Sensitivity analysis for stochastic computer codes and second-level sensitivity analysis

Agnès Lagnoux
Institut de Mathématiques de Toulouse
TOULOUSE - FRANCE

CIROQUO, July 2th 2021

OUTLINE OF THE LECTURES

Part I : From Sobol' indices to universal indices

Part II : Stochastic computer codes and an introduction to second-level sensitivity analysis

Part II
Stochastic computer codes and an introduction to second-level sensitivity analysis

Stochastic computer codes
State of the art
Sensitivity analysis for stochastic computer codes

Second-level sensitivity analysis
Introductory example
Second level sensitivity analysis
Link with stochastic computer codes
Numerical applications

Outline of the talk

Stochastic computer codes
- State of the art
- Sensitivity analysis for stochastic computer codes

Second-level sensitivity analysis
- Introductory example
- Second level sensitivity analysis
- Link with stochastic computer codes
- Numerical applications

General framework

Complicated function f depending on several variables :

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

where

1. the inputs x_i pour $i = 1, \dots p$ 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 $(x_1, \dots, x_p)$.*
2. *Identify the most important variables to be able to fix the less important ones to their nominal value.*

Introduction to stochastic codes

Here f is assumed to be a real-valued 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 for a given x.

Introduction to stochastic codes

Typical stochastic computer codes are

- agent-based models (Siebers et al. 2010), for instance simulating disease propagation (Boukouvalas and Cornford 2009) or atmospheric pollution (Reich et al. 2011) ;
- models involving partial differential equations applied to heterogeneous random media, for instance fluid flows in oil reservoirs (Zabalza et al. 1998) or acoustical wave propagation in turbulent fluids (Iooss et al. 2002) ;
- models involving stochastic differential equations (Le Maître et al. 2015) and (Etoré et al. 2020) ;

Introduction to stochastic codes

- the unitary simulations of Monte Carlo neutronic models (computing elementary particle trajectories in a nuclear reactor, Picheny et al. 2011) and the Lagrangian stochastic models (computing particle trajectories inside atmospheric or hydraulic turbulent media, Pope 1994). ;
- ...

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.

D. Bursztyn and D. M. Steinberg. Screening experiments for dispersion effects. *Screening*, pages 21–47. Springer, 2006.
B. Ankenman, B. L. Nelson, and J. Staum. Stochastic kriging for simulation metamodeling. *Winter Simulation Conf.*, pages 362–370. IEEE, 2008.
G. Dellino and C. Meloni. *Uncertainty management in simulation-optimization of complex systems*. Springer, 2015.
J. P. Kleijnen. Design and analysis of simulation experiments. *International Workshop on Simulation*, pages 3–22. Springer, 2015.

Traducing the randomness of the code (I)

Another approach is to consider that the stochastic code is of the form $f(X, D)$ where X contains the classical input variables and D is an extra unobserved random input.

A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM : Probability and Statistics*, 18 :342–364, 1 2014.

Such an idea is exploited to compare the estimation of the Sobol' indices in an "exact" model to the estimation of the Sobol' indices in an associated metamodel.

Traducing the randomness of the code (II)

G. Mazo. An optimal tradeoff between explorations and repetitions in global sensitivity analysis for stochastic computer models. *Submitted* 2019.

Mazo builds two different indices.

1. The first index is obtained by substituting $f(X, D)$ for $f(X)$ in the classical definition of the first order Sobol' index

$$S^i = \mathrm{Var}(\mathbb{E}[f(X)|X_i])/\mathrm{Var}(f(X)).$$

   In this case, D is considered as another input, even though it is not observable.

2. The second index is obtained by substituting $\mathbb{E}[f(X, D)|X]$ for $f(X)$ in the Sobol' index. The noise is then smoothed out.

Traducing the randomness of the code (III)

J. L. Hart, A. Alexanderian, and P. A. Gremaud. Efficient computation of Sobol'indices for stochastic models. *SIAM Journal on Scientific Computing*, 39(4) :A1514–A1530, 2017.

Their algorithm returns n realizations of the first-order Sobol' index S^i : $S^i_j(D_j)$ for $1 \leqslant j \leqslant n$ and $1 \leqslant i \leqslant p$.

