To appear in the *Journal of Statistical Computation and Simulation*
Vol. 00, No. 00, Month 20XX, 1–19

RESEARCH ARTICLE

Computing first-order sensitivity indices with contribution to the sample mean plot

N. Saint-Geours[a] † S. Tarantola[b] V. Kopustinskas[c] R. Bolado-Lavin[d]
[a] *Irstea, UMR TETIS, 500 rue J.F. Breton BP 5095, F-34196 Montpellier, France*
[b] *Joint Research Center of the European Commission, IPSC, Ispra, Italy*
[c] *Joint Research Center of the European Commission, IET, Ispra, Italy*
[d] *Joint Research Center of the European Commission, IET, Petten, Netherlands*

(*November 13, 2013*)

In this paper, we investigate the use of the Contribution to the Sample Mean plot (CSM plot) as a graphical tool for sensitivity analysis (SA) of computational models. We first provide an exact formula that links, for each uncertain model input X_j, the CSM plot $C_j(\cdot)$ with the first-order variance-based sensitivity index S_j. We then build a new estimate for S_j using polynomial regression of the CSM plot. This estimation procedure allows the computation of S_j from *given data*, without any SA-specific design of experiment. Numerical results show that this new S_j estimate is efficient for large sample sizes, but that at small sample sizes it does not compare well with other S_j estimation techniques based on given data, such as the EASI method.

Keywords: variance-based sensitivity analysis; importance measures; contribution to the sample mean plot; Monte Carlo simulation; Ishigami function

AMS Subject Classification: 49Q12; 65S05

1. Introduction

Global sensitivity analysis (GSA) is used to study how the variability of the output of a model can be apportioned to different sources of uncertainty in its inputs. Here, the term *model* denotes any computer code in which a response variable is calculated as a deterministic function of input variables. Originally developed in the 1990s [1], GSA is now recognized as an essential component of model building [2, 3] and has been gaining increasing acceptance in various fields over the last decade. In this paper, we focus on variance-based GSA, which relies on the decomposition of a model output variance into conditional variances [4]. So-called *first-order* and *total-order sensitivity indices* measure the main and total effect contribution of each uncertain model input to the model output variance. Uncertain model inputs are then ranked based on these sensitivity indices, to i) identify inputs that should be better scrutinized to reduce the variability of the model output (*variance-cutting*), but also to ii) simplify the model under study by fixing non-influential inputs (*factor-fixing*) [5]. More generally, variance-based GSA helps to explore the response surface of a black box computer code and to prioritize the possibly numerous processes that are involved in it.

While a great deal of GSA research has focused on the estimation of sensitivity indices, few papers have reported on an interesting issue: the computation of variance-based im-

†Corresponding author. Email: saintge@teledetection.fr

portance measures from *given data*. Indeed, most available variance-based GSA techniques require to evaluate the model on a specific, often sophisticated sampling scheme in the space of uncertain model inputs – e.g., Sobol' [6], FAST [7] or winding stairs [8, 9] sampling schemes. However, it often happens in real-world studies that the analyst has to work from a set of *pre-evaluated model runs* (given data), that is, a set of model runs that were initially carried out for a non-GSA purpose, such as uncertainty propagation or simple model exploration. In this case, model inputs may have been sampled using simple random sampling (SRS) or systematic sampling. Most of the available algorithms for sensitivity indices estimation are thus unsuitable, and other techniques are needed to allow the computation of sensitivity measures from pre-evaluated model runs.

To our knowledge, only few works have tried to tackle this issue – see [10] for a review. In 2009, Bolado–Lavin *et al.* revived a simple and versatile graphical tool named Contribution to the Sample Mean plot (CSM plot), originally proposed by Sinclair [11], and demonstrated how it could be used for GSA from a set of pre-evaluated model evaluations. More recently, Plischke [12] also proposed another graphical tool named CUSUNORO – for *cumulative sum of the normalised reordered output*, which is closely related to the CSM plot. He related this CUSUNORO plot to new estimators for first-order sensitivity indices based on correlation ratios. Before, Plischke [13] had also designed the EASI algorithm that computes first-order indices from given data using a Fast Fourier transformation approach.

In this paper, we will focus on the use of the CSM plot for variance-based GSA from given data. As we will explain later, the CSM plot is a curve in the $[0,1] \times [0,1]$ square, that visualises the contribution of a model input to the output mean across its uncertainty range. Apart from providing a profitable analysis of the input-output mapping, the CSM plot can be used for input prioritisation and regional sensitivity analysis. Indeed, Bolado–Lavin *et al.* explain that *the more the CSM curve deviates from the diagonal, the more important is the parameter* [14]. From this assertion, they developed a permutation-based statistical test for model input prioritisation, the test statistic being the maximum vertical distance of the CSM curve from the diagonal — in case of several crossings of the diagonal, sum of the maximums was proposed in [14]. They also tried to relate the CSM plot with first-order variance-based sensitivity indices, claiming from empirical considerations that *an input featuring a very low first-order effect will lead to a line close to the diagonal in the CSM plot* [14, 15]. Their numerical experiments corroborate this claim, and indicate that their CSM-based criterion is consistent with the ranking of input parameters provided by first-order variance-based sensitivity indices.

Nevertheless, to date, none of these studies has clearly demonstrated the formal link between the CSM plot and variance-based sensitivity indices. Our paper is an attempt is this direction. We will try to answer the following questions: how exactly are the CSM plot and first-order sensitivity indices related? Is it possible to compute first-order sensitivity indices from a CSM plot? By answering these questions, our goal is to open a new way to carry out variance-based GSA from pre-evaluated model evaluations. We start by defining the CSM plot (Section 2). We then give an explicit formula that links the CSM plot to first-order variance-based sensitivity indices (Section 3). From this formula, we propose new estimates for first-order indices based on given data, using polynomial regression or spline smoothing of the CSM plot (Section 4). In Section 5, we make a brief foray into the case of total-order sensitivity indices and examine their potential relation with the Contribution to the Sample Variance plot. Numerical experiments are then carried out (Section 6) on the classic Ishigami and G-Sobol functions, and the performance of our new estimators are compared with existing ones. We discuss the outcomes of our research, its limits and its connections to related works in Section 7.

2. The Contribution to the Sample Mean plot

The contribution to the sample mean plot (CSM plot) was first introduced by Sinclair [11], and then revived and improved by Bolado–Lavin *et al.* [14] as a graphical tool for GSA. The CSM plot is also known as *Lorenz curve* by social economists, who use it as a measure of inequality. In this section, we first reproduce, in a slightly different way, the presentation of Bolado–Lavin *et al.* on how to build a CSM plot from a set of model runs, then we offer to define the CSM plot in a more formal way. We also show the link between the CSM plot and the CUSUNORO curve, which was recently introduced by Plischke [12] as another graphical tool for GSA.

2.1 *CSM plotting procedure from a set of model runs*

2.1.1 Notations

Let consider a computational model $y = f(x_1, \ldots, x_m)$ with m input parameters and a scalar output y. In order to describe the uncertainty on input parameters, we treat them as independent random variables $X_1, \ldots, X_m$. The model output is also a random variable $Y = f(X_1, \ldots, X_m)$. We assume that Y has finite expectation $\mathrm{E}(Y)$ and finite variance $\mathrm{V}(Y)$. We assume that a sample of size $n \in \mathbb{N}^*$ of random variables $(X_1, \ldots, X_m)$ has been drawn. We will denote the i^{th} realisation of a parameter set with $\boldsymbol{x}_{i\cdot} = (x_{i1}, \ldots, x_{im})$, and the total sample is denoted with matrix notation $\mathbf{x} = (x_{ij})_{i=1,\ldots,n,\, j=1\ldots,m}$. The associated vector of model output is denoted by $\boldsymbol{y} = f(\mathbf{x})$, that is, $y_i = f(\boldsymbol{x}_{i\cdot})$, $i = 1, \ldots, n$.

2.1.2 Plotting procedure

The empirical Contribution to the Sample Mean plot (CSM plot) for the j^{th} input parameter X_j is built using the following procedure [14]:

(1) build the shifted output $\boldsymbol{y}^+ = \boldsymbol{y} - \min(\boldsymbol{y})$;
(2) compute the – shifted – output mean $\hat{\mu}^+ = 1/n \sum_i y_i^+$;
(3) sort the vector $(x_{1j}, \ldots, x_{nj})$ of random realisations of X_j increasingly to obtain a permuted vector $(x_{\pi_j(i)j})_{i=1,\ldots,n}$. The permutation π_j transforms the initial indices $i \in \{1, \ldots, n\}$ into shuffled indices $\pi_j(i)$ so that $x_{\pi_j(1)j} \leq \cdots \leq x_{\pi_j(n)j}$;
(4) build the permuted output vector $\pi_j(\boldsymbol{y}^+) = (y^+_{\pi_j(1)}, \ldots, y^+_{\pi_j(n)})$;
(5) for each n-quantile $q_i = \frac{i}{n}$, $i = 1, \ldots, n$, compute c_{ij}^+ as follows:

$$c_{ij}^+ = \frac{1}{n\hat{\mu}^+} \sum_{s=1}^{i} y^+_{\pi_j(s)} \tag{1}$$

The CSM plot for input parameter X_j (Fig. 1) is finally obtained by plotting the vector $(c_{1j}^+, \ldots, c_{nj}^+)$ against the n-quantiles $(q_1, \ldots, q_n)$. For convenience, we will also define the bottom-left corner of the CSM plot with $q_0 = 0$ and $c_{0j} = 0$.

2.1.3 Note on output vector shifting

Note that in the first step of the above procedure, the output vector $\boldsymbol{y}$ is shifted to get a positive vector $\boldsymbol{y}^+ = \boldsymbol{y} - \min(\boldsymbol{y})$. Thanks to this shift, the CSM values verify

$$\forall i = 1, \ldots, n, \quad c_{ij}^+ \leq c_{(i+1)j}^+$$

and all the CSM points $(q_i, c_{ij}^+)_{i=1,\dots,n}$ lie in the $[0,1] \times [0,1]$ square. Nevertheless, it is also possible to build a CSM plot without shifting the output vector $\boldsymbol{y}$. We will denote by $\mu = 1/n \sum_i y_i$ the non-shifted output mean, and by c_{ij} the non-shifted contribution to the sample mean associated to each n-quantile $q_i = \frac{i}{n}$, $i = 1, \dots, n$:

$$\forall i = 1, \dots, n, \qquad c_{ij} = \frac{1}{n\hat{\mu}} \sum_{s=1}^{i} y_{\pi_j(s)} \tag{2}$$

In this case, if the output vector $\boldsymbol{y}$ includes negative values, then the CSM values are not increasing anymore, and some of the CSM points $(q_i, c_{ij})_{i=1,\dots,n}$ may lie outside of the $[0,1] \times [0,1]$ square (Fig. 1, right). The relation between the shifted and non-shifted CSM points is simply derived from Eqn. (1) and (2), using the notation $\lambda = \min(\boldsymbol{y})/\hat{\mu}$:

$$\forall i = 1, \dots, n, \qquad c_{ij} = (1 - \lambda) c_{ij}^+ + \lambda q_i$$

In the following sections, we will focus on the non-shifted CSM plot, but we will also show why this distinction is, in the end, irrelevant regarding the computation of first-order sensitivity indices from the CSM plot.

