Sensitivity analysis: from variance analysis to Cramér Von Mises statistics

BRGM Conference

Fabrice Gamboa-Institut de Mathématiques de Toulouse

17th of January 2018

Obrigado

- Many thanks for the organizing comitee (with special thanks to Jeremy Rohmer)
- My co-authors for this work: A. Janon, T. Klein, A. Lagnoux, C. Prieur

Agenda

Sensitivity analysis-Sobol indices

- From Hoeffding decomposition to Sobol indices
- Pick-Freeze estimation of Sobol indices
- Application: Statistical testing
- Concentration inequality

Cramér-von Mises indices

- Motivation
- Indices based on Cramér-von Mises distance

Goal

Complicated function f valued in $\mathbb{R}^k$ depending on several variables

$$y = f(x_1, \ldots, x_p) \in \mathbb{R}^k.$$

- Generally
  1. f is not analiticaly known
  2. Given $(x_1, \ldots, x_p)$ the computer code give $y = f(x_1, \ldots, x_p)$.
  3. Computing $y = f(x_1, \ldots, x_p)$ may be costly
- Wishes

Identify the most important variables to be able to fix the less important

Probabilistic frame

Inputs are assumed to be random $X := (X_1, \ldots, X_p) \in E := E_1 \times \cdots \times E_p$
$f : E \to \mathbb{R}^k$ is a measurable function evaluable on runs
Y is the code output

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

Main assumptions $X_1, \ldots, X_p$ are independent , $\mathbb{E}(\|Y\|^2) < \infty$ and Y is scalar (here, for sack of simplicity).

The question is:
How one may quantify the amount of **randomness** *that a variable or a group of variable* **bring** *to Y ?*

Let have a look to a simple example:

$$(X_1, X_2, X_3) \mapsto X_1 + X_1 X_2.$$

Obviously

1. Y is not depending on X_3
2. X_1 should be more influent that X_2 as it appears once alone (term X_1) and once related to X_2 (term $X_1 X_2$).

An input variable is influent if its variations leads to strong variations on the output.
$\Rightarrow$ *Build an indice of influence on the variance of the output*

From Hoeffding decomposition to Sobol indices

Let $\mathbf{u}$ be a subset of $\{1,\dots,p\}$ and $\sim\mathbf{u}$ its complementary in $\{1,\dots,p\}$.
Example $p=3$, $\mathbf{u}=\{1\}$ et $\sim\mathbf{u}=\{2,3\}$.
Let denote $X_{\mathbf{u}}=(X_i, i\in\mathbf{u})$ and $X_{\sim\mathbf{u}}=(X_i, i\in\sim\mathbf{u})$.
Decomposition of the output f

$$Y:=f(X)=\underbrace{\mathbb{E}(Y)}_{\text{Mean effect}}+\underbrace{\mathbb{E}(Y|X_{\mathbf{u}})-\mathbb{E}(Y)+\mathbb{E}(Y|X_{\sim\mathbf{u}})-\mathbb{E}(Y)}_{\text{First order effects}}+\underbrace{Y-(\mathbb{E}(Y)+\mathbb{E}(Y|X_{\mathbf{u}})-\mathbb{E}(Y)+\mathbb{E}(Y|X_{\sim\mathbf{u}})-\mathbb{E}(Y))}_{\text{Second order effects or interaction:}=IA}.$$

Factors in the decomposition are orthogonal in L^2. One may compute the variance on both sides

$$\mathrm{Var}(Y)=\mathrm{Var}\left(\mathbb{E}(Y|X_{\mathbf{u}})\right)+\mathrm{Var}\left(\mathbb{E}(Y|X_{\sim\mathbf{u}})\right)+\mathrm{Var}\left(IA\right).$$

This is the so-called *Hoeffding decomposition* of f.
Dividing by $\mathrm{Var}(Y)$ one get

$$1 = \frac{\mathrm{Var}\left(\mathbb{E}(Y|X_{\mathbf{u}})\right)}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}\mathbb{E}(Y|X_{\sim\mathbf{u}})}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}\left(IA\right)}{\mathrm{Var}(Y)}.$$

Definition

The Sobol index with respect to the input $X_{\mathbf{u}} = (X_i, i \in \mathbf{u})$, *is the quantity* $S^{\mathbf{u}} = \frac{\mathrm{Var}(\mathbb{E}(Y|X_{\mathbf{u}}))}{\mathrm{Var}(Y)}$.

