Recent advances in Global Sensibility Analysis

Clémentine Prieur

Université Grenoble Alpes

Research School on Uncertainty in Scientific Computing
ETICS2021@Erdeven, September, 12-17, 2021

Part I

Global Sensitivity Analysis:
from variance-based to more general sensitivity indices, in the framework of independent inputs and for a deterministic code.

Overview

inputs $\longrightarrow$ model $\longrightarrow$ output

$$\left.\begin{matrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_d \end{matrix}\right\} \quad \text{---}\ \mathcal{M}\ \text{---} \quad Y = \mathcal{M}(X_1, \ldots, X_d)$$

Experimental design: planification, sampling $\longleftrightarrow$ **Sensitivity analysis:** sensitivity indices' inference

Background :

$$\mathcal{M} : \begin{cases} \mathbb{R}^d & \rightarrow & \mathbb{R} \\ \mathbf{x} & \mapsto & y = \mathcal{M}(x_1, \ldots, x_d) \end{cases}$$

Goal : find how model outputs vary with inputs changes.

Different strategies :

- Qualitative analysis : non-linear behaviors? possible interactions?
  ex. : screening .
- Quantitative analysis : factorial hierarchisation, statistical tests H_0 "negligible input"
  ex. : sensitivity Sobol' indices

Sensitivity analysis may help identifying inappropriate models.

Background :

$$\mathcal{M} : \begin{cases} \mathbb{R}^d & \rightarrow & \mathbb{R} \\ \mathbf{x} & \mapsto & y = \mathcal{M}(x_1, \ldots, x_d) \end{cases}$$

Goal : find how model outputs vary with inputs changes.

Different strategies :

- Qualitative analysis : non-linear behaviors? possible interactions?
  ex. : screening .
- Quantitative analysis : factorial hierarchisation, statistical tests H_0 "negligible input"
  ex. : sensitivity Sobol' indices

Sensitivity analysis may help identifying inappropriate models.

Background :

$$\mathcal{M} : \begin{cases} \mathbb{R}^d & \to & \mathbb{R} \\ \mathbf{x} & \mapsto & y = \mathcal{M}(x_1, \ldots, x_d) \end{cases}$$

Goal : find how model outputs vary with inputs changes.

Different strategies :

- Qualitative analysis : non-linear behaviors? possible interactions?
  ex. : screening .
- Quantitative analysis : factorial hierarchisation, statistical tests H_0 "negligible input"
  ex. : sensitivity Sobol' indices

Sensitivity analysis may help identifying inappropriate models.

Various approaches for quantitative sensitivity :

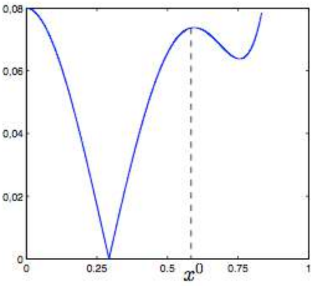
Local approaches :

$\mathcal{M}(\mathbf{x}) \approx \mathcal{M}(\mathbf{x^0}) + \sum_{i=1}^{d} \left(\frac{\partial \mathcal{M}}{\partial x_i}\right)_{\mathbf{x^0}} (x_i - x_i^0)$ (Taylor approximation).

First order sensitivity index for input i : $\left(\frac{\partial \mathcal{M}}{\partial x_i}\right)_{\mathbf{x^0}}$.

Pros : Low computational cost even for large d

Cons : local approaches, not well-suited for highly nonlinear models



Global approaches :

From expert knowledge or observations, we attribute a probability law to the inputs vector.

ex.: If independent inputs, then only margins are needed.

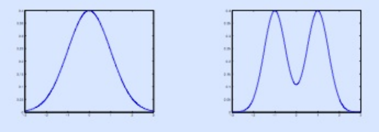


Figure: law (left) unimodal , (right) bimodal

Introduction

We vary inputs w.r.t. their probability distribution.

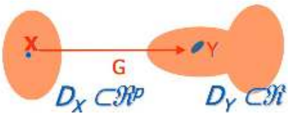


Figure: Local versus Global ($G := \mathcal{M}$)

We vary input w.r.t. their probability law

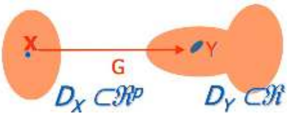


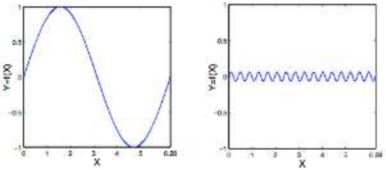
Figure: Local versus Global ($G := \mathcal{M}$), illustration.