2.2 *Formal definition of the CSM plot*

We now present a more formal definition of the Contribution to the Sample Mean plot, and explain how it relates to the plotting procedure described previously.

2.2.1 Definition

Definition 1 (Contribution to the sample mean plot) Let $\boldsymbol{X}$ be a m-dimensional random vector with joint pdf p. Let denote by $\boldsymbol{X}_{\sim j}$ the random vector of all components X_i except X_j, and $p_{\sim j}$ the associated joint pdf. The marginal pdf (resp. cdf) of random variable X_j is denoted with p_j (resp. F_j). Let $f : \mathbb{R}^m \mapsto \mathbb{R}$ be a function such that $f(\boldsymbol{X})$ has finite, non-zero expectation and finite variance. The CSM plot associated to X_j and f is a function $C_j : [0;1] \to [0;1]$ defined by:

$$\forall q \in [0;1], \qquad C_j(q) = \frac{\int_{-\infty}^{F_j^{-1}(q)} \left(\iiint_{\mathbb{R}^{m-1}} f(\boldsymbol{x})\, p_{\sim j}(\boldsymbol{x}_{\sim j}) d\boldsymbol{x}_{\sim j} \right) p_j(x_j) dx_j}{\iiint_{\mathbb{R}^m} f(\boldsymbol{x})\, p(\boldsymbol{x}) d\boldsymbol{x}} \tag{3}$$

$C_j(q)$ represents the fraction of the output mean due to the fraction q of smallest values of X_j. We notice that it is not defined if $\mathrm{E}\left[f(\boldsymbol{X})\right] = 0$.

2.2.2 Monte Carlo estimation

We will now show that the CSM plotting procedure described by Bolado–Lavin *et al.* [14] and reproduced in Section 2.1.2 leads to a Monte Carlo estimation of the quantities $C_j(q)$ for each n-quantile $q_i = \frac{i}{n}$, $i = 1, \dots, n$. To make it clear, let first rewrite Eqn. (3) as the ratio of two m-dimensional integrals, for the specific case $q = q_i = i/n$:

$$\forall i = 1, \dots, n, \qquad C_j(q_i) = \frac{\iiint_{\mathbb{R}^m} f(\boldsymbol{x})\, 1_{F_j(x_j) \leq q_i}\, p(\boldsymbol{x}) d\boldsymbol{x}}{\iiint_{\mathbb{R}^m} f(\boldsymbol{x})\, p(\boldsymbol{x}) d\boldsymbol{x}} \tag{4}$$

Let now consider a simple random sample of size $n \in \mathbb{N}^\star$ of random vector $\boldsymbol{X}$, with $\boldsymbol{x}_{s\cdot} = (x_{s1}, \ldots, x_{sm})$ the s^{th} realisation of the sample. To estimate the exact quantity $C_j(q_i)$, we may compute the Monte Carlo estimates of both the numerator and denominator in Eqn. (4):

$$\forall i = 1, \ldots, n, \qquad \widehat{C_j(q_i)} = \frac{\frac{1}{n}\sum_{s=1}^{n} f(\boldsymbol{x}_{s\cdot}) 1_{F_j(x_{sj}) \leq q_i}}{\frac{1}{n}\sum_{s=1}^{n} f(\boldsymbol{x}_{s\cdot})} \tag{5}$$

To go further, we need to approximate the indicator function $1_{F_j(x_{sj}) \leq q_i}$. A Monte Carlo estimate for the marginal cdf $F_j(x_{sj})$ is $\widehat{F_j(x_{sj})} = \pi_j(s)/n$ where π_j is the permutation previously defined in Section 2.1.2. Using $q_i = i/n$, the condition $F_j(x_{sj}) \leq q_i$ is then approximated by $\pi_j(s) \leq i$. If we note that the denominator in Eqn. (5) is equal to the output sample mean $\hat{\mu}$, we finally obtain the following Monte Carlo estimate for $C_j(q_i)$:

$$\forall i = 1, \ldots, n, \qquad \widehat{C_j(q_i)} = \frac{1}{n\hat{\mu}} \sum_{s=1}^{n} f(\boldsymbol{x}_{s\cdot}) 1_{\pi_j(s) \leq i} \tag{6}$$

We observe that this Monte Carlo estimate $\widehat{C_j(q_i)}$ is just a rewriting of the non-shifted CSM point c_{ij} defined in Eqn. (2). The three Monte Carlo estimates involved in the computation of $\widehat{C_j(q_i)}$ have a theoretical convergence speed of $1/\sqrt{n}$, yet we did not investigated further the asymptotic properties of $\widehat{C_j(q_i)}$.

2.3 *Link with the CUSUNORO curve*

Plischke [12] introduced the CUSUNORO curve – standing for *cumulative sum of the normalised reordered output* – as a new graphical tool for sensitivity analysis of computational models . The CUSUNORO curve related to model input X_j is defined by plotting the n-quantiles $q_i = \frac{i}{n}$, $i = 1, \ldots, n$ against the cumulative sums z_{ij}, $i = 1, \ldots, n$ of re-ordered model output, defined by:

$$\forall i = 1 \ldots n, \quad z_{ij} = \frac{1}{n\hat{\sigma}} \sum_{s=1}^{i} \left(y_{\pi_j(s)} - \hat{\mu} \right) \tag{7}$$

with $\hat{\sigma} = \sqrt{1/n \sum_{i=1}^{n} (y_i - \hat{\mu})^2}$ a biased estimator of the standard deviation of random output Y. As [12] already noticed it, the CSM plot and the CUSUNORO curves are just different ways of visualising the same information. Indeed, we have the following relation between the non-shifted CSM points c_{ij} and the CUSUNORO points z_{ij} [Eqn. (1) and (7)]:

$$\forall i = 1 \ldots n, \quad z_{ij} = \frac{c_{ij} - q_i}{\hat{c}_v}$$

with $\hat{c}_v = \hat{\sigma}/\hat{\mu}$ a biased estimator of the coefficient of variation $c_v = \sigma(Y)/\mathrm{E}(Y)$ of random output Y. Based on this relation, we suggest to define the CUSUNORO function

$Z_j : [0,1] \to [0,1]$ by:

$$\forall q \in [0,1], \quad Z(q) = \frac{C_j(q) - q}{c_v} \tag{8}$$

The main advantage of the CUSUNORO curve compared to the CSM plot is that it is well defined even if the model output expectation $\mathrm{E}(Y) = 0$. Its main drawback is that it is not scaled like the – shifted – CSM plot, whose points always lie in the $[0,1] \times [0,1]$ square, which is not the case for the CUSUNORO points z_{ij}. Because the CSM and CUSUNORO plots are closely related, most of the following discussion on the CSM plot will also hold for the CUSUNORO curve.

3. Link between CSM plot and first-order sensitivity indices

We will now give an exact relation between the CSM plot $C_j(\cdot)$ and the first-order variance-based sensitivity index S_j associated to model input X_j. Let first recall that S_j is defined by:

$$S_j = \frac{\mathrm{Var}_{X_j}\left(\mathrm{E}_{\boldsymbol{X}_{\sim j}}\left[Y \mid X_j\right]\right)}{\mathrm{V}(Y)}$$

S_j measures the main effect of input parameter X_j on the variance of model output. It is the expected share of output variance that would be reduced if input parameter X_j was fixed. First-order indices verify $S_j \in [0,1]$ and $\sum_j S_j \leq 1$. They can be used to identify the model inputs that account, on their own, for most of the model output variability. Please refer to [5] for an in-depth discussion on the definition and properties of first-order sensitivity indices.

PROPOSITION 1 *Let $C_j(q)$ (resp. $Z_j(q)$) denote the CSM plot (resp. the CUSUNORO plot) associated with the j^{th} input parameter X_j, S_j denote the first-order variance-based sensitivity index of X_j with respect to model output Y, and $c_v = \sigma(Y)/\mathrm{E}(Y)$ denote the coefficient of variation of model output Y. We have:*

$$S_j = \frac{1}{c_v^2} \cdot \int_0^1 \left[\frac{\mathrm{d}}{\mathrm{d}q}\left(C_j(q) - q\right)\right]^2 dq \tag{9a}$$

and

$$S_j = \int_0^1 \left[\frac{\mathrm{d}Z_j}{\mathrm{d}q}(q)\right]^2 dq \tag{9b}$$

Proof. The proof is straightforward using the definitions of the CSM plot and first-order sensitivity indices. The CSM plot $C_j(q)$ defined in Eqn. (3) can be written using conditional expectation:

$$\forall q \in [0;1], \qquad C_j(q) = \frac{1}{\mathrm{E}(Y)} \int_{-\infty}^{F_j^{-1}(q)} \mathrm{E}\left[Y \mid X_j = x_j\right] p_j(x_j) dx_j \tag{10}$$

By derivating Eqn. (10) with respect to q, using the fact that $(F_j^{-1})'(q) = 1/p_j\left(F_j^{-1}(q)\right)$, we obtain an expression of the derivative of $C_j(q)$:

$$\forall q \in [0;1], \quad \frac{\mathrm{d}C_j}{\mathrm{d}q}(q) = \frac{\mathrm{E}\left[Y \mid X_j = F_j^{-1}(q)\right]}{\mathrm{E}(Y)} \tag{11}$$

We can make the right member of Eqn. (11) appear from the definition of variance-based first-order sensitivity index S_j of input parameter X_j with respect to output Y:

$$\begin{aligned} S_j &= \frac{\mathrm{Var}_{X_j}\left(\mathrm{E}_{\boldsymbol{X}_{\sim j}}\left[Y \mid X_j\right]\right)}{\mathrm{V}(Y)} \qquad \text{(definition from Saltelli } \textit{et al.} \text{ [5])} \\ &= \frac{1}{\mathrm{V}(Y)}\mathrm{E}\left[\left(\mathrm{E}\left[Y \mid X_j\right] - \mathrm{E}(Y)\right)^2\right] \\ &= \frac{1}{\mathrm{V}(Y)}\int_{\mathbb{R}} \left(\mathrm{E}\left[Y \mid X_j = x_j\right] - \mathrm{E}(Y)\right)^2 p_j(x_j)dx_j \end{aligned}$$

Using transformation $x_j = F_j^{-1}(q)$ we have:

$$\begin{aligned} S_j &= \frac{1}{\mathrm{V}(Y)}\int_0^1 \left(\mathrm{E}\left[Y \mid X_j = F_j^{-1}(q)\right] - \mathrm{E}(Y)\right)^2 dq \\ &= \frac{\mathrm{E}(Y)^2}{\mathrm{V}(Y)}\int_0^1 \left(\frac{\mathrm{E}\left[Y \mid X_j = F_j^{-1}(q)\right]}{\mathrm{E}(Y)} - 1\right)^2 dq \end{aligned}$$

Using Eqn. (11) and denoting with $c_v = \sigma(Y)/\mathrm{E}(Y)$ the coefficient of variation of random output vector Y, we finally obtain the expected relation between first-order sensitivity index S_j and $C_j(q)$. The relation between S_j and the CUSUNORO plot $Z_j(q)$ is simply obtained by combining Eqn. (8) with Eqn. (9a). ■

Let us look briefly at what Proposition 1 means. We observe that Eqn. (9a) involves the derivative of $(C_j(q) - q)$, which is the vertical distance of the CSM plot to the diagonal in the $[0,1] \times [0,1]$ square. Hence, the more the CSM plot deviates from the diagonal, the larger the value of S_j, that is, the larger the contribution of input parameter X_j to the variance of the model output. This result is in line with previous assertions on the CSM plot [14, 15]. Equation (9a) also offers a new way to estimate first-order sensitivity indices from a CSM plot: we discuss this point in the following Section 4.

