbb{R} \times \mathbb{R}^3$ obeying the relation

$$Y_i = f(X_i) \ , \ i = 1, ..., n$$

where

$$f(x) = (2 + x_3^4) \sin(x_1) + 7 \sin^2(x_2) \ , \ x = (x_1, x_2, x_3) \in \mathbb{R}^3.$$

The X_i's are assumed independent with uniform distribution on $[-\pi, \pi]^3$. This function is commonly used in sensitivity analysis as a test case and is classically referred to as the Ishigami function.

From the two possible approaches discussed in the previous section, we choose the second one due to its faster computation time. Thus, the test statistics is given by

$$T = n\widehat{\xi}(\mathbf{x})^\top \widehat{\Gamma}(\mathbf{x})\widehat{\xi}(\mathbf{x})$$

where $\widehat{\Gamma}(\mathbf{x})$ is a regularized inverse of the empirical estimator $\widehat{\Sigma}(\mathbf{x})$, whose precise construction will be detailed below.

The experimental design $\mathbf{x} = (x_1, ..., x_K)$ is drawn from the same distribution as the original sample, namely a uniform distribution on $[-\pi, \pi]^3$. We draw $K = 10$ points to build the test. This seemingly small value ended up providing satisfactory results while keeping the computation time reasonable. The power of the test can be slightly improved by taking a larger experimental design $\mathbf{x}$ but the additional time is too much amplified in our framework where numerous replications were made to evaluate the performances of the test. For an actual application of the method where the algorithm is run only once, the computation takes at most a few minutes and the size of $\mathbf{x}$ is not much of a limiting factor.

The matrix $\widehat{\Gamma}(\mathbf{x})$ is obtained from a truncated singular value decomposition of $\widehat{\Sigma}(\mathbf{x})$. Precisely, let $\lambda_1 \geq ... \geq \lambda_K$ be the ordered eigenvalues of $\widehat{\Sigma}(\mathbf{x})$ and consider the singular value decomposition

$$\widehat{\Sigma}(\mathbf{x}) = P \, \mathrm{Diag}(\lambda_1, ..., \lambda_K) P^\top$$

where P is orthogonal (i.e. $PP^\top = P^\top P = I$). We define

$$\widehat{\Gamma}(\mathbf{x}) = P \, \mathrm{Diag}\left(g_t(\lambda_1), ..., g_t(\lambda_K)\right) P^\top$$

where g_t is the so-called truncated SVD filter function $g_t(x) = 1/x$ if $x > t$ and $g_t(x) = 0$ otherwise. The test statistics T is then compared to the quantile of the χ^2 distribution with r degrees of freedom, were $r = \mathrm{rank}(\widehat{\Gamma}(\mathbf{x}))$ is the number of eigenvalues of $\widehat{\Sigma}(\mathbf{x})$ larger than t. The hypothesis is rejected if

the observed value of T exceeds the $(1-\alpha)$-quantile of the $\chi^2(r)$ distribution. To ensure that $r > 0$, we choose t equal to a vanishing proportion $\tau_n \in (0,1)$ of the spectral radius λ_1 of $\widehat{\Sigma}(\mathbf{x})$:

$$t = \tau_n \lambda_1. \tag{4}$$

The rule of thumb $\tau_n = 0.1 n^{-1/3}$ is used in the simulations.

The test statistics and resulting p-values are calculated over $N = 10000$ replications of the experiments. Four different hypotheses are considered:

1. $H_0 : S^{(3)} = 0 \iff \mathbb{E}[Y|X_3] = \mathbb{E}[Y]$
2. $H_0 : S^{(2,3)} = S^{(2)} \iff \mathbb{E}[Y|X_2, X_3] = \mathbb{E}[Y|X_2]$
3. $H_0 : S^{(1)} = 0 \iff \mathbb{E}[Y|X_1] = \mathbb{E}[Y]$
4. $H_0 : S^{(1,3)} = S^{(1)} \iff \mathbb{E}[Y|X_1, X_3] = \mathbb{E}[Y|X_1]$

As discussed in the article, these hypotheses boil down to testing the non-parametric significance of some of the input variables, e.g. the first one reduces to testing the significance of X_3 to explain Y while the second one corresponds to testing the significance of X_3 in presence of X_2. The null hypothesis is true in the first two cases where the simulations aim to evaluate the actual significance level as function of the nominal value α the test is supposed to achieve. For the last two cases, the null hypothesis is false and the simulations aim to evaluate the power of the test.