"Globalized" local approaches : e.g. (1) $\mathbb{E}_X \left[\frac{\partial \mathcal{M}}{\partial x_i} \Big|_{\mathbf{X}} \right]$, ou (2) $\mathbb{E}_X \left[\left(\frac{\partial \mathcal{M}}{\partial x_i} \Big|_{\mathbf{X}} \right)^2 \right]$.

Avantages : particularly interesting if adjoint available

Cons :

(1) does not discriminate enough



(2) is known as Derivative-based Global Sensitivity Measures, see Sobol' & Gresham (1995), Sobol' & Kucherenko (2009). This index is more adapted for screening than for hierarchization (e.g. Lamboni *et al.*, 2013).

Screening tools based on gradients have been developed recently (see, e.g., references in Da Veiga *et al.* [Chapter 2], 2021).

The present lecture targets global approaches that allow to efficiently rank input factors.

However, let us provide, as an introduction, a first outlook to screening most usual methods.

A quick overlook on screening methods

Main objective : to screen among a large amount of inputs which ones are non influential on the quantity of interest (QoI).

Advantages : moderate computational cost.

Drawbacks : partial information, no hierarchisation.

A OAT screening method : Morris, 1991

OAT One At a Time we vary the factors one by one.
The screening method proposed by Morris is a global OAT approach.

Model $Y = \mathcal{M}(\mathbf{X})$, $\mathbf{X} = (X_1, \ldots, X_d)$ *with the* X_i*s independent uniform random variables on* $[0, 1]$.

More details on the method :

- input discretization on a grid with p values $\left\{0, \frac{1}{p-1}, \ldots, 1\right\}$.
- Δ a multiple of $1/(p-1)$, fixed once for all.
- $\Omega := \left\{0, \frac{1}{p-1}, \ldots, 1\right\}^d$.
- $\Omega_i^\Delta := \{x \in \Omega \text{ such that } (x_1, \ldots, x_{i-1}, x_i + \Delta, x_{i+1}, \ldots, x_d) \in \Omega\}$.

Definition

Elementary effect of X_i computed at $\mathbf{x} \in \Omega_i^\Delta$,

$$d_i(\mathbf{x}) = \frac{1}{\Delta} \left\{\mathcal{M}(x_1, \ldots, x_{i-1}, x_i + \Delta, x_{i+1}, \ldots, x_d) - \mathcal{M}(\mathbf{x})\right\} .$$

There are $p^{d-1}(p - \Delta(p-1))$ elementary effects to compute.

Steps :

- one draws uniformly a r-sample in Ω_i^Δ : $\mathbf{x}^1, \ldots, \mathbf{x}^r$;
- one computes $d_i(\mathbf{x}^j)$, $j = 1, \ldots, r$, $i = 1, \ldots, d$;
- one computes

$$\begin{cases} \mu_i &= \frac{1}{r} \sum_{j=1} d_i(\mathbf{x}^j) \\ \sigma_i^2 &= \frac{1}{r} \sum_{j=1}^r (d_i(\mathbf{x}^j) - \mu_i)^2 \,. \end{cases}$$

	σ_i^2 low	σ_i^2 high
$\lvert\mu_i\rvert$ low	non influential	nonlinearities and/or interactions
$\lvert\mu_i\rvert$ high	influential	nonlinearities and/or interactions

The efficiency of the method "number of elementary effects computed / number of model runs" is equal to $1/2$.

Morris (1991) presents an adaptation with an efficiency equal to $d/(d+1)$, with d the input space dimension.

A toy example

Advection-reaction-diffusion equation with Dirichlet boundary condition :

$$\begin{cases} \dfrac{\partial u}{\partial t} = -r.u - a\dfrac{\partial u}{\partial x} + \lambda\dfrac{\partial^2 u}{\partial x^2} + f \quad x \in [0, L],\ t \in [0, T] \\ u(x = 0, t) = \Psi_1(t) \quad t \in [0, T] \\ u(x = L, t) = \Psi_2(t) \quad t \in [0, T] \\ u(x, t = 0) = g(x) \quad x \in (0, L)\,. \end{cases}$$

A : energy norm of the solution at time $t = T$.

