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

Agnès Lagnoux
Institut de Mathématiques de Toulouse
TOULOUSE - FRANCE

SAMO, March 14-16 2022

Outline of the talk

Introduction to SA and Sobol' indices

The classical Pick-Freeze estimation

Mighty estimation based on ranks

Comparison of the different estimation procedures

Numerical applications

Framework

We consider a complicated regression function f defined on $E = E_1 \times E_2 \times \cdots \times E_p$ and valued in $\mathbb{R}^k$ depending on several variables :

$$y = f(x_1, \ldots, x_p), \tag{1}$$

where

1. the inputs x_i pour $i = 1, \ldots p$ are objects ;
2. f is deterministic and unknown. It is called a black-box.

Aim

Generally,

1. f is not analytically known ;
2. given $(x_1, \ldots, x_p)$, the computer code gives $y = f(x_1, \ldots, x_p)$;
3. computing $y = f(x_1, \ldots, x_p)$ may be costly.

Wishes :

1. evaluate y for any value of the p-uplet $(x_1, \ldots, x_p)$.
2. identify the most important variables to be able to fix the less important ones to their nominal value.

Probabilistic frame

In order to quantify the influence of a variable, it is common to assume that the inputs are random :

$$X := (X_1, \ldots, X_p) \in E = E_1 \times \ldots \times E_p.$$

Then $f : E \to \mathbb{R}^k$ is a measurable function that can be evaluated on runs and the output code Y becomes random too :

$$Y = f(X_1, \ldots, X_p).$$

In this presentation, the inputs X_i are assumed to be mutually independent.

Probabilistic frame

Main assumptions :

1. $X_1, \ldots, X_p$ are independent.
2. $\mathbb{E}[\|Y\|^2] < \infty$.
3. Y is scalar (here, for sake of simplicity).

The question is :

> *How one may quantify the amount of randomness that a variable or a group of variables bring to Y ?*

The simplest indicator of variability of a random variable is the variance.

The so-called Sobol' indices

Classically to quantify the amount of randomness that a variable or a group of variables bring to Y, one computes the so-called Sobol' indices.

For instance, the first order Sobol' index with respect to $X_{\mathbf{u}} = (X_i, i \in \mathbf{u})$ is given by

$$S^{\mathbf{u}} = \frac{\mathrm{Var}(\mathbb{E}[Y|X_{\mathbf{u}}])}{\mathrm{Var}(Y)}$$

(assuming Y is scalar).

Such indices stem from the Hoeffding decomposition of the variance of f (or equivalently Y) that is assumed to lie in L^2.

Some extensions of the Sobol' indices

- Multidimensional and functional outputs
  F. Gamboa, A. Janon, T. Klein, and A. Lagnoux. "Sensitivity analysis for multidimensional and functional outputs". *Electron. J. Stat*, (2014). Volume 8, no. 1, pp 575–603.
- Indices in general metric spaces - GMS indices
  F. Gamboa, T. Klein, A. Lagnoux, and L. Moreno. "Sensitivity analysis in general metric spaces ", *RESS*, 2021.
- Indices based on the whole distribution - Cramér-von Mises indices
  F. Gamboa, T. Klein, and A. Lagnoux. "Sensitivity analysis based on Cramér-von Mises distance ", *SIAM UQ*, 2018.
- Universal indices
  J.-C. Fort, T. Klein, and A. Lagnoux. "Global sensitivity analysis and Wasserstein spaces", *SIAM UQ*, 2021.

Estimation of the Sobol' indices

1. First approach - the classical Pick-Freeze estimation
2. Second approach - mighty estimation based on ranks

Outline of the talk

Introduction to SA and Sobol' indices

The classical Pick-Freeze estimation

Mighty estimation based on ranks

Comparison of the different estimation procedures

Numerical applications

Pick-Freeze estimation of Sobol' indices

To fix ideas assume for example $p = 5$, $\mathbf{u} = \{1, 2\}$ so that $\sim \mathbf{u} = \{3, 4, 5\}$.
We consider the Pick-Freeze variable $Y_{\mathbf{u}}$ defined as follows :

- draw $X = (X_1, X_2, X_3, X_4, X_5)$,
- build $X_{\mathbf{u}} = (X_1, X_2, X_3', X_4', X_5')$.