The results are compared with the test built from the Pick-Freeze estimators of the Sobol indices presented in [6]. For each scenario, the expression in Equation (3) is used, and the p-value for the unilateral test is calculated. To ease differentiate the results of the two methods in what follows, the Pick-Freeze based test will be abbreviated to PF, while the method introduced in this paper will be referred to as the Empirical Process (EP) test.

We represent the probability of rejecting the null hypothesis for all $\alpha \in [0,1]$ to give a global view of the distribution of the p-value, although, only the discrepancies between the actual and nominal values for α smaller than say 0.1 (the range of values typically used in practice) are relevant to measure the reliability of the test procedure for practical purposes. The results are computed for three sample sizes n which designate the number of calls to the function f. We emphasize that a specific sampling design is needed for the Pick-Freeze method, which is not the case for the EP test. In particular, all four hypotheses can be tested from a unique sample by the EP approach while individual samples need to be generated for each hypothesis for the PF test. In this aspect, the EP test provides a clear advantage to reduce the number of calls to f if multiple hypotheses are to be tested.

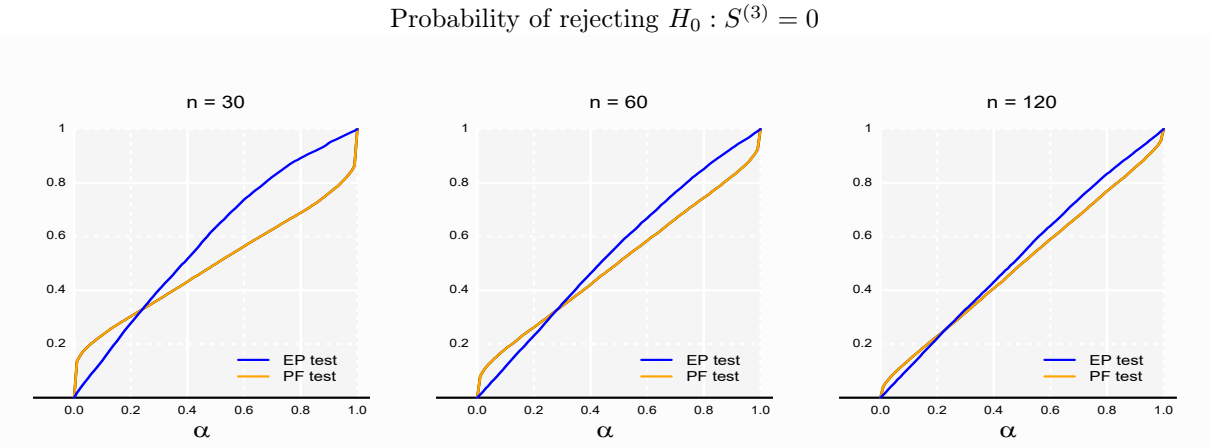

Figure 1: Estimated probability of rejecting the null hypothesis $H_0 : S^{(3)} = 0$ for the Empirical Process (EP) and Pick-Freeze (PF) tests. The empirical cdf of the tests' p-values are calculated on $N = 10000$ iterations and return the (estimated) actual significance level of the test as a function of the nominal level α.