Sensitivity of A with respect to (a, r, λ)? Uncertain input parameters are modeled as $a, r \sim \mathcal{U}([0.4, 0.6])$, $\lambda \sim \mathcal{U}([0.04, 0.06])$.

Scheme : 2-stemps Adams-Moulton, sample size equals 2^{13}.

Sensitivity measures based on variance : $S_a = 0.0188$, $S_\lambda = 0.7299$, $S_r = 0.2488$, $S_a + S_\lambda + S_r = 0.988$.

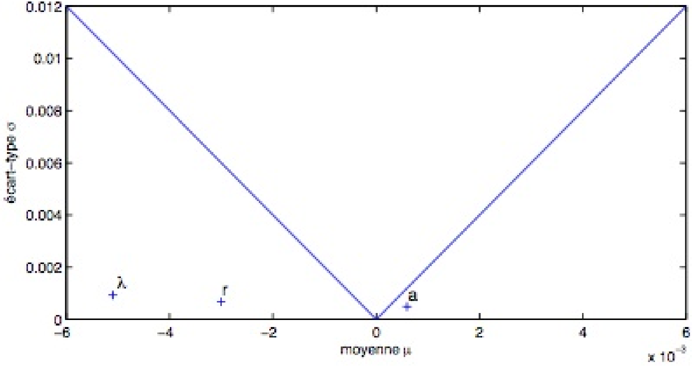


Figure: Morris with $p = 50$, $\Delta = 25/49$.

see jupyter Jupyter notebook Premiers–Pas.

Lecture outline

I- Functional variance analysis

II- Sobol' index inference

- Monte Carlo estimators
- Given data estimators
- Spectral estimators

III- Indices based on the Cramér-von-Mises distance

IV- Towards general metric space indices

V- Pick-freeze estimation procedure for Cramér-von Mises indices

VI- Indices "à la Borgonovo"

VII- Kernel based ANOVA decomposition

I- Functional variance analysis

General setup : (Hoeffding, 1948; Sobol', 1993)

$Y = \mathcal{M}(X_1, \ldots, X_d)$, $(X_1, \ldots, X_d) \sim P_{X_1,\ldots,X_d}$. In the following, we assume :

i) the X_i are independent ;

ii) $\forall\, i = 1, \ldots, d$, $X_i \sim \mathcal{U}([0, 1])$.

Assumption ii) is not restrictive : with the inverse technique, $Y = \mathcal{M}(X_1, \ldots, X_d)$ can be written as

$$Y = \mathcal{M}(F_{X_1}^{-1}(U_1), \ldots, F_{X_d}^{-1}(U_d)) = \widetilde{\mathcal{M}}(U_1, \ldots, U_d)$$

with $U_i, i = 1, \ldots d$ independent and for all i, $U_i \sim \mathcal{U}([0, 1])$, $F_{X_i}^{-1}$ inverse of the cumulative distribution function of X_i.

The complex case of correlated inputs will be mentioned at the end of this lecture and in the lecture on Shapley effects.

General setup : (Hoeffding, 1948; Sobol', 1993)

$Y = \mathcal{M}(X_1, \ldots, X_d)$, $(X_1, \ldots, X_d) \sim P_{X_1,\ldots,X_d}$. In the following, we assume :

i) the X_i are independent ;

ii) $\forall\, i = 1, \ldots, d$, $X_i \sim \mathcal{U}([0,1])$.

Assumption ii) is not restrictive : with the inverse technique, $Y = \mathcal{M}(X_1, \ldots, X_d)$ can be written as

$$Y = \mathcal{M}(F_{X_1}^{-1}(U_1), \ldots, F_{X_d}^{-1}(U_d)) = \widetilde{\mathcal{M}}(U_1, \ldots, U_d)$$

with $U_i, i = 1, \ldots d$ independent and for all i, $U_i \sim \mathcal{U}([0,1])$, $F_{X_i}^{-1}$ inverse of the cumulative distribution function of X_i.

The complex case of correlated inputs will be mentioned at the end of this lecture and in the lecture on Shapley effects.