Then, we compute

- $Y = f(X)$,
- $Y_{\mathbf{u}} = f(X_{\mathbf{u}})$.

A small miracle

$$\mathrm{Var}(\mathbb{E}[Y|X_{\mathbf{u}}]) = \mathrm{Cov}(Y, Y_{\mathbf{u}}) \text{ so that } S^{\mathbf{u}} = \frac{\mathrm{Cov}\,(Y, Y_{\mathbf{u}})}{\mathrm{Var}(Y)}.$$

Pick-Freeze estimation of Sobol' indices

In practice Generate two N-samples :

- one N-sample of X : $\left(X^i\right)_{i=1,\dots,N}$,
- one N-sample of $X_{\mathbf{u}}$: $\left(X^i_{\mathbf{u}}\right)_{i=1,\dots,N}$.

Compute the code on both samples :

- $Y^i = f(X^i)_{i=1,\dots,N}$,
- $Y^i_{\mathbf{u}} = f(X^i_{\mathbf{u}})_{i=1,\dots,N}$.

Then estimate $S^{\mathbf{u}}$ by

$$S^{\mathbf{u}}_{N,PF} = \frac{\frac{1}{N}\sum Y^i Y^i_{\mathbf{u}} - \left(\frac{1}{N}\sum Y^i\right)\left(\frac{1}{N}\sum Y^i_{\mathbf{u}}\right)}{\frac{1}{N}\sum (Y^i)^2 - \left(\frac{1}{N}\sum Y^i\right)^2}$$

Pick-Freeze estimation : some statistical questions

Is the Pick-Freeze estimator a "good" estimator of the Sobol' index ?

- Is it consistent ? Response : YES SLLN.
- If yes, at which rate of convergence ? Resp. : YES CLT (cv in $\sqrt{N}$).
- Is it asymptotically efficient ? Resp. : YES.
- Is it possible to measure its performance for a fixed N ? Response : YES Berry-Esseen and concentration inequalities.

Pick-Freeze estimation : consistency and CLT

$$S^{\mathbf{u}}_{N,PF} = \frac{\frac{1}{N}\sum Y^i Y^i_{\mathbf{u}} - \left(\frac{1}{N}\sum Y^i\right)\left(\frac{1}{N}\sum Y^i_{\mathbf{u}}\right)}{\frac{1}{N}\sum (Y^i)^2 - \left(\frac{1}{N}\sum Y^i\right)^2}, \; S^{\mathbf{u}} = \frac{\mathrm{Var}\left(\mathbb{E}\left[Y|X_{\mathbf{u}}\right]\right)}{\mathrm{Var}(Y)}.$$

Theorem (Janon, Klein, Lagnoux, Nodet, Prieur (2015))

1. *One has* $S^{\mathbf{u}}_{N,PF} \xrightarrow[N\to\infty]{a.s.} S^{\mathbf{u}}$.
2. *If* $\mathbb{E}[Y^4] < \infty$, *then*

$$\sqrt{N}\left(S^{\mathbf{u}}_{N,PF} - S^{\mathbf{u}}\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1\left(0, \sigma_S^2\right)$$

where $\sigma_S^2 = \dfrac{\mathrm{Var}((Y-\mathbb{E}[Y])[(Y^u-\mathbb{E}[Y])-S^{\mathbf{u}}(Y-\mathbb{E}[Y])])}{\left(\mathrm{Var}(Y)\right)^2}$.