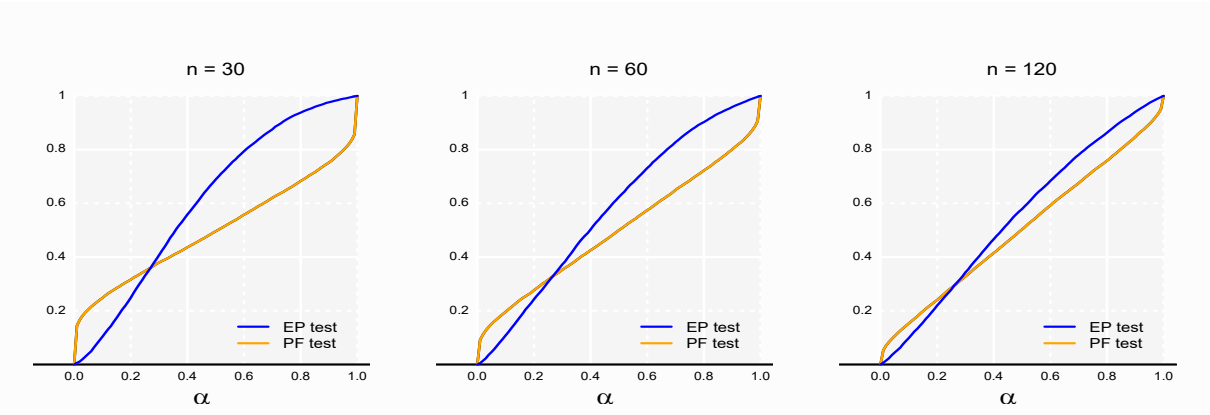

Figure 2: Estimated probability of rejecting the null hypothesis $H_0 : S^{(2,3)} = S^{(2)}$ for the EP and PF tests, as a function of the nominal significance level α.

As seen in Figures 1, 2, the EP method appears more reliable than the PF approach for the null hypotheses $H_0 : S^{(3)} = 0$ and $H_0 : S^{(2,3)} = S^{(2)}$, as the (estimated) actual significance level is closer to the nominal value. Here, the rule of thumb with $\tau_n = 0.1n^{-1/3}$ used for the TSVD regularization of $\widehat{\Sigma}(\mathbf{x})$ seems to yield a well calibrated test for a nominal significance level α below 10%. Unsurprisingly, the discrepancy is more pronounced for small sample sizes. The PF test is not well calibrated in these cases which is probably caused by a too slow convergence of the Sobol index estimator to a Gaussian distribution, on which the calculations of the critical regions of the PF test are based on.

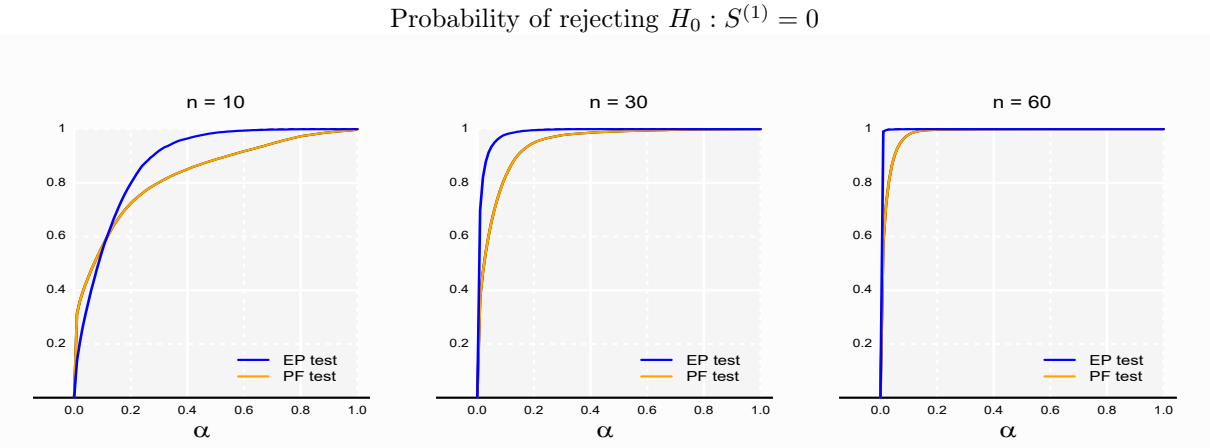

Figure 3: Estimated probability of rejecting the null hypothesis $H_0 : S^{(1)} = 0$ for the EP and PF tests. In this case where the null hypothesis is not verified, the empirical cdf of the tests' p-values return the estimated power of the test as a function of the nominal significance level α.

