Universal Sensitivity Indices - Application to Stochastic Codes and Second Level SA

Agnès Lagnoux

Institut de Mathématiques de Toulouse
Université Toulouse Jean Jaurès
TOULOUSE - FRANCE

Journée du GST "Mécanique et Incertain"
October 19th, 2023

Outline of the talk

General framework

Complex function f depending on several variables :

$$y = f(x_1, \ldots, x_p)$$

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.

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 :

$$X := (X_1, \ldots, X_p) \in \mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_p.$$

Then f is a measurable function that can be evaluated on runs and the output code Y becomes random too :

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

The question is :

how one may quantify the amount of randomness that a variable or a group of input variables bring to the output Y ?

Toy example

Let have a look on a simple example :

$$(X_1, X_2, X_3, X_4) \mapsto Y = X_1 + X_2 + X_1 X_3$$

where X_1, X_2, X_3, X_4 are independent, centered and so that

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_3) = \mathrm{Var}(X_4) = 1, \ \mathrm{Var}(X_2) = 2.$$

Obviously,

1. Y is not depending on X_4 ;
2. X_2 should be more influent than X_1 at first order since its variance is greater than the one of X_1 ;
3. X_1 should be more influent than X_3 as it appears once alone (term X_1) and once related to X_3 (term $X_1 X_3$).

The so-called Sobol' indices

An input variable is influent if its variations lead to strong variations on the output Y.

$\Rightarrow$ *Build an index of influence on the variance of the output Y.*

For instance, the first order Sobol' index with respect to $X_{\mathbf{u}} = (X_i, i \in \mathbf{u})$ where $\mathbf{u} \subset \{1, \cdots, p\}$ 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.

Toy example (continued)

We consider again

$$Y = f(X) = f(X_1, X_2, X_3, X_4) = X_1 + X_2 + X_1 X_3$$

where X_1, X_2, X_3, X_4 are independent, centered and so that

$$\mathrm{Var}(X_1) = \mathrm{Var}(X_3) = \mathrm{Var}(X_4) = 1,\ \mathrm{Var}(X_2) = 2.$$

Then

$$S^1 = 1/4,\ S^2 = 1/2,\ S^{13} = 1/4,$$

and

$$S^3 = S^4 = S^{12} = S^{14} = S^{23} = S^{24} = S^{34} = 0, S^{ijk} = 0\ \forall i, j, k.$$

Motivatory example for second level SA

Let us consider the linear model

$$Y = X_1 + X_2,$$

where X_1 and X_2 are two independent centered random variables with respective variance θ^2 and $1-\theta^2$. Donc $\mathrm{Var}(Y) = 1$.

Motivatory example for second level SA

Let us consider the linear model

$$Y = X_1 + X_2,$$

where X_1 and X_2 are two independent centered random variables with respective variance θ^2 and $1-\theta^2$. Donc $\mathrm{Var}(Y) = 1$.

Naturally, the first order Sobol' indices are given by

$$S^1 = \frac{\mathrm{Var}(\mathbb{E}[Y|X_1])}{\mathrm{Var}(Y)} = \theta^2 \quad \text{and} \quad S^2 = \frac{\mathrm{Var}(\mathbb{E}[Y|X_2])}{\mathrm{Var}(Y)} = 1-\theta^2$$

so that

$$S^1 < S^2 \quad \text{if} \quad \theta < 1/\sqrt{2} \quad \text{and} \quad S^1 \geqslant S^2 \quad \text{if} \quad \theta \geqslant 1/\sqrt{2}.$$

Second level sensitivity analysis

The second level uncertainty corresponds to the uncertainty on the type of the input distributions and/or on the parameters of the input distributions.

Outline of the talk

Introduction to stochastic codes

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

Here f is assumed to be a stochastic code : two evaluations of the code for the same input $x^* = (x_1^*, \ldots, x_p^*)$ lead to two different outputs.