Pick-Freeze estimation : concentration inequality

The Central Limit Theorem is a limit result. In real life, the number of experiments is finite. Concentration inequalities allow to quantify the error between the estimate and the index true value for a fixed value of N.
Using soundly Bennett inequality, one gets

Proposition (Gamboa, Janon, Klein, Lagnoux, Prieur (2015))

Let $\mathbf{u}$ *be a subset of* $\{1,\ldots,p\}$. *Then,*

$$\mathbb{P}\left(|S_N^{\mathbf{u}} - S^{\mathbf{u}}| \geqslant t\right) \leqslant 2\exp\left(-\frac{N\mathrm{Var}(Y)^2}{128}\left(1-\frac{1}{N}\right)^2\left(\frac{t}{3+2t}\right)^2\right).$$

Pick-Freeze estimation of Sobol indices

Références

- A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. " Asymptotic normality et efficiency of a Sobol index estimator", *ESAIM P&S*, 2013.
- F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. " Statistical Inference for Sobol pick freeze Monte Carlo method", *Statistics*, 2015.

Drawbacks of the Pick-Freeze estimation

- The cost (=number of evaluations of the function f) of the estimation of the p first-order Sobol' indices is quite expensive : $(p+1)N$.
- This methodology is based on a particular design of experiment that may not be available in practice. For instance, when the practitioner only has access to real data.

$\Rightarrow$ *We are then interested in an estimator based on a N-sample only.*

Outline of the talk

Introduction to SA and Sobol' indices

The classical Pick-Freeze estimation

Mighty estimation based on ranks

Comparison of the different estimation procedures

Numerical applications

Mighty estimation based on ranks

Here we assume that the inputs X_i for $i = 1, \dots, p$ are scalar and we want to estimate the Sobol' index S^1 with respect to X_1 :

$$S^1 = \frac{\mathrm{Var}\left(\mathbb{E}[Y|X_1]\right)}{\mathrm{Var}(Y)}.$$

To do so, we consider a N-sample of the input/output pair (X_1, Y) given by

$$(X_{1,1}, Y_1), (X_{1,2}, Y_2), \dots, (X_{1,N}, Y_N).$$

Mighty estimation based on ranks

The pairs $(X_{1,(1)}, Y_{(1)}), (X_{1,(2)}, Y_{(2)}), \dots, (X_{1,(N)}, Y_{(N)})$ are rearranged in such a way that

$$X_{1,(1)} < \dots < X_{1,(N)}.$$

Example

- $N = 6$
- Original sample $(1,5), (2,9), (-2,3), (6,-4), (0,8)$
- Rearranged sample $(-2,3), (0,8), (1,5), (2,9), (6,-4)$.

Mighty estimation based on ranks

We introduce

$$S^1_{N,Rank} = \frac{\frac{1}{N}\sum_{i=1}^{N-1} Y_{(i)} Y_{(i+1)} - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right)^2}{\frac{1}{N}\sum_{i=1}^{N} Y_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} Y_i\right)^2}.$$

Theorem (Gamboa, Gremaud, Klein, Lagnoux, 2021)

1. *One has* $S^1_{N,Rank} \xrightarrow[N\to\infty]{a.s.} S^1$.
2. *If the* X_i*'s are uniformly distributed and under some mild assumptions on* f*, then*

$$\sqrt{N}\left(S^1_{N,Rank} - S^1\right) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1\left(0, \sigma_R^2\right).$$

Sketch of the proof of the CLT

We consider the estimation of $\mathbb{E}[\mathbb{E}[Y|X_1]^2]$ only and its estimation

$$\frac{1}{N}\sum_{j=1}^{N} Y_{(j)}Y_{(j+1)}.$$

For $j=1,\dots N-1$, we note $X = X_1$, $W = (X_2,\dots,X_p)$ and introduce

$$\boxed{\Delta_{N,j} := f\left(X_{(j)}, W_j\right) - f\left(\frac{j}{N+1}, W_j\right),\ W_{N,j} := \left(\frac{j}{N+1}, W_j\right).}$$