4. Estimation of first-order sensitivity indices from the CSM plot

4.1 *Overview*

We now assume that a CSM plot (i.e., a set of points $(q_i, c_{ij})_{i=1,\ldots,n}$) has been built for the j^{th} input parameter from a set of n model runs, following the plotting procedure described in Section 2.1. We can use Eqn. (9a) to approximate first-order sensitivity index S_j from this CSM plot. A key point in Eqn. (9a) is that the sensitivity index S_j is invariant under any linear transformation of model output $Y \to aY + b$, $a \in \mathbb{R}^\star$, $b \in \mathbb{R}$. As

a consequence, S_j can be computed from a CSM plot built on the original output vector $\boldsymbol{y}$, or from a CSM plot built on any scaled and shifted output vector $\boldsymbol{y}^\star = a\boldsymbol{y} + b$. In particular, S_j can be equally computed from the shifted output vector $\boldsymbol{y}^+ = \boldsymbol{y} - \min(\boldsymbol{y})$, or from the non-shifted output vector $\boldsymbol{y}$. Hence, we can equally use the shifted CSM points c_{ij}^+ [Eqn. (1)] or the non-shifted CSM points c_{ij} [Eqn. (2)] to estimate S_j: we will now stop distinguishing between these two options. Besides, in order to keep the notation short, the dependency of $C_j(q)$ and c_{ij} on j will now be dropped, and we will simply write $C(q)$ or c_i.

The coefficient of variation c_v in Eqn. (9a) can be straightforwardly estimated from the output vector $\boldsymbol{y}$: $\hat{c}_v = \hat{\sigma}/\hat{\mu}$ with $\hat{\mu}$ and $\hat{\sigma}^2$ the unbiased estimators of $\mathrm{E}(Y)$ and $\mathrm{V}(Y)$, respectively. Alternatively, it can also be estimated from the set of CSM points $(c_i)_{i=1,\dots,n}$ using Eqn. (1):

$$\widehat{c_v} = \sqrt{\frac{n^2}{n-1}\sum_{i=1}^{n-1}(c_{i+1} - c_i - \frac{1}{n})^2} \tag{12}$$

Next, the integral $I = \int_0^1 (C'(q) - 1)^2 \, dq$ can be numerically approximated from the set of points $(q_i, c_i)_{i=1,\dots,n}$. Sensitivity index S_j is then approximated by:

$$\hat{S}_j = \frac{\hat{I}}{\hat{c}_v^2} \tag{13}$$

We describe in the following subsections 4.2 to 4.4 three different techniques to compute $\hat{I}$ and $\hat{S}_j$: i) finite difference scheme and Simpsons's rule; ii) polynomial regression; and iii) spline smoothing.

4.2 *Finite difference scheme*

In this approach, we follow three steps. We first approximate the CSM derivative $C'(q_i)$ at n-quantiles q_i from the sample points $(q_i, c_i)_{i=1,\dots,n}$, using a finite difference scheme. In order to get enough smoothing of the derivative, we choose a 11 points finite difference scheme. The approximated derivative $c'_i \approx C'(q_i)$ at quantile $q_i = \frac{i}{n}$ is given by:

$$\begin{aligned}
&\forall i \in \{5, \dots, n-5\}, \\
&\quad c'_i \approx \frac{n}{2520} \cdot [2 \cdot c_{i+5} - 25 \cdot c_{i+4} \\
&\qquad + 150 \cdot c_{i+3} - 600 \cdot c_{i+2} \\
&\qquad + 2100 \cdot c_{i+1} - 2100 \cdot c_{i-1} + 600 \cdot c_{i-2} \\
&\qquad - 150 \cdot c_{i-3} + 25 \cdot c_{i-4} - 2 \cdot c_{i-5}]
\end{aligned} \tag{14}$$

Then, we use the composite Simpson's rule to approximate the integral $I = \int_0^1 (C'(q) - 1)^2 \, dq$ from the set of sample points $(c'_i)_{i=5,\dots,n-5}$. Finally, we compute the approximate sensitivity index with $\hat{S}_j = \hat{I}/\hat{c}_v^2$.

4.3 *Polynomial regression of the CSM plot*

In this second method, we proceed in two steps. First, we fit a polynomial model to the CSM sample points $(q_i, c_i)_{i=1,\dots,n}$. Instead of using the canonical polynomial basis $(q^k)_{k\in\mathbb{N}}$, which is known to lead to unstable regression results, we use the shifted Legendre polynomials $(P_k(q))_{k\in\mathbb{N}}$, which form an orthonormal basis of $L^2([0;1])$[1]:

$$\forall i = 1, \dots, n, \quad c_i = \sum_{k=0}^{d} \alpha_k P_k(q_i) + \epsilon_i \tag{15}$$

with residuals ϵ_i assumed to follow a zero-mean normal distribution. The coefficients α_k are estimated with least squares regression. The maximal polynomial order $d \in \mathbb{N}^*$ can be selected by minimizing the corrected AIC information criterion:

$$\mathrm{AICc}(d) = \frac{n}{2} + \frac{n \cdot (d+2)}{n-d-3} + \frac{n}{2} \cdot \log\left(\frac{2\pi}{n}\sum_{i=1}^{n} \epsilon_i(d)^2\right)$$

with $(\epsilon_i(d))_{i=1,\dots,n}$ the regression residuals obtained with a polynomial model of maximal degree d. Then, using the approximation $C(q) \approx \sum_k \alpha_k P_k(q)$, we get an approximation of the integral $I = \int_0^1 (C'(q)-1)^2\, dq$:

$$\hat{I} = \int_0^1 \left[\left(\sum_{k=1}^{d} \alpha_k P_k'(q)\right) - 1\right]^2 dq \tag{16}$$

In order to obtain a more pleasant notation, we use the fact that $P_1'(q) = 2$ to define modified coefficients $(\tilde{\alpha}_k)_{k=1,\dots,d}$ as equal to coefficients $(\alpha_k)_{k=1,\dots,d}$ except for $\tilde{\alpha}_1 = \alpha_1 - \frac{1}{2}$, and we obtain:

$$\begin{aligned}\hat{I} &= \int_0^1 \left[\sum_{k=1}^{d} \tilde{\alpha}_k P_k'(q)\right]^2 dq \\ &= \sum_{k,l=1}^{d} \tilde{\alpha}_k \tilde{\alpha}_l \int_0^1 P_k'(q) P_l'(q) dq\end{aligned}$$

The exact value of the integral $I_{kl} = \int_0^1 P_k'(q)P_l'(q)dq$ can be computed from classical results on Legendre polynomials (see Appendix A for a proof):

$$\forall (k,l) \in \mathbb{N}^2,\ k \le l \qquad I_{kl} = 2k\,(k+1)\,1_{\{(k+l)\,\in\, 2\mathbb{N}\}}$$

---

[1]Shifted Legendre polynomial P_k are defined by $P_k(q) = P_k^{(s)}(2q-1)$ with $P_k^{(s)}$ the standardized Legendre polynomials, which are given by the Rodrigue's formula [16, p.785, Eqn. 22.11.5] :

$$\forall k \in \mathbb{N}, \forall q \in [-1,1], \quad P_k^{(s)}(q) = \frac{(-1)^k}{2^k \cdot k!} \frac{\mathrm{d}^k}{\mathrm{d}q^k}\left[(q^2-1)^k\right]$$

We thus obtain the following estimator for first-order sensitivity index S_j:

$$\hat{S}_j = \frac{2}{\hat{c_v}^2} \sum_{\substack{k,l=1 \\ k+l \in 2\mathbb{Z}}}^{d} \tilde{\alpha}_k \tilde{\alpha}_l \cdot \min(k,l) \left[\min(k,l)+1\right] \tag{17}$$

Using Eqn. (8), we can find a similar relation to estimate S_j from the polynomial regression of the CUSUNORO curve $Z_j(q)$:

$$\hat{S}_j = 2 \sum_{\substack{k,l=1 \\ k+l \in 2\mathbb{Z}}}^{d} \beta_k \beta_l \cdot \min(k,l) \left[\min(k,l)+1\right] \tag{18}$$

where $(\beta)_{k\in\mathbb{N}}$ are the regression coefficients of the polynomial model $Z(q) = \sum_k \beta_k P_k(q)$.

4.4 *Spline smoothing*

In this third method, we proceed in four steps. We first fit a non-parametric regression model $\widehat{C}(q)$ (spline of order $o \in \mathbb{N}$) to the set of CSM points $(q_i, c_i)_{i=1,\dots,n}$. We then use this spline model to estimate CSM derivative at N new points $(q_s = \frac{s}{N})_{s=1,\dots,N}$ with $N = 10000$. Next, we approximate the integral $\hat{I} \approx I$ from the N points $(q_s, \widehat{C}'(q_s))_{s=1,\dots,N}$ using composite Simpson's rule. We finally approximate sensitivity index with $\hat{S}_j = \hat{I}/\hat{c}_v^2$.

5. Contribution to the sample variance plot and total-order sensitivity indices

In the light of our findings on the CSM plot, it is temptating to formulate a new question: is there any graphical tool, analogous to the CSM plot, that would allow the estimation of total-order sensitivity indices ST_j? A possible track to investigate is that of the Contribution to Sample Variance plot (CSV plot) introduced by Tarantola *et al.* [15]. Hence, before moving to the presentation of two numerical test cases in the next Section, we will briefly discuss the possible use of the CSV plot to compute total-order sensitivity indices.

5.1 *Definition of the CSV plot*

Definition 2 (Contribution to the sample variance plot) Let $\boldsymbol{X}$ be a m-dimensional random vector with joint pdf p. Let denote by $\boldsymbol{X}_{\sim j}$ the random vector of all components X_i except X_j, and $p_{\sim j}$ the associated joint pdf. The marginal pdf (resp. cdf) of random variable X_j is denoted with p_j (resp. F_j). Let $f : \mathbb{R}^m \mapsto \mathbb{R}$ be a function such that $Y = f(\boldsymbol{X})$ has finite, non-zero expectation and finite variance. The CSV plot associated to X_j and f is a function $CSV_j : [0;1] \to [0;1]$ defined by:

$$\forall q \in [0,1], \qquad CSV_j(q) = \frac{1}{\mathrm{V}(Y)} \int_{-\infty}^{F_j^{-1}(q)} \int_{\mathbb{R}^{m-1}} \left[f(\boldsymbol{x}) - \mathrm{E}(Y)\right]^2 p(\boldsymbol{x}) d\boldsymbol{x}_{\sim j} dx_j \tag{19}$$

5.2 *Interpretation of the CSV plot*

The CSV plot is a useful tool to analyse the effect of reduced range of an input parameter to the variance of the model output. The CSV plot works along similar lines with the CSM plot: if it is close to diagonal, it indicates that the contribution to the output variance is equal throughout the range of the input parameter X_j.