The practitioner is then interested in the distribution μ_{x^*} of the output Y for a given x^*.

How to perform ?

A first step to deal with stochastic codes

A natural way to handle stochastic computer codes is definitely

- to consider the expectation of the output code
- and to perform GSA on this expectation.

Our procedure

This type of code can be traduced in terms of a deterministic code by considering an extra input D which is not chosen and not observed by the practitioner itself but which is a latent variable generated randomly by the computer code and independently of the classical input $(X_1, \cdots, X_p)$.

References for this work

F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux.
Global Sensitivity Analysis : a new generation of mighty estimators based on rank statistics.
Bernoulli. 2022.

J.-C. Fort, T. Klein, and A. Lagnoux.
Global sensitivity analysis and Wasserstein spaces.
SIAM UQ, 2021.

Two related (deterministic) applications

Thus one considers

1. a first (deterministic) code

$$f_s : \begin{array}{ccc} \mathcal{E} \times \mathcal{D} & \to & \mathbb{R} \\ (x, d) = (x_1, \cdots, x_p, d) & \mapsto & f_s(x, d) = f_s(x_1, \cdots, x_p, d); \end{array}$$

2. a second (deterministic) code whose output is a probability measure

$$f : \begin{array}{cc} \mathcal{E} & \to \mathcal{M}_2(\mathbb{R}) \\ x & \mapsto \mu_x. \end{array}$$

Obviously, in practice, one does not assess the output of f but one can only obtain an empirical approximation of the measure μ_x given by n evaluations of f_s at x.
Further, f can be seen as an ideal version of f_s.

In practice...

Concretely, for a single random input $X^* \in \mathcal{E} = \mathcal{E}_1 \times \cdots \times \mathcal{E}_p$, we evaluate n times f_s (so that the code will generate independently n hidden variables D_1, ..., D_n) and one may observe

$$f_s(X^*, D_1), \ldots, f_s(X^*, D_n)$$

leading to the measure

$$\mu_{X^*,n} = \frac{1}{n} \sum_{k=1}^{n} \delta_{f_s(X^*, D_k)}$$

approximating the distribution $\mu_{X^*} = f(X^*)$.

Remind the random variables $D_1, \ldots, D_n$ are not observed.

In practice...

Finally, the general design of experiments is the following :

$$\begin{array}{ccc} (X_1, D_{1,1}, \dots, D_{1,n}) & \to & f_s(X_1, D_{1,1}), \dots, f_s(X_1, D_{1,n}), \\ & \vdots & \\ (X_N, D_{N,1}, \dots, D_{N,n}) & \to & f_s(X_N, D_{N,1}), \dots, f_s(X_N, D_{N,n}), \end{array}$$

where $N \times n$ is the total number of evaluations of the stochastic computer code f_s. Then we construct the approximations of μ_{X_j} for any $j = 1, \dots, N$ given by

$$\mu_{X_j,n} = \frac{1}{n} \sum_{k=1}^{n} \delta_{f_s(X_j, D_{j,k})}.$$

Framework and notation

Here, the output of the code f is a probability measure (or equivalently a density or a cumulative distribution function) on $\mathbb{R}$.

Then we introduce the Wasserstein metric W_2 of order 2 on the output space : for two probability measures μ and ν with c.d.f. F_μ and F_ν respectively, one has

$$
\begin{aligned}
W_2^2(\mu,\nu) &= \inf\left\{\mathbb{E}\left[d(X,Y)^2\right]\ ;\ \mathbb{P}_X=\mu \textit{ and } \mathbb{P}_Y=\nu\right\} \\
&= \int_0^1 (F_\mu^{-1}(t)-F_\nu^{-1}(t))^2 dt = \mathbb{E}[|F_\mu^{-}(U)-F_\nu^{-}(U)|^2].
\end{aligned}
$$

Here F_μ^{-1} and F_ν^{-1} are the generalized inverses of the increasing functions F_μ and F_ν and $U \sim \mathcal{U}([0,1])$.

