Sobol' sensitivity analysis for stochastic numerical codes

Bertrand, Iooss
EDF Lab Chatou, France

Thierry, Klein
Institut de Mathématiques de Toulouse, France

Agnès, Lagnoux
Institut de Mathématiques de Toulouse, France

Most of the results in sensitivity analysis consider deterministic computer codes, that is codes providing the same output values for the same input variables (Iooss and Lemaître, 2015). For instance, the sensitivity indices of Sobol makes it possible to know the part of the variance output explained by each of the model input. Formally, let us consider the model

$$Y = g(X),$$

where $X = (X_1, \dots X_p)$ is a random vector of independent input parameters (for $i = 1, \dots p$, X_i belongs to some probability space $\mathcal{X}_i$), $Y \in \mathbb{R}$ is the code output and $g(\cdot)$ is a deterministic function representing the computer code.

In this work, we propose to deal with a stochastic computer code denoted by

$$Y = f(X, \varepsilon),$$

where $f(\cdot)$ is the computer code and ε is a random variable representing the physical system randomness (see Marrel et al., 2012, for a typological description of this kind of models). When performing a Sobol' sensitivity analysis on such a code, two different situations occur:

1. We are interested by the full probability density function (pdf) of the outputs. Transformation of this pdf to a few scalar quantities of interest (*e.g.* the first statistical moments of the studied variable) is a first simple solution, while metrics between pdfs can also be used (Douard and Iooss, 2013). Aggregated Sobol' indices (Gamboa et al., 2013) propose a more elegant solution as shown in Le Gratiet et al. (2016) on an application involving probability of detection curves (which look like cumulative distribution functions).

2. We are only interested by the mean value relative to the inherent randomness of the code. In this case (called "Monte Carlo calculation codes" in several engineering domains), we substitute the code by its empirical mean (called "simulator" in this paper).

We focus our analysis on the second situation. In this context, the computer code does not provide the true value of the model (noticed $g(\cdot)$) at x but instead a value $f(x, \epsilon)$ where ε represents the physical system randomness. A standard technique assumes that ε is a random variable such that $\mathbb{E}(f(x, \varepsilon)^2) < \infty$. Hence the real value of g at x can be represented as

$$Y = g(X) := g(X_1, \dots, X_p) = \mathbb{E}(f(X, \varepsilon)|X). \tag{1}$$

For deterministic computer code, by assuming that Y is square integrable and $\mathrm{Var}Y \neq 0$, the corresponding vector of closed Sobol' indices is then

$$S^{\mathbf{u}}_{\mathrm{Cl}}(g) := \left(\frac{\mathrm{Var}(\mathbb{E}(Y|X_i, i \in u_1))}{\mathrm{Var}(Y)}, \dots, \frac{\mathrm{Var}(\mathbb{E}(Y|X_i, i \in u_k))}{\mathrm{Var}(Y)} \right), \tag{2}$$

where $\mathbf{u} := (u_1, \dots, u_k)$ are k subsets of $I_p := \{1, \dots, p\}$. For X and for any subset v of I_p we define X^v by the vector such that $X^v_i = X_i$ if $i \in v$ and $X^v_i = X'_i$ if $i \notin v$ where X' and X are two

independent and identically distributed vectors. We then set $Y^v := g(X^v)$. We also define

$$Z^{\mathbf{u}}_{(i)} = \frac{1}{k+1}\left(Y_{(i)} + \sum_{j=1}^{k} Y^{u_j}_{(i)}\right), \quad M^{\mathbf{u}}_{(i)} = \frac{1}{k+1}\left(Y^2_{(i)} + \sum_{j=1}^{k} (Y^{u_j}_{(i)})^2\right).$$