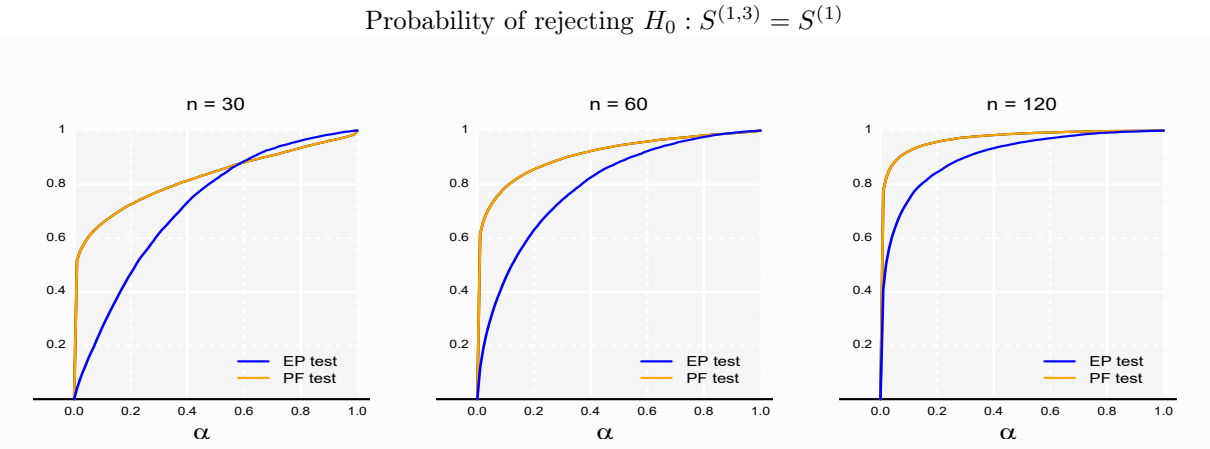

Figure 4: Power of the EP and PF tests for the null hypothesis $H_0 : S^{(1,3)} = S^{(1)}$ as a function of the significance level α.

Figures 3, 4 display the estimated probability of rightfully rejecting $H_0 : S^{(1)} = 0$ as a function of the nominal significance level α for the EP and PF tests. The EP test performs better for the simple hypothesis $H_0 : S^{(1)} = 0$. The power rapidly converges towards 1 for both tests, which conveys the high (non-parametric) significance of X_1 in this situation. On the contrary, the PF test rightfully rejects the null hypothesis $H_0 : S^{(1,3)} = S^{(1)}$ more often than the EP test. However, the higher efficiency of the PF test may be inflated by its propensity to underestimate its actual significance level, as pointed out in Figures 1, 2.