Natural origin

Recall that the Sobol' index is given by

$$S^{\mathbf{u}} = \frac{\mathrm{Var}(\mathbb{E}[Y|X_{\mathbf{u}}])}{\mathrm{Var}(Y)} = \frac{\mathbb{E}[(\mathbb{E}[Y] - \mathbb{E}[Y|X_{\mathbf{u}}])^2]}{\mathrm{Var}(Y)}$$

that generalizes in the Cramér-von Mises index defined as

$$S^{\mathbf{u}}_{2,CVM} = \frac{\int_{\mathbb{R}} \mathbb{E}\left[(\mathbb{E}[\mathbb{1}_{Y\leqslant t}] - \mathbb{E}[\mathbb{1}_{Y\leqslant t}|X_{\mathbf{u}}])^2\right] dF(t)}{\int_{\mathbb{R}} \mathrm{Var}(\mathbb{1}_{Y\leqslant t}) dF(t)} = \frac{\int_{\mathbb{R}} \mathbb{E}\left[(F(t) - F^{\mathbf{u}}(t))^2\right] dF(t)}{\int_{\mathbb{R}} F(t)(1-F(t)) dF(t)}.$$

Sensitivity index

Let us denote by $\mathbb{F}$ the c.d.f. of the output of the code (it depends on the input variables).

The universal index $S^{\mathbf{u}}_{2,W_2}(\mathbb{F})$ with respect to $X_{\mathbf{u}} = (X_i, i \in \mathbf{u})$ is :

$$\frac{\int_{\mathcal{W}_2(\mathbb{R})^2} \mathbb{E}\left[\left(\mathbb{E}[\mathbb{1}_{W_2(F_1,\mathbb{F})\leqslant W_2(F_1,F_2)}] - \mathbb{E}[\mathbb{1}_{W_2(F_1,\mathbb{F})\leqslant W_2(F_1,F_2)}|X_{\mathbf{u}}]\right)^2\right] d\mathbb{P}^{\otimes 2}(F_1,F_2)}{\int_{\mathcal{W}_2(\mathbb{R})^2} \mathrm{Var}(\mathbb{1}_{W_2(F_1,\mathbb{F})\leqslant W_2(F_1,F_2)}) d\mathbb{P}^{\otimes 2}(F_1,F_2)}.$$

Estimation procedure

In order to compute explicitly our estimator, it remains to compute terms of the form :

$$W_2(\mu_{n,X_i}, \mu_{n,X_j}).$$

Actually, such quantities are easy to compute since for two discrete measures supported on a same number of points and given by

$$\nu_1 = \frac{1}{n}\sum_{k=1}^{n} \delta_{x_k},\ \nu_2 = \frac{1}{n}\sum_{k=1}^{n} \delta_{y_k},$$

the Wasserstein distance between ν_1 and ν_2 simply writes

$$W_2^2(\nu_1, \nu_2) = \frac{1}{n}\sum_{k=1}^{n} (x_{(k)} - y_{(k)})^2,$$

where $z_{(k)}$ is the k-th order statistics of z.

Numerical study (I)

Let X_1, X_2, X_3 be 3 independent random variables Bernoulli distributed with parameter p_1, p_2, and p_3 respectively. We consider the c.d.f.-valued code, the output of which is given by

$$\mathbb{F}_{(X_1,X_2,X_3)}(t) = \frac{t}{1 + X_1 + X_2 + X_1X_3} \mathbb{1}_{0 \leqslant t \leqslant 1+X_1+X_2+X_1X_3}$$

$$+ \mathbb{1}_{1+X_1+X_2+X_1X_3<t},$$

so that

$$\mathbb{F}^{-1}_{(X_1,X_2,X_3)}(v) = v\Big(1 + X_1 + X_2 + X_1X_3\Big).$$

Numerical study (II)