Then, by a Taylor expansion (allowed by the regularity of f),

$$Y_{(j)}Y_{(j+1)} \approx f\left(W_{N,j}\right) f\left(W_{N,j+1}\right) + \Delta_{N,j} f\left(W_{N,j+1}\right) + \Delta_{N,j+1} f\left(W_{N,j}\right).$$

Sketch of the proof of the CLT

First part :

$$B_N := \frac{1}{N}\sum_{j=1}^{N-1} f\left(W_{N,j}\right) f\left(W_{N,j+1}\right).$$

We use the CLT for 1-dependent random variables of Orey *et al.* (1958) together with

Lemma (Key lemma 1)

There exists a measurable set $\Pi \subset \Omega_W$ *with* $\mathbb{P}_W$*-probability one such that for any* $\omega_W \in \Pi$,

$$\frac{1}{N}\sum_{j=1}^{N-2}\delta_{\left(\frac{j-1}{N+1},\frac{j}{N+1},\frac{j+1}{N+1},\frac{j+2}{N+1},W_{j-1}(\omega_W),W_j(\omega_W),W_{j+1}(\omega_W)\right)}$$
$$\Rightarrow \mathcal{L}_{(X,X,X)} \otimes \mathcal{L}_W \otimes \mathcal{L}_W \otimes \mathcal{L}_W,$$

as $N \to \infty$ *where as before* X *is uniformly distributed on* $[0,1]$.

Sketch of the proof of the CLT

Second part :

$$C_N := \frac{1}{N}\sum_{j=1}^{N-1}\left(\Delta_{N,j} f\left(W_{N,j+1}\right) + \Delta_{N,j+1} f\left(W_{N,j}\right)\right)$$

$$\approx \frac{1}{N}\sum_{j=1}^{N-1}\left(X_{(j)} - \frac{j}{N+1}\right) f_x\left(W_{N,j}\right)\left(f\left(W_{N,j-1}\right) + f\left(W_{N,j+1}\right)\right)$$

by a Taylor expansion.

We work conditionally to $\mathcal{F}_W$ the σ-algebra generated by the W_j's and we recall that

$$W_{N,j} := \left(\frac{j}{N+1}, W_j\right).$$

Sketch of the proof of the CLT

The next lemma is a generalization of the CLT for a L-statistics.

Lemma (Key lemma 2)

Let $(U, \mathbb{B}(U))$ be a Polish space where $\mathbb{B}(U)$. Let $(\chi_j)_{1 \leqslant j \leqslant n,\ n \in \mathbb{N}^*}$ valued in U and Q a proba. measure on $U \times [0,1]$ such that

$$\frac{1}{N} \sum_{j=1}^{N-1} \delta_{\frac{j}{N}, \chi_j} \Rightarrow Q.$$

Let ψ be a bounded measurable real function on $U \times [0,1]$. We assume that the set of discontinuity points of ψ has null Q-probability. Then,

$$\frac{1}{\sqrt{N}} \sum_{j=1}^{N-1} \left(X_{(j)} - \frac{j}{N+1} \right) \psi \left(\chi_j, \frac{j}{N} \right) \xrightarrow[N \to \infty]{\mathcal{L}} \mathcal{N}\left(0, s_\psi^2\right).$$

Sketch of the proof of the CLT

We use the representation

$$X_{(j)} \overset{\mathcal{L}}{=} \frac{\sum_{i=1}^{j} E_i}{\sum_{i=1}^{N+1} E_i}, \text{ where } E_i \sim \mathcal{E}(1) \text{ and are independent.}$$

We apply Key lemma 1 with $\chi_j = (W_{j-1}, W_j, W_{j+1})$ to get

$$\frac{1}{n}\sum_{j=1}^{n-1} \delta_{\frac{j-1}{n+1},\frac{j}{n+1},\frac{j+1}{n+1},\chi_j}$$

$$\Rightarrow Q = \mathcal{L}_{(X,X,X)} \otimes \mathcal{L}_W \otimes \mathcal{L}_W \otimes \mathcal{L}_W$$

and we conclude applying Key lemma 2.