Following Eqn. (9a), let us have a look at the derivative, or the slope, of the CSV plot. From Eqn. (19), the slope of the CSV plot between two points $(q_1, CSV_j(q_1))$ and $(q_2, CSV_j(q_2))$ can be written as:

$$\frac{CSV_j(q_2) - CSV_j(q_1)}{q_2 - q_1} = \frac{\frac{1}{(q_2-q_1)} \int_{F_j^{-1}(q_1)}^{F_j^{-1}(q_2)} \int_{\mathbb{R}^{m-1}} \left[f(\boldsymbol{x}) - \mathrm{E}(Y)\right]^2 p(\boldsymbol{x}) d\boldsymbol{x}_{\sim j} dx_j}{\mathrm{V}(Y)} \tag{20}$$

The numerator of the right member of Eqn. (20) can be identified as the variance of the model output when the range of the parameter X_j is reduced to $[F_j^{-1}(q_1), F_j^{-1}(q_2)]$, but with respect to constant mean $\mathrm{E}(Y)$ over the full range of all parameters. Using conditional expectations notations, we have:

$$\forall (q_1, q_2) \in [0,1]^2, \quad \frac{CSV_j(q_2) - CSV_j(q_1)}{q_2 - q_1} = \frac{\mathrm{E}\left[(Y - \mathrm{E}(Y))^2 \mid F_j(X_j) \in [q_1, q_2]\right]}{\mathrm{V}(Y)} \tag{21}$$

Similarly, by derivating Eqn. (19) with respect to q, using the fact that $(F_j^{-1})'(q) = 1/p_j\left(F_j^{-1}(q)\right)$, we obtain an expression of the CSV plot derivative:

$$\forall q \in [0,1], \quad CSV_j'(q) = \frac{\mathrm{E}\left[(Y - \mathrm{E}(Y))^2 \mid F_j(X_j) = q\right]}{\mathrm{V}(Y)} \tag{22}$$

Again, we identify the numerator of the right member of Eqn. (22) as the variance of the model output Y when the input parameter X_j is fixed to $F_j^{-1}(q)$, but with respect to constant mean $\mathrm{E}(Y)$ over the full range of all parameters. Eqn. (21) and (22) provide a good interpretation of the meaning of the CSV plot. Indeed, the CSV plot conveys information about the effect of a reduced range of input parameter X_j on the variance of the model output Y. The slope of the CSV plot $CSV_j(\cdot)$ between quantiles q_1 and q_2 is equal to the ratio between the reduced variance of model output Y when X_j varies in the reduced range $[F_j^{-1}(q_1), F_j^{-1}(q_2)]$, and the total variance $\mathrm{V}(Y)$. The slope of the tangent to the CSV plot $CSV_j(\cdot)$ at quantile q is equal to the ratio between the reduced variance of model output Y when X_j is fixed to $F_j^{-1}(q)$, and the total variance $\mathrm{V}(Y)$.

5.3 *Relation with total-order sensitivity indices?*

The total-order variance-based sensitivity index associated to input parameter X_j is defined by [5]:

$$ST_j = \frac{\mathrm{E}_{X_{\sim j}}\left[\mathrm{Var}_{X_j}\left(Y \mid \boldsymbol{X}_{\sim j}\right)\right]}{\mathrm{V}(Y)}$$

ST_j measures the total contribution of input parameter X_j, and its interactions with other input parameters $\boldsymbol{X}_{\sim j}$, to the output variance. It is the expected residual part of

output variance if all model inputs but X_j were fixed. Total-order indices verify $\sum_j ST_j \geq 1$. They are useful for model simplification by identifying model inputs that have little influence on the model output variance.

ST_j can be written as:

$$ST_j = \frac{1}{\mathrm{V}(Y)} \iiint_{\mathbb{R}^{m-1}} \mathrm{E}\left[\left(Y - \mathrm{E}\left[Y \mid \boldsymbol{X}_{\sim j} \right] \right)^2 \mid \boldsymbol{X}_{\sim j} = \boldsymbol{x}_{\sim j} \right] p_{\sim j}(\boldsymbol{x}_{\sim j})\, d\boldsymbol{x}_{\sim j} \qquad (23)$$

Let now compare the expression of the CSV plot derivative $CSV'_j(\cdot)$ [Eqn. (22)] and that of total-order sensitivity index ST_j [Eqn. (23)]. The key point is that the expression of $CSV'_j(\cdot)$ involves the computation of a conditionnal expectation $\mathrm{E}(\cdot \mid X_j = x_j)$ where input parameter is fixed, while the expression of ST_j first requires the computation of a conditionnal expectation $\mathrm{E}(\cdot \mid \boldsymbol{X}_{\sim j} = \boldsymbol{x}_{\sim j})$ where all input parameters but X_j are fixed. The space of uncertain input parameters is thus not explored in the same order in these two expressions. This difference supports the conclusion that, in spite of the engaging analogy between the CSM plot and the CSV plot, further research is needed to find a possible link of CSV to sensitivity indices.

6. Numerical test cases

6.1 *Ishigami test case*

6.1.1 Presentation

We consider the usual Ishigami test case, with three uncertain inputs X_1, X_2 and X_3 treated as i.i.d random variables with uniform pdf in $[-\pi, \pi]$, and two fixed parameters $a = 7$, $b = 0.1$:

$$Y = \sin(X_1) + a \cdot \sin(X_2)^2 + b \cdot X_3^4 \cdot \sin(X_1) \qquad (24)$$

6.1.2 Scatterplots, CSM plots and CUSUNORO plots

We first draw a simple random sample $\mathbf{x}$ of size $n = 300$ and compute the associated output vector $\boldsymbol{y} = f(\mathbf{x})$. We then build, for each input X_1, X_2, X_3, the scatterplot of output vector $\boldsymbol{y}$ against input vector X_j (Fig. 2), the CSM plot $(q_i, c_{ij}^+)_{i=1,\ldots,n}$, and the CUSUNORO plot $(q_i, z_{ij})_{i=1,\ldots,n}$.

It first appears on Fig. 3 that the CSM curve for input X_1 deviates a lot from the diagonal, while CSM curves for X_2 and X_3 stay closer to it, which we interpret to mean that input parameter X_1 has the largest first-order contribution to output variance. The same observation can be derived from the CUSUNURO plots (Fig. 4), excepts that one then have to consider the departure of $Z_j(q)$ from the horizontal 0 line. This illustration of CSM/CUSUNORO plots on the Ishigami function also shows that the CSM plot is somehow easier to read than the CUSUNORO curve, mainly because it is scaled in such a way that it always lies in the $[0, 1] \times [0, 1]$ square. However, one may also argue that the relative importance of the inputs is better highlighted in the CUSUNORO curve.

To account for sampling variability, the above procedure was repeated $r = 30$ times, for increasing sample sizes $n = 10, 100, 1000$. Fig. 5 displays the resulting multiple CSM plots $(q_i, c_{i1}^+)_{i=1,\ldots,n}$ for first input parameter X_1 over the r replicas. It clearly suggests that at small sample size ($n = 10$), the CSM plot fails to capture the behaviour of the Ishigami function across the range of X_1 values.

6.1.3 Estimation of first-order sensitivity indices

We now compare, on the Ishigami function, the three estimation methods for first-order sensivity indices S_j based on the CSM plot: finite difference (Sections 4.2), polynomial regression (Section 4.3) and spline smoothing (Section 4.4). Again, we first draw a simple random sample $\mathbf{x}$ of size $n = 300$, compute the associated output vector $\boldsymbol{y} = f(\mathbf{x})$, and build the CSM plot $(q_i, c_{ij}^+)_{i=1,\dots,n}$ for each input parameter X_1, X_2,X_3.

Fig. 6 shows the approximation of CSM derivative $\frac{\mathrm{d}C_j}{\mathrm{d}q}(q), j = 1, 2, 3$ obtained with a 11-points finite difference scheme. The approximation appears to be extremely noisy and cannot be used to estimate first-order sensitivity indices.

To fit polynomial regressions (second method) on CSM sample points $(q_i, c_i)_{i=1,\dots,n}$, we used the `lm` function on `R` software along with the `orthopolynom` library. Model selection in the range $d = 2, \dots, 20$ was based on the AICc information criterion: $d = 8$ was selected as the maximum polynomial degree for all input parameters X_1 to X_3 (Fig. 7). Fig. 8 shows the fitted polynomial model on the (q_i, c_i) CSM points. The quality of the regression is good, as it can be seen by looking at the regression residuals ϵ_i (Fig. 9 and Fig. 10), even if a slight auto-correlation structure must be noted.

To perform the CSM spline smoothing (third method), we used the `spline` package on `R` software (spline order $o = 5$, smoothing parameter $\lambda = 1$). Fig. 11 shows the fitted spline model on the (q_i, c_i) CSM points. Fig. 12 then displays the estimates of CSM derivatives $C'(q)$ as derived from the spline regression model (these estimates were also computed using `spline` package). In line with Eqn. (11), we observe that the CSM derivative is equal to the – shifted and scaled – conditional mean $\mathrm{E}(Y|X_j)$.

To account for sample variability, the whole estimation procedure described above was repeated for increasing sample sizes from $n = 30$ to $n = 10000$. For each sample size n, we drew $r = 100$ random samples, computed 100 estimates of first-order sensitivity indices $\hat{S}_1, \hat{S}_2, \hat{S}_3$ with both polynomial regression and spline smoothing, then computed the mean value and standard deviation of sensitivity indices estimates over the 100 replicas (Fig. 13 and Table 1). These numerical results suggest that, on average, the polynomial regression procedure performs better than the spline and finite difference techniques.

6.1.4 Influence of input sampling scheme

In the previous numerical experiments, the input sample $\mathbf{x}$ has been drawn using Simple Random Sampling (SRS). However, it is well-known from the GSA literature than using more sophisticated input sample schemes can improve the estimation of sensitivity indices [5]. While it is clearly not the emphasis of our research — we rather focus on the estimation of S_j from given data — to discuss what is the best sampling scheme for GSA, we still found interesting to compare the estimation of S_j from the CSM plot using three different sampling scheme: i) the SRS scheme; ii) an optimised Latin Hypercube Sampling (optLHS), using the dedicated `lhs` package in `R`, and iii) a low-discrepancy sampling scheme (namely, scrambled Sobol' sequences), using the `randtoolbox` package in `R`. Fig. 14 displays the first-order indices estimates for input paramater X_1, obtained with polynomial regression of the CSM plot, for increasing sample size n over $r = 100$ replicas. It supports the conclusion that both the LHS and scrambled Sobol' sampling schemes lead to better estimates of S_j than SRS. It also indicates that, for the Ishigami function, the scrambled Sobol' sequences perform slightly better than the optimised LHS schemes.