Taking two independent samples $\left(X_{(i)}\right)_{i=1\ldots N}$ and $\left(X'_{(i)}\right)_{i=1\ldots N}$, where N is the elementary samples size, the Janon-Monod estimator of $S^{\mathbf{u}}_{\mathrm{Cl}}(g)$ is then defined as (Janon et al., 2014):

$$T^{\mathbf{u}}_{N,\mathrm{Cl}}(g) = \left(\frac{\frac{1}{N}\sum Y_{(i)}Y^{u_1}_{(i)} - \left(\frac{1}{2N}\sum (Y_{(i)} + Y^{u_1}_{(i)})\right)^2}{\frac{1}{N}\sum M^{\mathbf{u}}_{(i)} - \left(\frac{1}{N}\sum Z^{\mathbf{u}}_{(i)}\right)^2}, \ldots, \frac{\frac{1}{N}\sum Y_{(i)}Y^{u_k}_{(i)} - \left(\frac{1}{2N}\sum (Y_{(i)} + Y^{u_k}_{(i)})\right)^2}{\frac{1}{N}\sum M^{\mathbf{u}}_{(i)} - \left(\frac{1}{N}\sum Z^{\mathbf{u}}_{(i)}\right)^2}\right). \tag{3}$$

As the computer code cannot provide values of g, we use the Sobol' indices associated to f (instead of g) and study either they are close to the Sobol' indices of g or not. It is then natural to approximate $\mathbb{E}(f(X,\varepsilon)|X=x)$ by its empirical mean. Thus we define what we call a simulator:

$$\widetilde{Y} := \tilde{g}(X,\varepsilon,n) = \frac{1}{n}\sum_{i=1}^{n} f(x,\varepsilon_{(i)}) = g(X) + \delta_n(X,\varepsilon),$$

where n is the sample size (called the number of particles) and $\delta_n(x,\varepsilon)$ is the perturbation. We then define the Sobol' indices associated to $\tilde{g}$ and their estimators using Eqs. (2) and (3) with $\widetilde{Y}$ instead of Y. Moreover, we can prove that the estimator $T^{\mathbf{u}}_{N,\mathrm{Cl}}(\widetilde{g})$ can be used to approximate the true Sobol' indices $S^{\mathbf{u}}_{\mathrm{Cl}}(g)$. Indeed, following the proofs of Janon et al. (2014), we can derive a Central Limit Theorem for this estimator (not shown here in this short abstract).

We will also numerically study the convergence of the Sobol' indices estimates with respect to the sample sizes n (number of particles) and N (size od the elementary samples for Sobol' estimates) considering the following toy function:

$$f(X_1, X_2, \varepsilon) = \sin(X_1(\varepsilon_1 + \varepsilon_2 X_2)) + \varepsilon_3,$$

with the independent random variables $X_1 \sim \mathcal{U}[0,1]$, $X_2 \sim \mathcal{U}[0,1]$, $\varepsilon_1 \sim \mathcal{N}(1,1)$, $\varepsilon_2 \sim \mathcal{N}(2,1)$ and $\varepsilon_1 \sim \mathcal{U}[0,1]$. This leads to a function g defined by

$$g((x_1,x_2)) = \mathbb{E}\left(f(X_1,X_2,\varepsilon)|X_1=x_1, X_2=x_2\right) = \frac{1}{2} + \sin(x_1(1+2x_2))e^{-\frac{x_1^2}{2}(1+x_2^2)}.$$

Finally, an application will consider a Monte Carlo simulator of industrial asset management strategies where the variable of interest is an economic indicator (the Net Present Value).

References:

F. Douard and B. Iooss (2013), Dealing with uncertainty in technical and economic studies of investment strategy optimization. *Proceedings of MASCOT-SAMO 2013 Conference*, Nice, France, July 2013.

F. Gamboa, A. Janon, T.Klein and A. Lagnoux (2014), Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics*, 8:575-603.

B. Iooss and P. Lemaître (2015), A review on global sensitivity analysis methods. In *Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, C. Meloni and G. Dellino (eds), Springer.

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

L. Le Gratiet, B. Iooss, T. Browne, G. Blatman, S. Cordeiro and B. Goursaud (2016). Model assisted probability of detection curves: New statistical tools and progressive methodology, *Submitted*.

A. Marrel, B. Iooss, S. da Veiga and M. Ribatet (2012), Global sensitivity analysis of stochastic computer models with joint metamodels. *Statistics and Computing*, 22:833-847.

[Bertrand Iooss; EDF R&D, 6 Quai Watier, 78401 Chatou, France]
[`bertrand.iooss@edf.fr` – `http://www.gdr-mascotnum.fr/doku.php?id=iooss1`]