Sketch of the proof of the CLT

Finally, we have a CLT for

$$B_N = \frac{1}{N}\sum_{j=1}^{N-1} f\left(W_{N,j}\right) f\left(W_{N,j+1}\right)$$

and, conditionnally to $\mathcal{F}_W$, a CLT for

$$C_N = \frac{1}{N}\sum_{j=1}^{N-1} \left(\Delta_{N,j} f\left(W_{N,j+1}\right) + \Delta_{N,j+1} f\left(W_{N,j}\right)\right)$$

For any s and $t \in \mathbb{R}$,

$$\mathbb{E}\left[e^{i(\sqrt{N}s(B_N-\mathbb{E}[B_N])+\sqrt{N}tC_N)}\right] = \mathbb{E}\left[e^{i\sqrt{N}s(B_n-\mathbb{E}[B_N])}\mathbb{E}\left[e^{i\sqrt{n}tC_N}\middle|\mathcal{F}_W\right]\right]$$

- $\mathbb{E}\left[e^{i\sqrt{N}tC_N}\middle|\mathcal{F}_W\right] \to \exp\{-\sigma_C^2 t^2/2\}$ a.s. not random ;
- $\sqrt{N}s(B_N - \mathbb{E}[B_N]) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma_B^2)$.

Sketch of the proof of the CLT

By Slutsky's lemma,

$$\left(\sqrt{N}s(B_N - \mathbb{E}[B_N]), \mathbb{E}\left[e^{i\sqrt{N}tC_N}\middle|\mathcal{F}_W\right]\right) \xrightarrow[N\to\infty]{\mathcal{L}} (B_s, \exp\{-\sigma_C^2 t^2/2\}).$$

We consider $h\colon (u,v) \in \mathbb{R} \times D(0,1) \mapsto e^{iu}v \in \mathbb{C}$ where $D(0,1)$ is the unit disc in $\mathbb{C}$: $e^{i\sqrt{N}s(B_N-\mathbb{E}[B_N])}\left[e^{i\sqrt{N}tC_N}\middle|\mathcal{F}_W\right]$ cv in distribution.
Finally,

$$\sqrt{N}(B_N - \mathbb{E}[B_N], C_N) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_2\left(0, \begin{pmatrix} \sigma_B^2 & 0 \\ 0 & \sigma_C^2 \end{pmatrix}\right).$$

We conclude using the delta method.

Mighty estimation based on ranks

Références

- F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux. "Global Sensitivity Analysis : a new generation of mighty estimators based on rank statistics", *Bernoulli.* 2021.
- S. Chatterjee. "A new coefficient of Correlation", *JASA*, 2020.
- S. Da Veiga, and F. Gamboa. "Efficient estimation of sensitivity indices", *Journal of Nonparametric Statistics* 2013.

Outline of the talk

Introduction to SA and Sobol' indices

The classical Pick-Freeze estimation

Mighty estimation based on ranks

Comparison of the different estimation procedures

Numerical applications

Comparison of the different estimation procedures

Example

We consider the following linear model

$$Y = f(X_1, \ldots, X_p) = \alpha X_1 + X_2 + \ldots + X_p,$$

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

We consider the estimation of $\mathbb{E}[\mathbb{E}[Y|X_1]^2]$ only.