Thus we consider the ideal code :

$$\begin{array}{rccc} f: & \mathcal{E} & \to & \mathcal{W}_2(\mathcal{E}) \\ & (X_1, X_2, X_3) & \mapsto & \mu_{(X_1,X_2,X_3)} \sim \mathbb{F}_{(X_1,X_2,X_3)} \end{array}$$

where $\mu_{(X_1,X_2,X_3)} \sim \mathcal{U}([0, 1 + X_1 + X_2 + X_1 X_3])$ and its stochastic counterpart :

$$\begin{array}{rccc} f_s: & \mathcal{E} \times \mathcal{D} & \to & \mathbb{R} \\ & (X_1, X_2, X_3, D) & \mapsto & f_s(X_1, X_2, X_3, D) \end{array}$$

where $f_s(X_1, X_2, X_3, D)$ is a realization of $\mu_{(X_1,X_2,X_3)}$.

Numerical study (III)

Hence, we do not assume that one may observe N realizations of $\mathbb{F}$ associated to N initial realizations of (X_1, X_2, X_3). Instead, for any of the N initial realizations of (X_1, X_2, X_3), we assess n realizations of a uniform random variable on $[0, T]$ where $T = 1 + X_1 + X_2 + X_1 X_3$.

We assume that only $N = 450$ calls of the computer code f are allowed to estimate the indices $S^{\mathbf{u}}_{2,W_2}$ for $\mathbf{u} = \{1\}$, $\{2\}$, and $\{3\}$.

The empirical c.d.f. based on the empirical measures $\mu_{i,n}$ for $i = 1, \ldots, n$ are constructed with $n = 500$ evaluations. We repeat the estimation procedure 200 times.

Numerical study (IV)

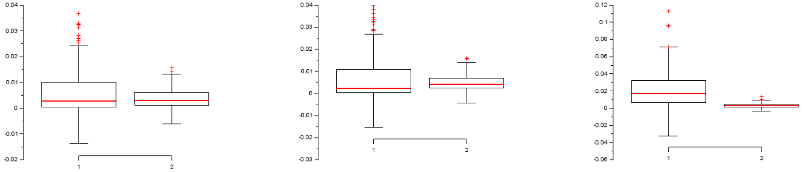

Figure – Boxplot of the mean square errors of the estimation of the Wasserstein indices $S^{\mathbf{u}}_{2,W_2}$. The indices with respect to $\mathbf{u} = \{1\}$, $\{2\}$, and $\{3\}$ are displayed from left to right. The results of the Pick-Freeze estimation procedure with $N = 64$ are provided in the left side of each graphic. The results of the rank-based methodology with $N = 450$ are provided in the right side of each graphic.
Here, $p_1 = 1/3$, $p_2 = 2/3$, and $p_3 = 3/4$

Outline of the talk

Link with stochastic computer codes

We denote by μ_i $(i = 1, \ldots, p)$ the distribution of the input X_i and we assume that each μ_i belongs to some parametric family $\mathcal{P}_i$ of probability measures endowed with a probability measure $\mathbb{P}_{\mu_\mathbf{i}}$:

$$\mathcal{P}_i := \{\mu_\theta, \theta \in \Theta_i \subset \mathbb{R}^{d_i}\}$$

where Θ_i is endowed with a probability measure ν_{Θ_i}.

Link with stochastic computer codes

Consider the stochastic mapping f_s from $\mathcal{P}_1 \times \ldots \times \mathcal{P}_p$ to $\mathcal{Y}$ defined by

$$f_s(\mu_1, \ldots, \mu_p) = f(X_1, \ldots, X_p)$$

where $X_1, \ldots, X_p$ are independently drawn according to the distribution $\mu_1 \times \ldots \times \mu_p$.

Hence f_s is a stochastic computer code from $\mathcal{P}_1 \times \ldots \times \mathcal{P}_p$ to $\mathcal{Y}$ and we can perform sensitivity analysis using the indices defined previously.