This quantity measure the proportion of the output variance coming from variables $X_{\mathbf{u}}$ alone.

Pick-Freeze estimation of Sobol indices

To fix the idea assume for example $p = 5$, $u = \{1, 2\}$ so that $\sim u = \{3, 4, 5\}$

We consider the Pick-Freeze variable Y^u defined as follows

- Draw $X = (X_1, X_2, X_3, X_4, X_5)$.
- Build $X^u = (X_1, X_2, X_3', X_4', X_5')$.

Freeze X_i when $i \in u$ and regenerate realizations for the others Then, compute

- $Y = f(X)$.
- $Y^u = f(X^u)$.

A small miracle

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

In practice: Generate two N-samples

- One N-sample of X: $\left(X^i\right)_{i=1,\dots,N}$
- One N-sample of X^u: $\left(X^{u,i}\right)_{i=1,\dots,N}$

Compute the code on the sample

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

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

$$S_N^u = \frac{\frac{1}{N}\sum Y_i Y_i^u - \left(\frac{1}{N}\sum Y_i\right)\left(\frac{1}{N}\sum Y_i^u\right)}{\frac{1}{N}\sum Y_i^2 - \left(\frac{1}{N}\sum Y_i\right)^2}$$

Consistency, Central limite Theorem

$$S_N^u = \frac{\frac{1}{N}\sum Y_i Y_i^u - \left(\frac{1}{N}\sum Y_i\right)\left(\frac{1}{N}\sum Y_i^u\right)}{\frac{1}{N}\sum Y_i^2 - \left(\frac{1}{N}\sum Y_i\right)^2}, \; S^u = \frac{\mathrm{Var}\left(\mathbb{E}\left(Y|X^u\right)\right)}{\mathrm{Var}(Y)}.$$

Proposition (Janon, Klein, Lagnoux, Nodet, Prieur.)

1. *One has* $S_N^u \xrightarrow[N\to\infty]{p.s.} S^u$.
2. *If* $\mathbb{E}(Y^4) < \infty$, *then*

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

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

Application: Statistical testing

In aircraft context, the necessary amount of fuel to travel between two fixed cities is given by Bréguet formula

$$M_{fuel} = (M_{empty} + M_{pload}) \left(e^{\frac{SFC \cdot g \cdot Ra}{V \cdot F} 10^{-3}} - 1 \right) .$$

The fixed variables are

- M_{empty}: Unloaded weight
- M_{pload}: Maximal weight
- g: Gravitational constant
- Ra: Travel distance

Uncertain variables

- V: Speed during take off and landing
- F: Aeronautical coefficient
- SFC: Engine characteristic

variable	law	parameters
V	*Uniform*	$[226, 234]$
F	*Beta* (7,2)	$[18.7, 19.05]$
SFC	$\theta_2\, e^{-\theta_2(u-\theta_1)}\, 1_{[\theta_1,+\infty[}$	$\theta_1 = 17.23, \theta_2 = 3.45$

Table: Uncertainty model

The aircraft builder may ask if it is better to improve the engine (SFC) or the aerodynamic properties (F). So that we should rank the effect of F and SFC on M_{fuel}. Hence, one should test

$$H_0 : \; S^{SFC} > S^F \quad \text{or} \quad H_1 : \; S^{SFC} \leq S^F.$$

As the asymptotic distribution of an estimate of (S^{SFC}, S^F) is known, one may build a statistical decision with risk α.

Concentration inequality

CLT is a limit result. In real life, the number of experiments is finited. Concentration inequalities allow to quantify the error between the the estimate and the index true value. Using soundly Bennett inequality one get

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

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

$$\mathbb{P}\left(|S_N^{\mathbf{u}} - S^{\mathbf{u}}| \geq t\right) \leq 2 \exp\left(-\frac{NV^2}{128}\left(1 - \frac{1}{N}\right)^2 \left(\frac{t}{5 + 2t}\right)^2\right).$$

Cramér-von Mises indices-Motivation

Sobol indices are based on a variance decomposition.