Comparison of the different estimation procedures

Pick-Freeze

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

Ranks

$$V_{\mathrm{Rank}}^{1,1} = \frac{4}{45}\alpha^4 + \frac{1}{3}m_{1,p}\alpha^3 + \frac{1}{3}\left(4v_p + m_{1,p}^2\right)\alpha^2 + 4m_{1,p}v_p\alpha + v_p\left(v_p + 4m_{1,p}^2\right)$$

Efficient

$$V_{\mathrm{Eff}}^{1} = \frac{4}{45}\alpha^4 + \frac{1}{3}m_{1,p}\alpha^3 + \frac{1}{3}\left(4v_p + m_{1,p}^2\right)\alpha^2 + 4m_{1,p}v_p\alpha + 4v_p m_{1,p}^2,$$

where $m_{1,p}$ and v_p stand for the expectation and the variance of $X_2 + \ldots + X_p$.

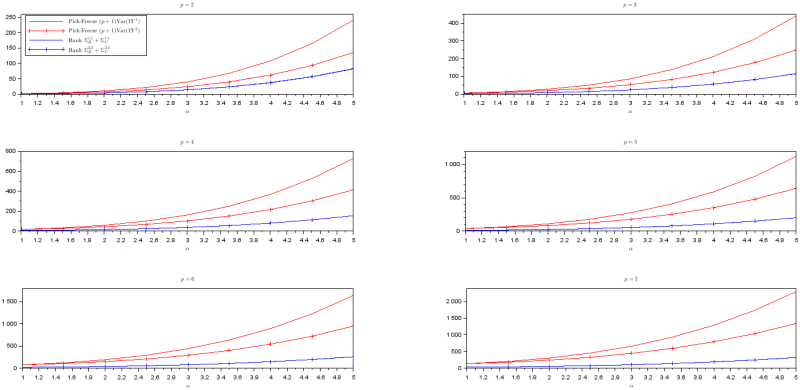


Figure – Limiting variances with respect to X_1 (−) and to X_2 (−+) for $p = 2$ to $p = 7$. The rank-based variances are represented in blue while the Pick-Freeze variances are represented in red.

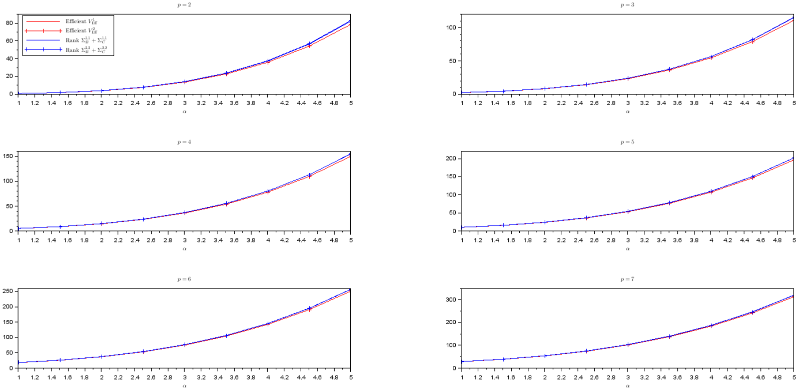


Figure – Limiting variances with respect to X_1 (–) and to X_2 (–+) for $p = 2$ to $p = 7$. The rank-based variances are represented in blue while the efficient variances are represented in red.

Outline of the talk

Introduction to SA and Sobol' indices

The classical Pick-Freeze estimation

Mighty estimation based on ranks

Comparison of the different estimation procedures

Numerical applications

A non-linear model (I)