Then, for any $i = 1, \ldots, p$, they approximate the statistical properties of S^i by considering the sample r-th moments :

$$\hat{\mu}^i_r = \frac{1}{n} \sum_{j=1}^{n} (S^i_j(D_j))^r$$

noticing that

$$\mathbb{E}_D[\hat{\mu}^i_r] = \mathbb{E}_D[(S^i)^r] \quad \text{and} \quad \mathrm{Var}_D(\hat{\mu}^i_r) = \frac{1}{n}\mathrm{Var}_D((S^i)^r).$$

Our procedure

Remind f is assumed to be a real-valued 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 for a given x.

This type of code can be traduced in terms of a deterministic code by considering an extra input which is not chosen by the practitioner itself but which is a latent variable generated randomly by the computer code and independently of the classical input.

References for this section

F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux.
Global Sensitivity Analysis : a new generation of mighty estimators based on rank statistics.
Preprint Arxiv. 2020.

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

Two related (deterministic) applications

Thus one considers

1. a first (determisnistic) code

$$\begin{aligned} f_s: \quad E \times D &\to \mathbb{R} \\ (x, d) &\mapsto f_s(x, d); \end{aligned}$$

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

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

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 textcolorblueideal version of f_s.

In practice...

Concretely, for a single random input $X \in E = E_1 \times \cdots \times E_p$, 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 of $f_s(X)$.

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

In practice...

Finally, the general design of experiments is the following :

$$
\begin{array}{rcl}
(X_1, D_{1,1}, \ldots, D_{1,n}) & \rightarrow & f_s(X_1, D_{1,1}), \ldots, f_s(X_1, D_{1,n}), \\
 & \vdots & \\
(X_N, D_{N,1}, \ldots, D_{N,n}) & \rightarrow & f_s(X_N, D_{N,1}), \ldots, f_s(X_N, D_{N,n}),
\end{array}
$$

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 $j = 1, \ldots, N$ 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 Wassertein 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

$$W_2^2(\mu,\nu) = \int_0^1 (F_\mu^{-1}(t) - F_\nu^{-1}(t))^2 dt = \mathbb{E}[|F_\mu^-(U) - F_\nu^-(U)|^2].$$

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

Sensitivity index

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

Then the universal index $S^{\mathbf{u}}_{2,W_2}$ with respect to $X^{\mathbf{u}}$ is :

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

Estimation procedure

1. Generate a Pick-Freeze sample of size N : $(X_i, X_i^{\mathbf{u}})_{1\leqslant i\leqslant N}$
2. For each input $(X_i, X_i^{\mathbf{u}})$, compute the corresponding output n times :

$$Z_{i,j} = f_s(X_i, D_j) \text{ and } Z_{i,j}^{\mathbf{u}} = f_s(X_i^{\mathbf{u}}, D_j'),\ 1 \leqslant i \leqslant N,\ 1 \leqslant j \leqslant n.$$

3. Approximate the measures by the corresponding empirical measures

$$\mu_{X_i} \approx \mu_{n,X_i} = \frac{1}{n}\sum_{j=1}^{n} \delta_{Z_{i,j}} \quad \textit{and} \quad \mu_{X_i^{\mathbf{u}}} \approx \mu_{n,X_i^{\mathbf{u}}} = \frac{1}{n}\sum_{j=1}^{n} \delta_{Z_{i,j}^{\mathbf{u}}}.$$

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 application (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 f, the output of which is given by

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

Numerical application (II)

Thus we consider the (ideal) code :

$$\begin{array}{rcl} f: & E & \to \mathcal{M}_2(E) \\ & (X_1, X_2, X_3) & \mapsto \mu_{(X_1,X_2,X_3)} \end{array}$$

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 :

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

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

Numerical application (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 $[0, 1 + X_1 + X_2 + X_1 X_3]$.