- It may occur (eg symmetric function variables with identical two first moments), that the Sobol indices are not suitable to discriminate the role of the inputs
- Sobol indices *only quantify the influence around the mean*

Hence

- $\Rightarrow$ Build indices taking into account the whole distributions

Example Let X_1 et X_2 be two independent random variable with distinct distributions sharing the two first moments. Assume

$$Y = X_1 + X_2 + X_1^2 X_2^2.$$

Then

$$\begin{aligned} \mathrm{Var}\left(\mathbb{E}\left[Y|X_1\right]\right) = \mathrm{Var}(X_1 + X_1^2 \mathbb{E}\left[X_2^2\right]) &= \mathrm{Var}(X_2 + X_2^2 \mathbb{E}\left[X_1^2\right]) \\ &= \mathrm{Var}\left(\mathbb{E}\left[Y|X_2\right]\right). \end{aligned}$$

Y is a symetrical function of X_1, X_2 as X_1 and X_2 have different distribution, X_1 et X_2 should act differently. It seems important to consider sensitivity indices that take into account not only the two first moments but the whole distributions.

Indices based on Cramér-von Mises distance

Let $Z := f(X_1, \ldots, X_d) \in \mathbb{R}^k$ be the code output and F its repartition function defined for $t = (t_1, \ldots, t_k) \in \mathbb{R}^k$ by

$$F(t) = \mathbb{P}(Z \leq t) = \mathbb{E}\left[\mathbb{1}_{\{Z \leq t\}}\right] := \mathbb{E}\left[Y_t\right].$$

Let $F^{\mathbf{u}}(t)$ be the conditional repartition function (conditionally Z knowing $X_{\mathbf{u}}$):

$$F^{\mathbf{u}}(t) = \mathbb{P}(Z \leq t | X_{\mathbf{u}}) = \mathbb{E}\left[\mathbb{1}_{\{Z \leq t\}} | X_{\mathbf{u}}\right] = \mathbb{E}\left[Y_t | X_{\mathbf{u}}\right]$$

.

$$\mathbb{E}\left[F^{\mathbf{u}}(t)\right] = F(t) \Longrightarrow \mathbb{E}\left[\left(F(t) - F^{u}(t)\right)^2\right] = \mathrm{Var}\left(\mathbb{E}\left[Y_t | X_{\mathbf{u}}\right]\right)$$

The Cramér-von Mises distance between $\mathcal{L}(Z)$ and $\mathcal{L}(Z|X_u)$ is

$$D^{u}_{2,CVM} := \int_{\mathbb{R}^k} \mathbb{E}\left[\left(F(t) - F^{u}(t)\right)^2\right] dF(t). \quad (1)$$

We aim now to estimate $D^u_{2,CVM} = \mathbb{E}\left[\mathbb{E}\left[(F(Z) - F^u(Z))^2\right]\right]$ and to study the asymptotic properties of the estimator
The estimation of $D^u_{2,CVM}$ stands on a double Monte Carlo.
Design

1. Two N-sample of Z: $(Z_j^{u,1}, Z_j^{u,2})$, $1 \leq j \leq N$; (Pick-Freeze)
2. A third independent N-sample of Z: W_k, $1 \leq k \leq N$.

Monte Carlo empirical estimate of $D^u_{2,CVM}$

$$\widehat{D}^u_{2,CVM} = \frac{1}{N}\sum_{k=1}^{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}_{\{Z_j^{u,1}\leq W_k\}}\mathbb{1}_{\{Z_j^{u,2}\leq W_k\}} - \left[\frac{1}{2N}\sum_{j=1}^{N}\left(\mathbb{1}_{\{Z_j^{u,1}\leq W_k\}} + \mathbb{1}_{\{Z_j^{u,2}\leq W_k\}}\right)\right]^2\right\}.$$

One can show

Theorem (Gamboa, Klein, Lagnoux)

$\widehat{D}^u_{2,CVM}$ is strongly convergent as N goes to infinity. The sequence $\widehat{D}^u_{2,CVM}$ is asymptotically Gaussian. More precisely, $\sqrt{N}\left(\widehat{D}^u_{2,CVM} - D^u_{2,CVM}\right)$ converge in law to a centred Gaussian variable with explicit variance.