Numerical study - model

We consider the synthetic example defined on $[0,1]^3$ by

$$f(X_1, X_2, X_3) = 2X_2 e^{-2X_1} + X_3^2,$$

where X_i are independent uniform random variables.

We are interested in the uncertainty in the support of the random variables X_1, X_2 and X_3.

Numerical study - model

We consider the synthetic example defined on $[0,1]^3$ by

$$f(X_1, X_2, X_3) = 2X_2 e^{-2X_1} + X_3^2,$$

where X_i are independent uniform random variables.

We are interested in the uncertainty in the support of the random variables X_1, X_2 and X_3.

Thus we assume

- $X_i \sim \mu_i = \mathcal{U}([A_i, B_i])$;
- $A_i \sim \mathcal{U}([0, 0.1])$;
- $B_i \sim \mathcal{U}([0.9, 1])$.

Numerical study - SA

1. For all i, we produce a N-sample $([A_{i,j}, B_{i,j}])_{j=1,\dots,N}$ of intervals $[A_i, B_i]$.
2. For all i and, for $1 \leqslant j \leqslant N$, we generate a n-sample $(X_{i,j,k})_{k=1,\dots,n}$ of X_i, where $X_{i,j,k} \sim \mathcal{U}([A_{i,j}, B_{i,j}])$.
3. For $1 \leqslant j \leqslant N$, we compute the n-sample $(Y_{j,k})_{k=1,\dots,n}$ of the output using

$$Y = f(X_1, X_2, X_3) = 2X_2 e^{-2X_1} + X_3^2.$$

   Thus we get a N-sample of the empirical measures of the distribution of the output Y given by :

$$\mu_{j,n} = \frac{1}{n} \sum_{k=1}^{n} \delta_{Y_{j,k}}, \quad \text{for } j = 1, \dots, N.$$

4. Finally, it remains to compute the indicators $S^{\mathbf{u}}_{2,W_2}$ and their means to get the Pick-Freeze estimators of $S^{\mathbf{u}}_{2,W_2}$, for $\mathbf{u} = \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}$, and $\{2,3\}$.

Numerical study - First illustration

We compute the estimators of $S^{\mathbf{u}}_{2,W_2}$ following the previous procedure with $N = 500$ and $n = 500$ and

1. Case 1 : $A_i \sim \mathcal{U}([0, 0.1])$ and $B_i \sim \mathcal{U}([0.9, 1])$,
2. Case 2 : $A_i \sim \mathcal{U}([0, 0.45])$ and $B_i \sim \mathcal{U}([0.55, 1])$.

	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
Case 1	0.07022	0.08791	0.09236	0.14467	0.21839	0.19066
Case 2	0.11587	0.06542	0.169529	0.22647	0.40848	0.34913

Numerical study - Second illustration

We run another simulations allowing for more variability on the upper bound related to the third input X_3 only :

$$B_3 \sim \mathcal{U}([0.5, 1]).$$

$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
0.01196	0.06069	0.56176	-0.01723	0.63830	0.59434

Reminder

	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
$A_i \in [0, 0.1]$ $B_i \in [0.9, 1]$	0.07022	0.08791	0.09236	0.14467	0.21839	0.19066

Numerical study - Third illustration

We perform a classical GSA on the inputs rather than on the parameters of their distributions : we estimate the index $S^{\mathbf{u}}_{2,CVM}$ with a sample size $N = 10^4$.

$\mathbf{u}$	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
$\hat{S}^{\mathbf{u}}_{2,\mathrm{CVM}}$	0.13717	0.15317	0.33889	0.33405	0.468163	0.53536

Reminder for $\hat{S}^{\mathbf{u}}_{2,W_2}$

	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
$A_i \in [0, 0.1]$ $B_i \in [0.9, 1]$	0.07022	0.08791	0.09236	0.14467	0.21839	0.19066

Thanks for your attention !

Any questions ?