6.2 *G-Sobol test case*

As a second test case, we consider the usual G-Sobol function, with $m = 8$ uncertain input variables X_1 to X_8 treated as i.i.d random variables with uniform pdf in $[0, 1]$, and a fixed parameter vector $a = (0, 1, 4.5, 9, 99, 99, 99, 99)$:

$$Y = \prod_{j=1}^{8} \frac{|4X_j - 2| + a_j}{1 + a_j} \tag{25}$$

We follow exactly the same steps as the ones described for the Ishigami function (Section 6.1). Fig. 15 shows scatterplots of output vector $\boldsymbol{y}$ against inputs X_j for an input sample $\mathbf{x}$ of size $n = 300$. Fig. 16 shows the CSM plots associated to each input factor: inputs X_1, X_2 and X_3 prove to bring a significant contribution to the variance of Y, while the other inputs X_4 to X_8 are almost non-influential. Table 2 and Fig. 17 give the approximate sensitivity indices $\hat{S}_1$ to $\hat{S}_8$ obtained with each method for increasing sample sizes from $n = 30$ to $n = 10000$, along with their exact values. Again, it appears that the polynomial regression technique leads to the best $\hat{S}_j$ estimates.

7. Discussion

7.1 *Relation between the CSM plot and first-order sensitivity indices*

The Contribution to the Sample Mean plot was first introduced by Bolado-Lavin *et al.* [14] as an empirical and graphical technique for sensitivity analysis. They presented how to build a CSM plot from a sample of n model runs, using an empirical plotting procedure that we described in Section 2.1. Our first contribution is to supplement this approach by providing a more formal definition for the CSM plot [Eqn. (3)], which, in return, allows us to take a fresh look at the original CSM plotting procedure: it appears that the empirical CSM points c_{ij} are simply Monte Carlo estimates of the exact CSM quantities $C_j(q_i)$ for the n-quantiles $q_i = i/n$, $i = 1, \ldots, n$. We also indicated how the CSM plot and the CUSUNORO curve, previously introduced by Plischke [12], are related.

Our main goal was then to clarify the alleged but unproven link between CSM plots and variance-based sensitivity indices. We provide with Eqn. (9) an exact relation between the CSM plot $C_j(\cdot)$ associated to input parameter X_j and the first-order sensitivity index S_j. This formula shows that the value of S_j, and thus the contribution pf X_j to the variance of the model output, depends on how much $C_j(q)$ deviates from the diagonal line in the $[0, 1] \times [0, 1]$ square. This investigation was strongly motivated by prior, non fully proven assertions of Bolado–Lavin *et al.* and Tarantola *et al.* [14, 15], that *"an input featuring a very low first-order effect will lead to a line close to the diagonal in the CSM plot"*, and that *"global importance measures could be derived from the CSM plot and provide the same ranking of model inputs as first-order sensitivity indices"*. Our findings corroborate these empirical claims. However, they also indicate that the initial CSM-based importance measure suggested by Bolado–Lavin *et al.* – that is, the maximum distance of the CSM plot from the diagonal, $\max(|C(q) - q|; q \in [0, 1])$ – can be easily and usefully replaced by an estimate of S_j that we derive from Eqn. (9).

7.2 *Estimation of S_j from the CSM plot*

Our research also sought to design a new estimation procedure for first-order variance-based sensitivity indices S_j, based on the CSM plot. Thanks to the explicit formula link-

ing S_j and $C_j(\cdot)$ [Eqn. (9)], we designed three new estimates $\widehat{S_j}$ using: i) finite difference sheme; ii) polynomial expansion of the CSM plot; and iii) spline smoothing of the CSM plot. Numerical results on the Ishigami and G-Sobol functions support the conclusion that the polynomial regression-based estimates $\widehat{S_j}$ are the most effective. Equation (17) provides an explicit formula for $\widehat{S_j}$ from the polynomial regression coefficients.

This estimation procedure can be used to compute first-order sensitivity indices S_j from *given data* [12], that is, from any set of model runs. Unlike most variance-based sensitivity analysis technique, it does not require any SA-specific design of experiment to sample uncertain model inputs. Even so, using an optimized sampling of model inputs to build the CSM plot will obviously improve the estimation of S_j; on the Ishigami test cases, we found that both Sobol' low-discrepancy sequences and LHS designs could improve the estimation of S_j.

7.3 *Limits and further research*

It should be noted that this new $\widehat{S_j}$ estimate displays a poor accuracy for small sample sizes n, which may restrict its practical use in real-world studies. In particular, for $n \leq 1000$, it does not compare well the EASI technique introduced by Plishcke to compute S_j from given data [13]. This discrepancy may be explained by the fact that our CSM-based estimation technique proceeds in two steps: i) first the CSM plot is built from the vector $\boldsymbol{y}$ of model output – with knowledge of Eqn. (3), this first step can be understood as a *numerical integration* of the usual scatterplot of $\boldsymbol{y}$ against $\boldsymbol{x}_j$; then ii) first-order sensitivity index S_j is estimated from the CSM plot using Eqn. (9a), which requires the computation of the CSM *derivative*. Our two-steps estimation technique thus involves both numerical integration and numerical differentiation. This may explain why it is less efficient than the more straightforward EASI technique, that directly computes S_j from the output vector $\boldsymbol{y}$. Further research is thus needed to try and improve the accuracy and convergence of our CSM-based $\widehat{S_j}$ estimate, and to explore its asymptotic properties.

Finally, we also briefly investigated in Section 5 if the Contribution to the Sample Variance plot [15], which is analogous to the CSM plot, could allow the estimation of total-order sensitivity indices ST_j in a similar way. Our first results suggest that it is probably not the case, because the CSV plot $CSV_j(\cdot)$ and total-order index ST_j do not involve the same variances. More lessons should be learnt in the future on this issue by analyzing CSV plots and values of total-order indices for a range of simple test functions.

8. Conclusion

This work was performed with a view towards promoting the use of the CSM plot as a convenient graphical tool for sensitivity analysis from given data. We first provided an exact formula linking, for each uncertain model input X_j, the CSM plot $C_j(\cdot)$ with the first-order variance-based sensitivity index S_j. We then designed a new first-order sensitivity index estimate $\widehat{S_j}$, based on the polynomial regression of the CSM plot. Because this estimate requires the computation of the CSM plot derivative, its performance does not compare well with other estimation methods for first-order sensitivity indices from given data, such as the EASI method. Further research is thus needed to improve the accuracy of $\widehat{S_j}$, and to investigate its asymptotic properties. Another challenging issue would be to design a new graphical tool, analogous to the CSM plot, that would allow the computation of total-order sensitivity indices: our first findings indicate that the CSV plot may be a misleading track, and that other solutions should now be sought after.

Acknowledgements

The first author thanks the MASCOT–NUM and the PEER networks for funding the international mobility grant that supported this research, as well as the Joint Research Center of the European Commission in Ispra, Italy, for their warm welcome.

Appendix A. Proofs on Legendre polynomials

PROPOSITION 2 *Let consider the shifted Legendre polynomials* $(P_k(q))_{k\in\mathbb{N}}$ *defined by* $P_k(q) = P_k^{(s)}(2q-1)$*, with* $P_k^{(s)}$ *the standardized Legendre polynomials. We have:*

$$\forall (k,l) \in \mathbb{N}^2,\, k \leq l, \qquad \int_0^1 P_k'(q)P_l'(q)dq = 2k(k+1)1_{\{(k+l)\,\in\, 2\mathbb{N}\}}$$

Proof. Consider $(k,l) \in \mathbb{N}^2$ such that $k \leq l$. Let denote by $I_{k,l}$ the integral $I_{k,l} = \int_0^1 P_k'(q)P_l'(q)dq$. Using an integration by parts we have :

$$I_{k,l} = \left[P_k'(q)P_l(q)\right]_0^1 - \int_0^1 P_k''(q)P_l(q)dq$$

P_k'' is a polynom of degree $k-2$: it can be decomposed on the finite orthogonal basis $(P_i)_{i=1,\ldots,k-2}$. As $k-2 < l$, using the orthogonality of shifted Legendre polynomials $(P_k)_{k\in\mathbb{N}}$ on $[0,1]$, we find that the integral $\int_0^1 P_k''(q)P_l(q)dq$ is equal to 0. Hence :

$$I_{k,l} = P_k'(1)P_l(1) - P_k'(0)P_l(0)$$

The values of $P_k(q)$ and its derivative $P_k'(q)$ at $q=0$ and $q=1$ can be found from the corresponding values of non-shifted Legendre polynomial $P_k^{(s)}(q)$ at $q=-1$ and $q=1$, which are given in [16, p.777], Eqn. (22.4.6), (22.5.37) and (22.4.2). Using the relations $P_k(q) = P_k^{(s)}(2q-1)$ and $P_k'(q) = 2(P_k^{(s)})'(2q-1)$ we have:

$$\forall k \in \mathbb{N} \quad \begin{cases} P_k(1) = 1 \\ P_k'(1) = k(k+1) \\ P_k(0) = (-1)^k \\ P_k'(0) = (-1)^{k-1}k(k+1) \end{cases}$$

We finally obtain:

$$I_{k,l} = k(k+1)[1 + (-1)^{k+l}]$$

which we can also write this way:

$$I_{kl} = 2k(k+1)1_{\{(k+l)\,\in\, 2\mathbb{N}\}}$$

■

References

[1] Sobol' I. Sensitivity analysis for non-linear mathematical model. Matem Mod. 1993;1:407–414.

[2] European Commission. Impact assessment guidelines; 2009. Tech Rep SEC(2009) 92.

[3] CREM. Guidance on the development, evaluation, and application of environmental models. US Environmental Protection Agency, Council for Regulatory Environmental Modeling; 2009. Tech Rep. Available from: http://www.epa.gov/crem/library/cred_guidance_0309.pdf.

[4] Hoeffding W. A class of statistics with asymptotically normal distribution. Ann Math Stat. 1948;19:293–325.

[5] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S. Global Sensitivity Analysis - The Primer. Wiley; 2008.

[6] Sobol' I. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simulat. 2001;55:271–280.

[7] Xu C, Gertner G. Understanding and comparisons of different sampling approaches for the Fourier Amplitudes Sensitivity Test (FAST). Comput Stat Data Anal. 2011;55(1):184–198.

[8] Jansen M. Analysis of variance designs for model output. Comput Phys Commun. 1999; 117(1):35–43.

[9] Chan K, Saltelli A, Tarantola S. Winding stairs: a sampling tool to compute sensitivity indices. Stat Comput. 2000;10:187–196.

[10] Plischke E, Borgonovo E, Smith C. Global sensitivity measures from given data. Eur J Oper Res. 2013;226(3):536–550.

[11] Sinclair J. Response to the PSACOIN Level S exercise. PSACOIN Level S intercomparison. Nuclear Energy Agency, Organisation for Economic Co-operation and Development; 1993. Tech Rep.

[12] Plischke E. An adaptive correlation ratio method using the cumulative sum of the reordered output. Reliab Eng Syst Saf. 2012;107:149–156.

[13] Plischke E. An effective algorithm for computing global sensitivity indices (EASI). Reliab Eng Syst Saf. 2010;95(4):354–360.