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 application (IV)

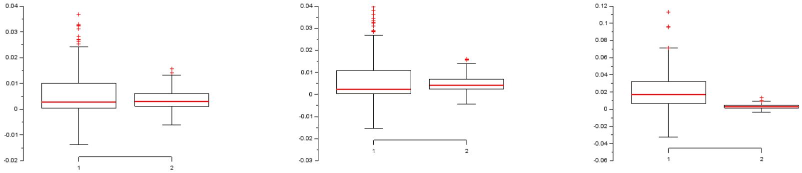

Figure – Boxplot of the mean square errors of the estimation of the Wasserstein indices $S^{\mathbf{u}}_{2,W_2}$. The indices with respect to $\mathbf{u} = \{1\}$, $\{2\}$, and $\{3\}$ are displayed from left to right. The results of the Pick-Freeze estimation procedure with $N = 64$ are provided in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

Red-thread example : 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. Here we assume that

$$X_2 = \frac{G_1 + G_2}{\sqrt{2}},$$

where G_1 and G_2 are independent standard Gaussian random variables, independent of X_1. In addition, the practitioner has access only to X_1 and G_1.

Red-thread example : a non-linear model (II)

Code TP_Sto.ipynb

Outline of the talk

Stochastic computer codes

- State of the art
- Sensitivity analysis for stochastic computer codes

Second-level sensitivity analysis

- Introductory example
- Second level sensitivity analysis
- Link with stochastic computer codes
- Numerical applications

Introductory example

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 $1-\theta^2$.

Naturally, the first order Sobol indices are given by

$$S^1 = \theta^2 \quad \text{and} \quad S^2 = 1-\theta^2$$

so that

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

Second level sensitivity analysis

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

A first natural step consists in studying the expectation with respect to the distribution of the parameters of the conditional output.

More interestingly, such uncertainties can be handled in two different manners :

1. aggregating them with no distinction (like, e.g. in Vincent Chabridon's thesis),
2. separating them (like, e.g. in Anouar Meynaoui's thesis).

Reference for this section

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

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 $\mathcal{X}$ defined by

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

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 $\mathcal{X}$ 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

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

and introduced in Gremaud et al. (2019). Here we are interested in the uncertainty in the support of the random variables X_1, X_2 and X_3.

Numerical study - first results

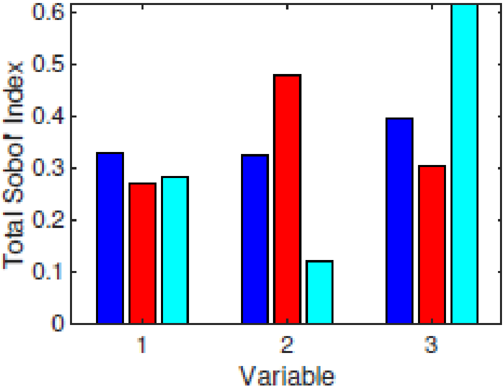

Figure 2. *Total Sobol' indices of (7). Blue bars denote the indices computed using the nominal distribution, red bars denote the indices when the distribution is perturbed to maximize T_2, cyan bars denote the indices when the distribution is perturbed to minimize T_3.*

Numerical study - methodology

Here, we adopt the methodology explained previously and we consider the stochastic code given by :

$$f_s(\mu_1, \mu_2, \mu_3) = 2X_2 e^{-2X_1} + X_3^2,$$

where

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

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 :
$$\mu_{X_j,n} = \frac{1}{n}\sum_{k=1}^{n} \delta_{Y_{j,k}}, \quad \text{for } j = 1, \dots, N.$$
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. with $A_i \sim \mathcal{U}([0, 0.1])$ and $B_i \sim \mathcal{U}([0.9, 1])$,
2. with $A_i \sim \mathcal{U}([0, 0.45])$ and $B_i \sim \mathcal{U}([0.55, 1])$.

	$\{1\}$	$\{2\}$	$\{3\}$	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$
$A_i \in [0, 0.1]$ $B_i \in [0.9, 1]$	0.07022	0.08791	0.09236	0.14467	0.21839	0.19066
$A_i \in [0, 0.45]$ $B_i \in [0.55, 1]$	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 $S_{2,CVM}^{\mathbf{u}}$ with a sample size $N = 10^4$.

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

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

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

Red-thread example : 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 Gaussian random variables. Here we assume that X_1 is centered and normally distributed with variance σ_1^2 and X_2 is centered and normally distributed with variance σ_2^2.

The aim here is to perform a second-level sensitivity analysis. The distributions of X_1 and X_2 are allowed to vary through their variances.

Red-thread example : a non-linear model (II)

Code

Your turn ! ! !

Thanks for your attention !
Any questions ?