Let us consider the following non-linear model

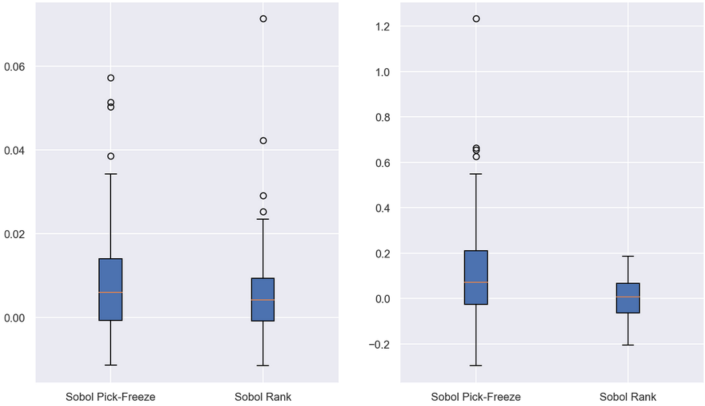
$$Y = \exp\{X_1 + 2X_2\},$$

where X_1 and X_2 are independent standard Gaussian random variables. Then tedious computations lead to the Sobol' indices S^1 and S^2 :

$$S^1 = (e-1)/(e^5-1) \approx 0.0117$$
$$S^2 = (e^4-1)/(e^5-1) \approx 0.3636$$

A non-linear model (II)

Comparison of the estimation procedures with $N = 10^5$ and $n_{\mathrm{rep}} = 100$.



The so-called Sobol' g-function

The g-function is defined by

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

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

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

As expected, the lower the coefficient a_i, the more significant the variable X_i. In the sequel, we simply fix $a_i = i$.

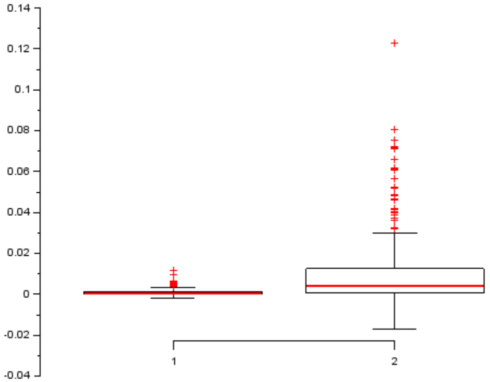


Figure – The Sobol' g-function model. Boxplot of the square errors of the estimation of S^1 with a fixed sample size and 500 replications. Rank methodology with $n = 700$ - left. Pick-Freeze estimation procedure with $N = 100$ - right.

	Ratio Pick-Freeze/Rank		
	$N = 10/n = 70$	$N = 50/n = 350$	$N = 100/n = 70$
mse S^1	10.35%	13.32%	10.69%
mse S^2	11.76%	15.11%	16.38%
mse S^3	11.95%	14.76%	16.37%
mse S^4	10.02%	17.29%	17.56%
mse S^5	09.63%	13.41%	16.62%
mse S^6	11.37%	13.62%	16.22%

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

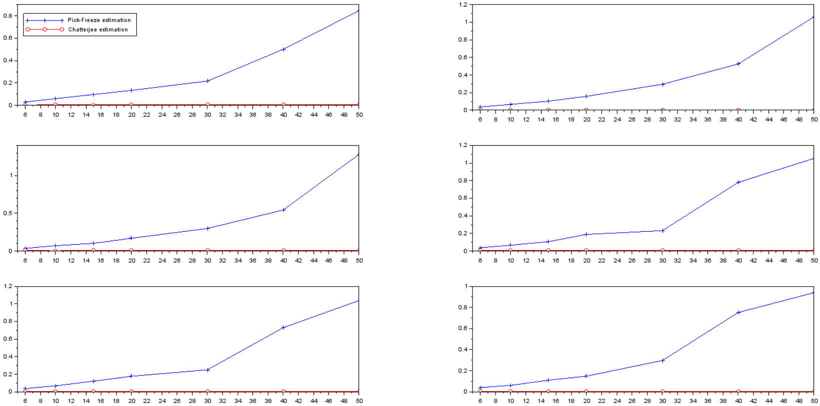


Figure – The Sobol' g-function model. Mean square errors of the estimation of the six first-order Sobol' indices with respect to p (6,10,15,20,30,40,50), with a fixed sample size (rank - red - $n = 200$; Pick-Freeze - blue - $N = n/(p+1)$) and 500 replications.

Thanks for your attention !
Any questions ?