[14] Bolado-Lavin R, Castaings W, Tarantola S. Contribution to the sample mean plot for graphical and numerical sensitivity analysis. Reliab Eng Syst Saf. 2009;94:1041–1049.

[15] Tarantola S, Kopustinskas V, Bolado-Lavin R, Kaliatka A, Ušpuras E, Vaišnoras M. Sensitivity analysis using contribution to sample variance plot: Application to a water hammer model. Reliab Eng Syst Saf. 2012;99(0):62–73.

[16] Abramowitz M, Stegun IA, editors. Handbook of mathematical functions with formulas, graphs, and mathematical tables. New York: Dover Publications; 1972.

Table 1. Ishigami toy function: exact and approximate values of first-order sensitivity indices. Sample size $n = 300$, mean and $\pm$ standard deviation over $r = 100$ replicas.

Method	$\hat{S}_1$	$\hat{S}_2$	$\hat{S}_3$
Exact values	0.3190	0.4511	0.004
Finite difference	0,950 ± 0,321	0,893 ± 0,418	0,836 ± 0,625
Polynomial regression	0,293 ± 0,038	0,536 ± 0,058	0,038 ± 0,028
Spline smoothing	0,285 ± 0,037	0,683 ± 0,069	0,032 ± 0,015

Table 2. G-Sobol test case: exact and approximate values of first-order sensitivity indices. Sample size $n = 300$, mean and $\pm$ standard deviation over $r = 100$ replicas.

Index	Exact values	Poly. regression	Spline smoothing
$\hat{S}_1$	0.716	0.734 ± 0.047	0.669± 0.037
$\hat{S}_2$	0.179	0.208 ± 0.049	0.207± 0.040
$\hat{S}_3$	0.024	0.058 ± 0.034	0.060± 0.024
$\hat{S}_4$	0.007	0.013 ± 0.023	0.011± 0.014
$\hat{S}_5$	7.10^{-5}	0.023 ± 0.021	0.024± 0.012
$\hat{S}_6$	7.10^{-5}	0.034 ± 0.024	0.029± 0.013
$\hat{S}_7$	7.10^{-5}	0.033 ± 0.018	0.023± 0.011
$\hat{S}_8$	7.10^{-5}	0.079 ± 0.021	0.038± 0.012

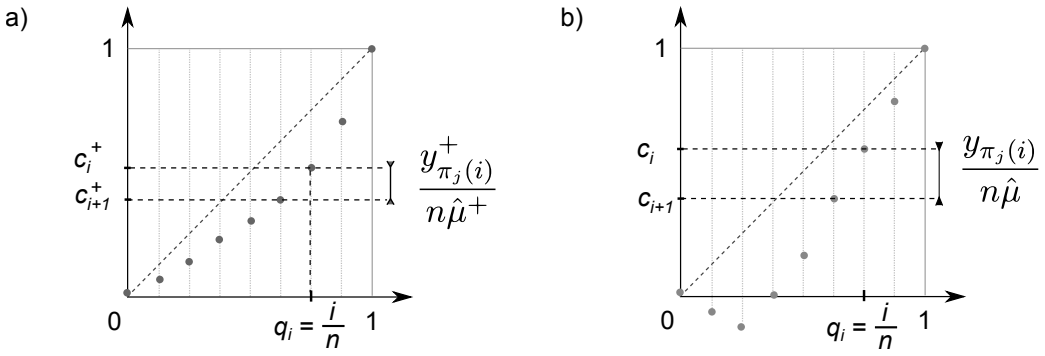

Figure 1. Empirical CSM points (q_i, c_{ij}^+) (left) and (q_i, c_{ij}) (right), from a sample of $n = 8$ model evaluations.

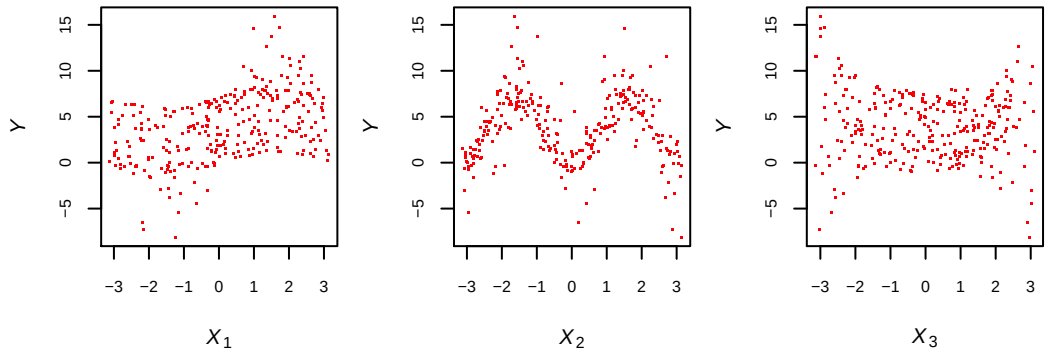

Figure 2. Ishigami test case: output Y against inputs X_1, X_2 and X_3. Sample size $n = 300$.

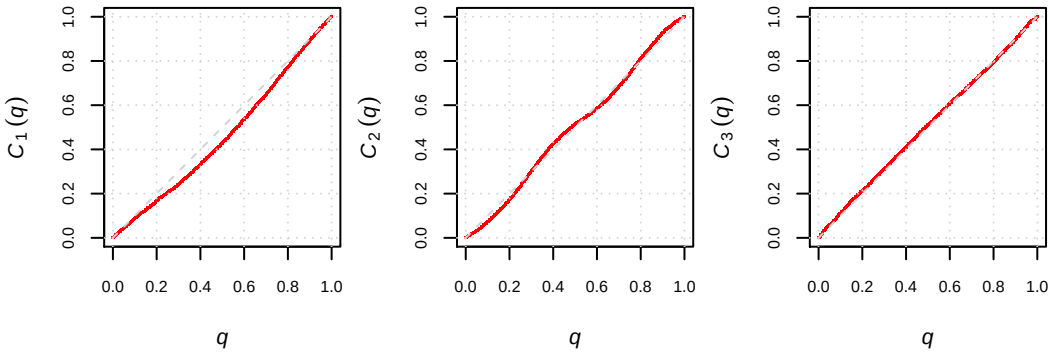

Figure 3. Ishigami test case: CSM plot of X_1, X_2 and X_3 with respect to Y; sample size $n = 300$.

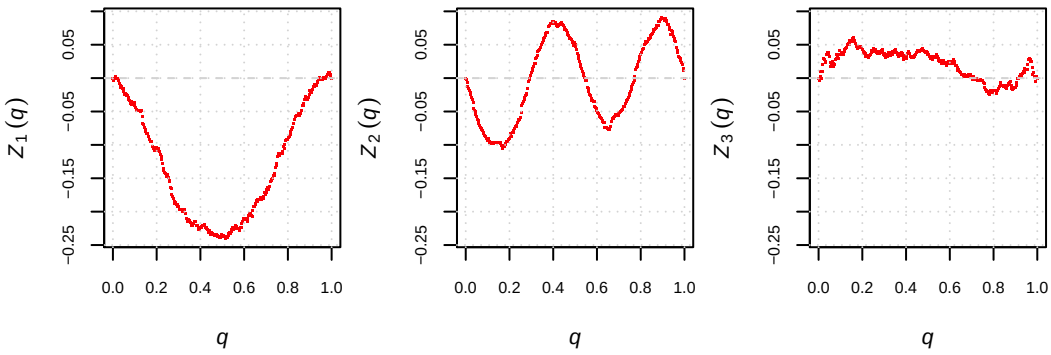

Figure 4. Ishigami test case: CUSUNORO plot of X_1, X_2 and X_3 with respect to Y; sample size $n = 300$.

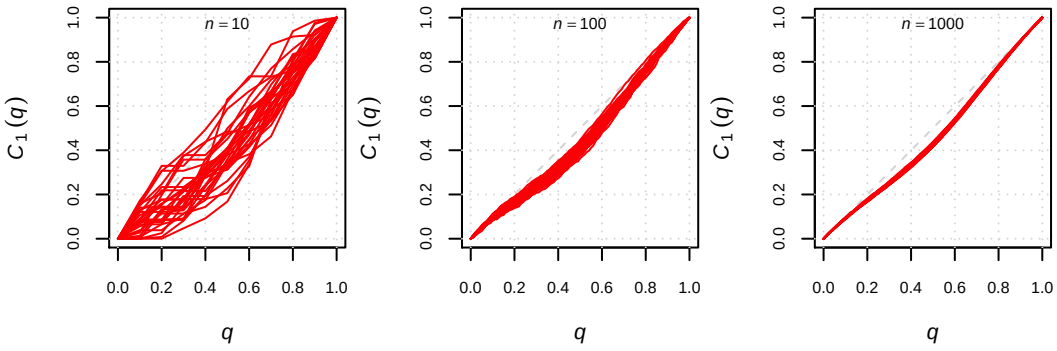

Figure 5. Ishigami test case: CSM sampling distribution over $r = 30$ replicas, for input parameter X_1 and for increasing sample sizes $n = 10, 100, 1000$.

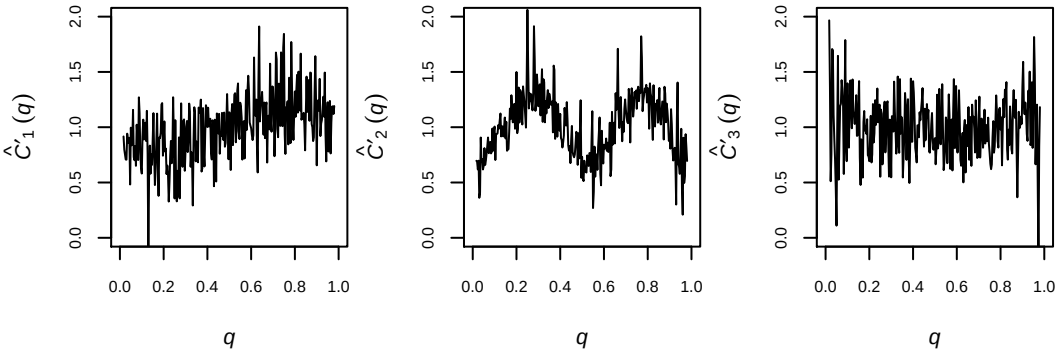

Figure 6. Ishigami test case: estimates c_i' of CSM derivative computed with finite difference scheme.