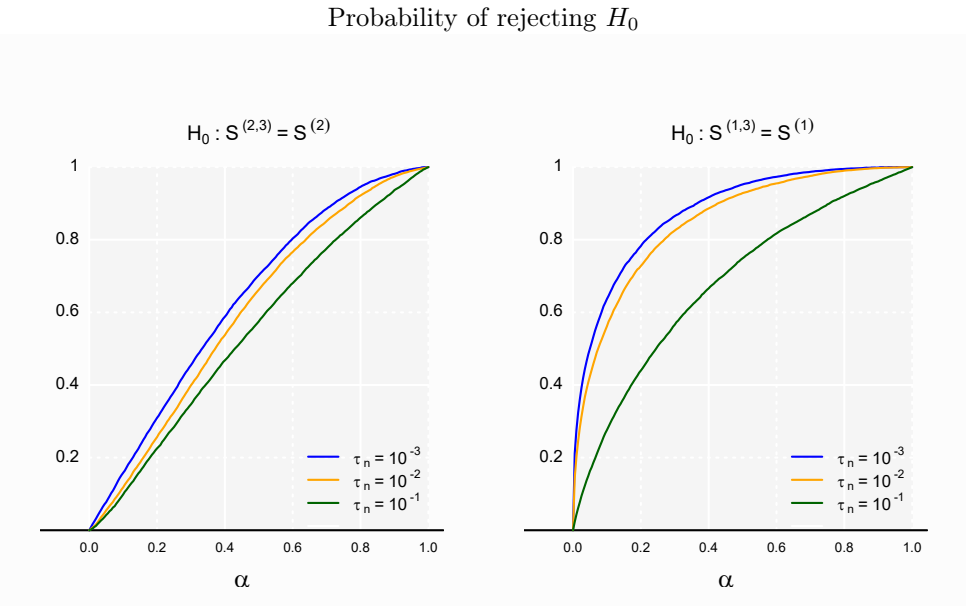

Figure 5: Estimated significance level for $H_0 : S^{(2,3)} = S^{(2)}$ (left) and power for $H_0 : S^{(1,3)} = S^{(1)}$ (right) for a TSVD threshold obtained with $\tau_n = 10^{-3}, 10^{-2}$ and 10^{-1} (see Eq. (4)). The sample size is $n = 60$ and design size $K = 10$.

Finally, the calibration of the regularization threshold used in the estimation of $\Gamma(\mathbf{x})$ has a non negligible impact on the quality of the test. In Figure 5, we show the difference in both power and significance level for three different values of τ_n. In this case, the rule of thumb gives the somewhat conservative $\tau_n = 0.1n^{-1/3} \approx 0.026$, which ensures a reliable test in term of significance level. Remark that although both thresholds $\tau_n = 10^{-2}$ and $\tau_n = 10^{-1}$ lead to similar and somewhat accurate levels, we observe a significant improvement in term of power. This suggests that the EP test procedure has room for improvement, at least through optimizing the choice of the regularization threshold.

References

[1] José Daniel Betancourt, François Bachoc, Thierry Klein, Déborah Idier, Rodrigo Pedreros, and Jeremy Rohmer. Gaussian process metamodeling of functional-input code for coastal flood hazard assessment. *Reliability Engineering and System Safety*, 198, June 2020.

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

[3] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[4] 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.

[5] Fabrice Gamboa, Pierre Gremaud, Thierry Klein, and Agnès Lagnoux. Global Sensitivity Analysis: a new generation of mighty estimators based on rank statistics. Working paper or preprint, February 2020.

[6] Fabrice Gamboa, Alexandre Janon, Thierry Klein, A Lagnoux, and Clémentine Prieur. Statistical inference for sobol pick-freeze monte carlo method. *Statistics*, 50(4):881–902, 2016.

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

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

[9] Déborah Idier, Axel Aurouet, François Bachoc, Audrey Baills, José Daniel Betancourt, Jonathan Durand, Renaud Mouche, Jeremy Rohmer, Fabrice Gamboa, Thierry Klein, Jérôme Lambert, Gonéri Le Cozannet, Sylvestre Leroy, Jessie Louisor, Rodrigo Pedreros, and Anne-Lise Véron. Toward a User-Based, Robust and Fast Running Method for Coastal Flooding Forecast, Early Warning, and Risk Prevention. *Journal of Coastal Research, Special Issue*, 95:11–15, 2020.

[10] 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.

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

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

[13] Karl 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.

[14] Nicolas Peteilh, Thierry Klein, Thierry Druot, Nathalie Bartoli, and Rhea P Liem. Challenging Top Level Aircraft Requirements based on operations analysis and data-driven models, application to take-off performance design requirements. In *AIAA AVIATION 2020 FORUM*, AIAA AVIATION 2020 FORUM, Reno, NV, United States, June 2020. American Institute of Aeronautics and Astronautics, American Institute of Aeronautics and Astronautics.

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

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

[17] 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.

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

[19] Aad W Van Der Vaart, Aad van der Vaart, Adrianus Willem van der Vaart, and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.