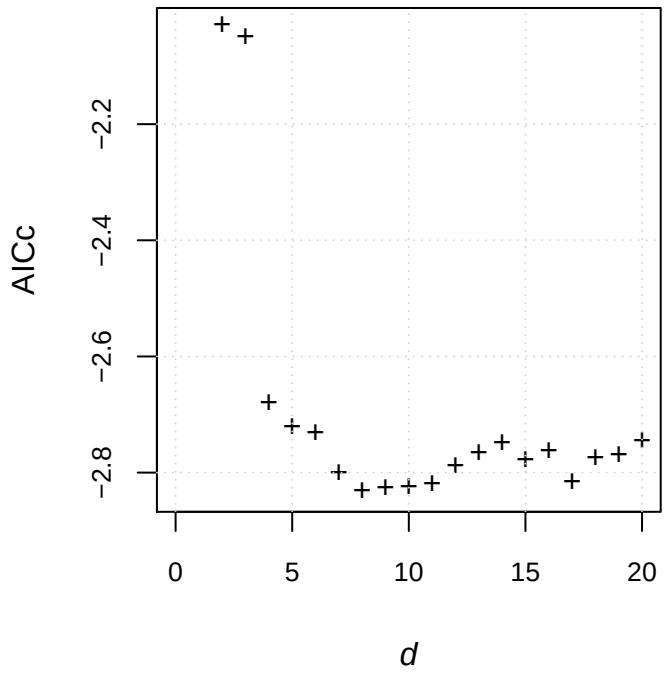

Figure 7. Ishigami test case: AICc information criteria for increasing maximum degree d of the polynomial regression. Best model is obtained for $d = 8$.

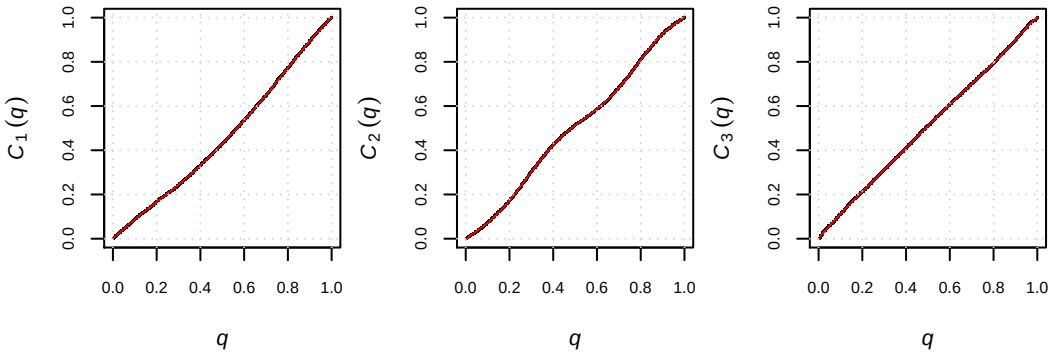

Figure 8. Ishigami test case: polynomial fit of degree $d = 8$ on the CSM points $(q_i, c_i)_{i=1,\ldots,n}$ for input factors X_1 (left) to X_3 (right).

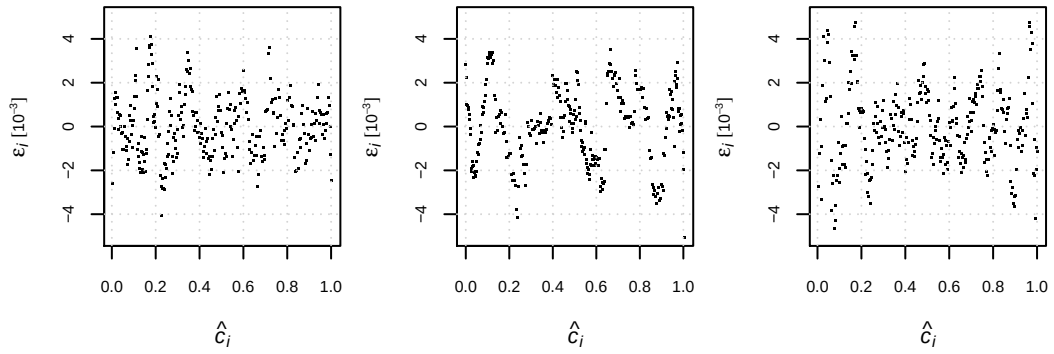

Figure 9. Ishigami test case: polynomial regression of the CSM curve. Residuals ϵ_i against fitted values $\hat{c}_i$ for input factors X_1 (left) to X_3 (right).

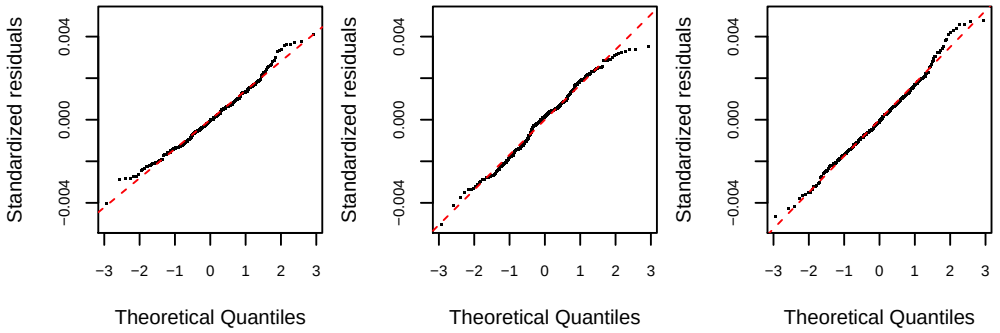

Figure 10. Ishigami test case: polynomial regression of the CSM curve. QQnorm plot of residuals $\epsilon_i = \hat{c}_i - c_i$ for input factors X_1 (left) to X_3 (right).

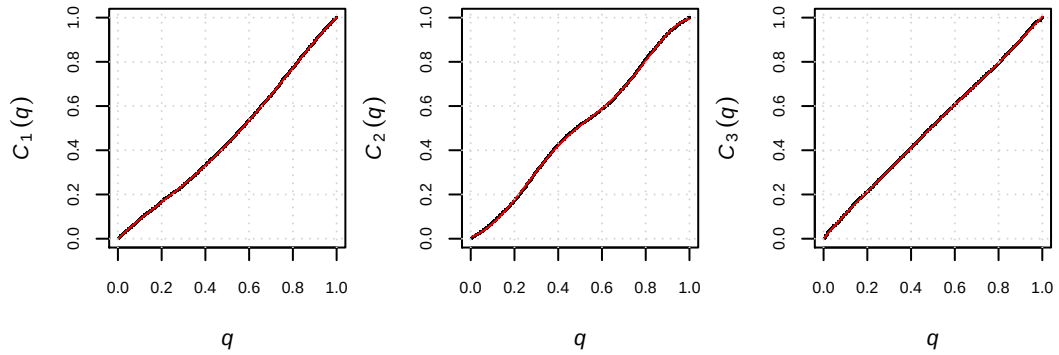

Figure 11. Ishigami test case: spline smoothing on the CSM points $(q_i, c_i)_{i=1,\ldots,n}$ for input factors X_1 (left) to X_3 (right).

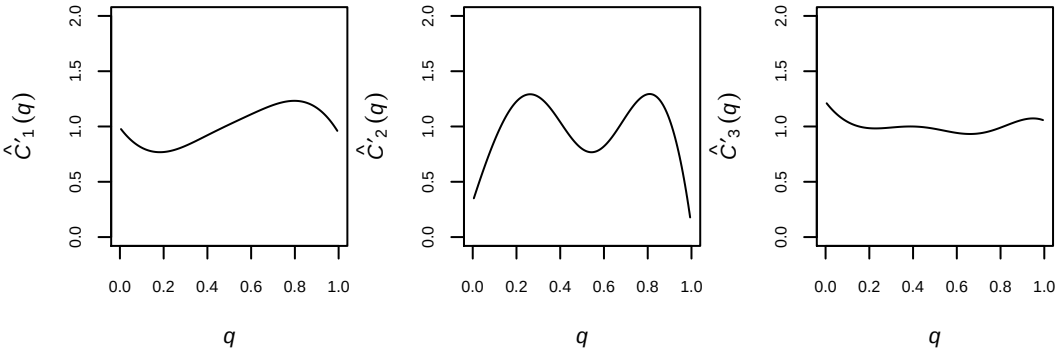

Figure 12. Ishigami test case: estimation of CSM derivative $C'(q)$ with spline smoothing (solid line) and exact conditional expectation $\mathrm{E}\left(Y|X_j = F_j^{-1}(q)\right)$ (dashed line).

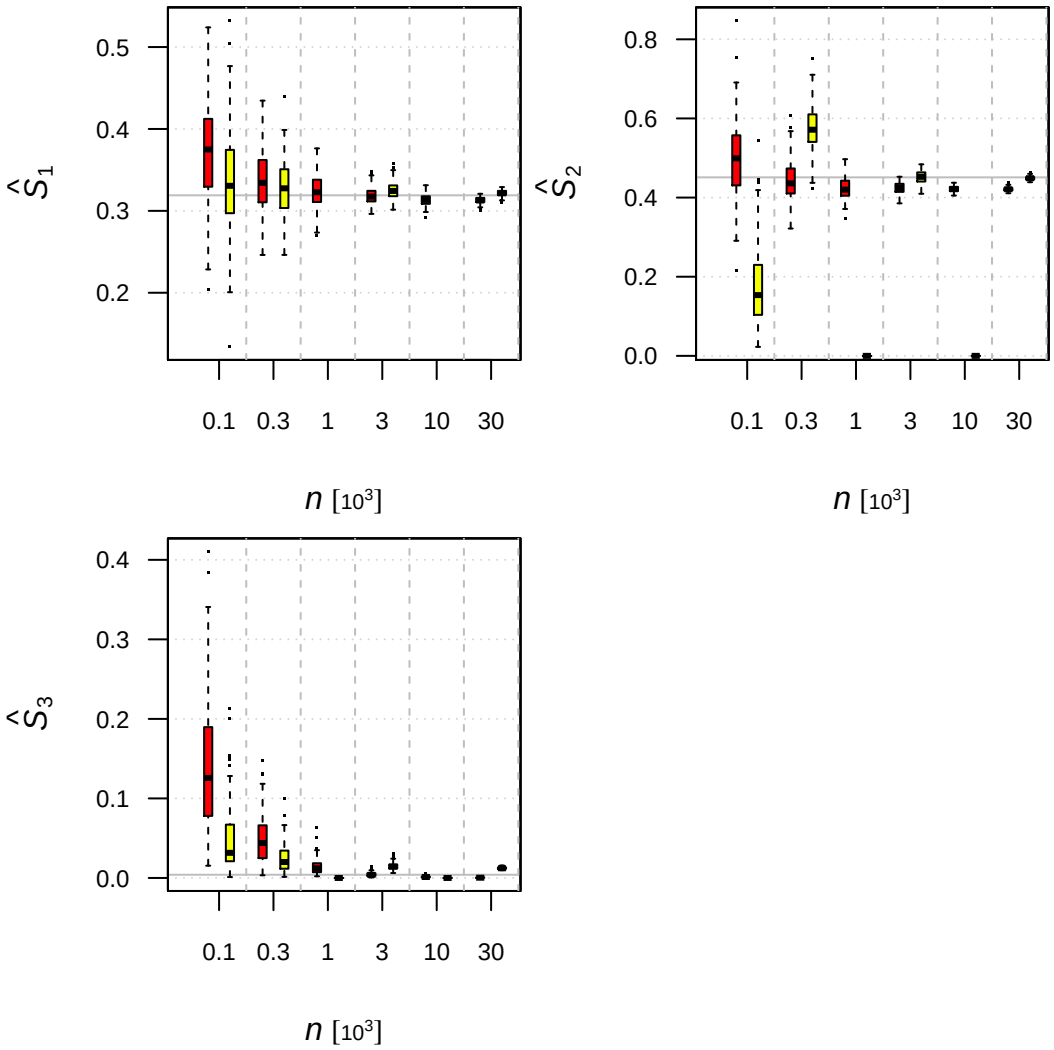

Figure 13. Ishigami test case: estimates of first-order sensitivity indices $\hat{S}_1$ (left) to $\hat{S}_3$ (right) for increasing sample size n with polynomial regression (left boxplots) or spline smoothing (right boxplots). Horizontal solid line shows the exact value of S_j.

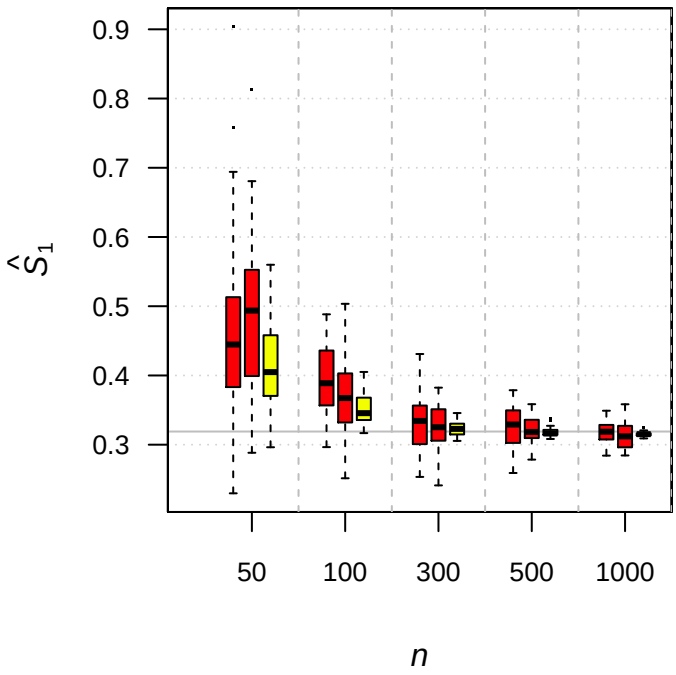

Figure 14. Ishigami test case: convergence of sensitivity index estimate $\hat{S}_1$ for increasing sample sizes and different sampling schemes: Simple Random Sampling (left boxplots), LHS (middle boxplots) and Sobol' sequences (right bloxplots).

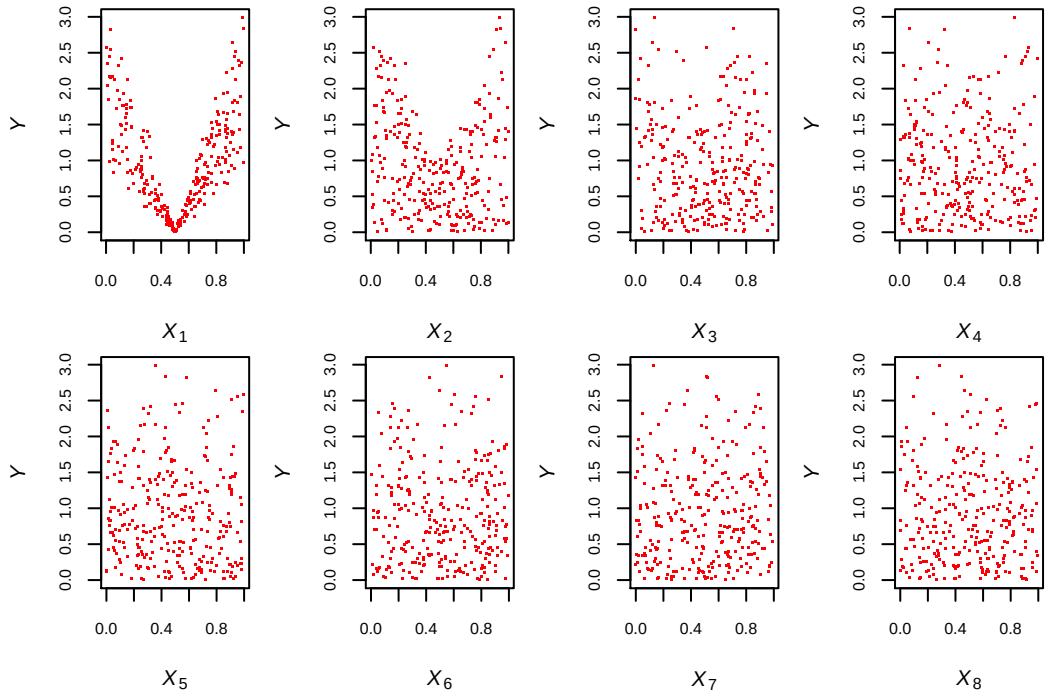

Figure 15. G-Sobol test case: output Y against inputs X_1 to X_8; sample size $n = 300$.

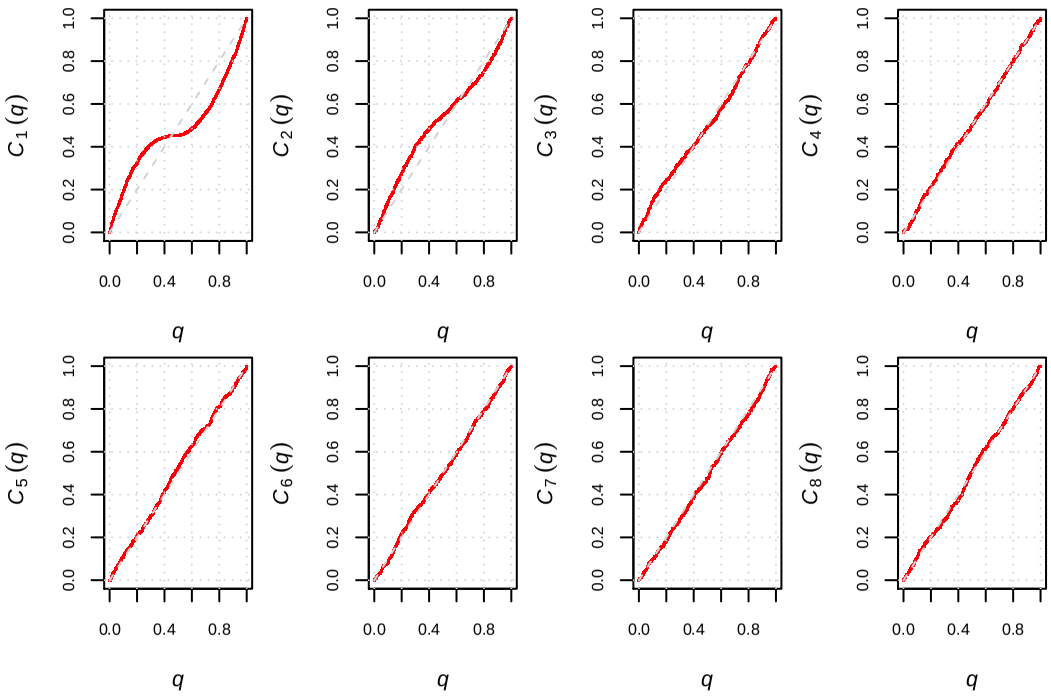

Figure 16. G-Sobol test case: CSM of X_1 to X_8 with respect to Y; sample size $n = 300$.

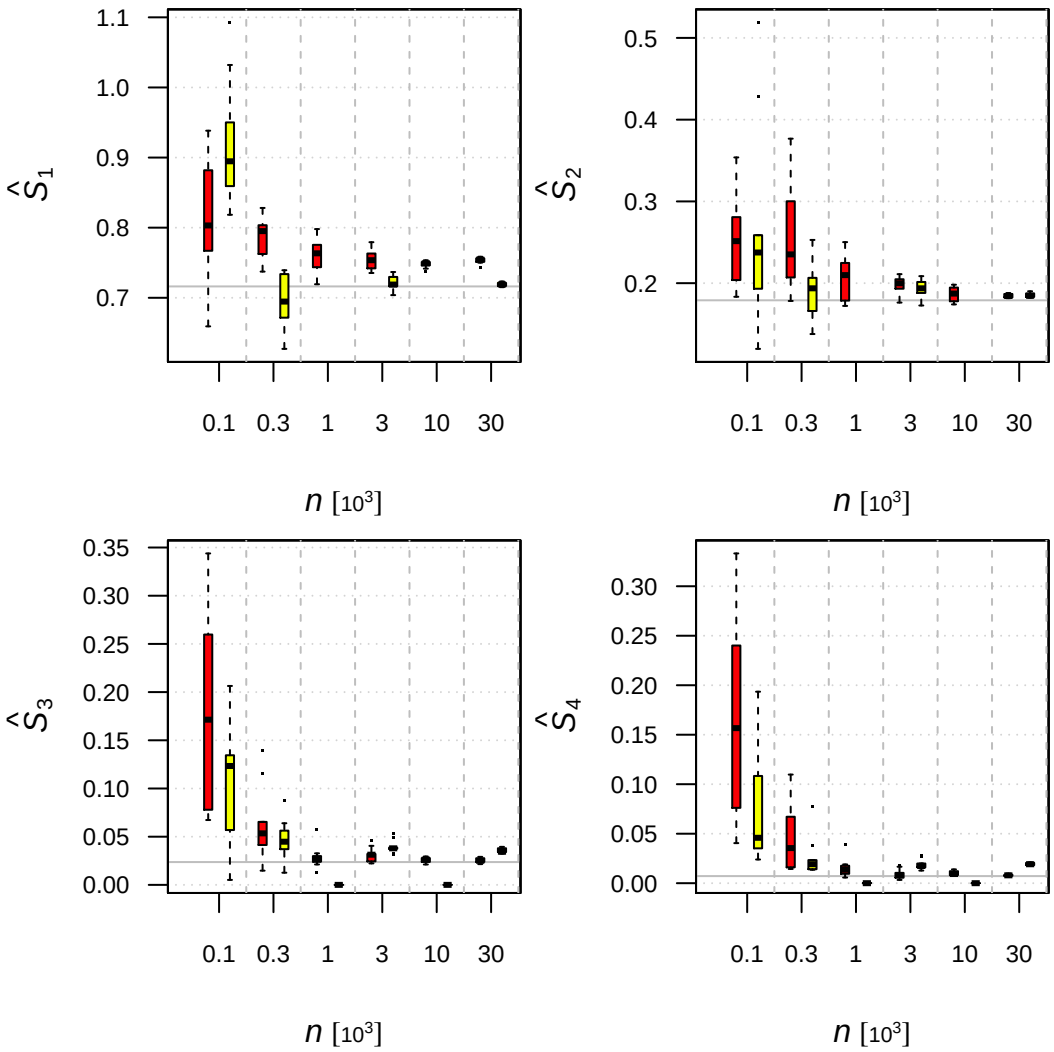

Figure 17. G-Sobol test case: estimates of first-order sensitivity indices $\hat{S}_1$ (top left) to $\hat{S}_4$ (bottom right) for increasing sample size n with polynomial regression (left boxplots) or spline smoothing (right boxplots) – mean and standard devation over 100 replicas. Horizontal solid line shows the exact value of S_j.