General setup : (Hoeffding, 1948; Sobol', 1993)

$Y = \mathcal{M}(X_1, \ldots, X_d)$, $(X_1, \ldots, X_d) \sim P_{X_1,\ldots,X_d}$. In the following, we assume :

i) the X_i are independent ;

ii) $\forall\, i = 1, \ldots, d$, $X_i \sim \mathcal{U}([0,1])$.

Assumption ii) is not restrictive : with the inverse technique,
$Y = \mathcal{M}(X_1, \ldots, X_d)$ can be written as

$$Y = \mathcal{M}(F_{X_1}^{-1}(U_1), \ldots, F_{X_d}^{-1}(U_d)) = \widetilde{\mathcal{M}}(U_1, \ldots, U_d)$$

with $U_i, i = 1, \ldots d$ independent and for all i, $U_i \sim \mathcal{U}([0,1])$, $F_{X_i}^{-1}$ inverse of the cumulative distribution function of X_i.

The complex case of correlated inputs will be mentioned at the end of this lecture and in the lecture on Shapley effects.

Towards Sobol sensitivity indices

Is the output Y more or less variable when input are fixed?

$\mathrm{Var}\,(Y|X_i = x_i)$, how to choose x_i? $\Rightarrow E\,[V\,(Y|X_i)]$

the smaller this quantity, (i.e. fixing X_i), the smaller is the variance of Y when fixing the ith input: variable X_i has a strong impact.

Theorem (Total variance)

$\mathrm{Var}(Y) = \mathrm{Var}\,[\mathbb{E}\,(Y|X_i)] + \mathbb{E}\,[\mathrm{Var}\,(Y|X_i)].$

Definition (First order Sobol' Index)

$i = 1, \ldots, d$

$$0 \leq S_i = \frac{V\,[E\,(Y|X_i)]}{\mathrm{Var}(Y)} \leq 1$$

ex. : linear output $Y = \sum_{i=1}^{d} \beta_i X_i$, we get $S_i = \frac{\beta_i^2 \mathrm{Var}(X_i)}{\mathrm{Var}(Y)} = \rho_i^2$, with ρ_i linear correlation coefficient.

Towards Sobol sensitivity indices

Is the output Y more or less variable when input are fixed?

$\mathrm{Var}\left(Y|X_i = x_i\right)$, how to choose x_i? $\Rightarrow E\left[V\left(Y|X_i\right)\right]$

the smaller this quantity, (i.e. fixing X_i), the smaller is the variance of Y when fixing the ith input: variable X_i has a strong impact.

Theorem (Total variance)

$\mathrm{Var}(Y) = \mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right] + \mathbb{E}\left[\mathrm{Var}\left(Y|X_i\right)\right].$

Definition (First order Sobol' Index)

$i = 1, \ldots, d$

$$0 \leq S_i = \frac{V\left[E\left(Y|X_i\right)\right]}{\mathrm{Var}(Y)} \leq 1$$

ex. : linear output $Y = \sum_{i=1}^{d} \beta_i X_i$, we get $S_i = \frac{\beta_i^2 \mathrm{Var}(X_i)}{\mathrm{Var}(Y)} = \rho_i^2$, with ρ_i linear correlation coefficient.

Towards Sobol sensitivity indices

Is the output Y more or less variable when input are fixed?

$\mathrm{Var}\,(Y|X_i = x_i)$, how to choose x_i? $\Rightarrow E\,[V\,(Y|X_i)]$

the smaller this quantity, (i.e. fixing X_i), the smaller is the variance of Y when fixing the ith input: variable X_i has a strong impact.

Theorem (Total variance)

$\mathrm{Var}(Y) = \mathrm{Var}\,[\mathbb{E}\,(Y|X_i)] + \mathbb{E}\,[\mathrm{Var}\,(Y|X_i)]$.

Definition (First order Sobol' Index)

$i = 1, \ldots, d$

$$0 \le S_i = \frac{V\,[E\,(Y|X_i)]}{\mathrm{Var}(Y)} \le 1$$

ex. : linear output $Y = \sum_{i=1}^{d} \beta_i X_i$, we get $S_i = \frac{\beta_i^2 \mathrm{Var}(X_i)}{\mathrm{Var}(Y)} = \rho_i^2$, with ρ_i linear correlation coefficient.

Towards Sobol sensitivity indices

Is the output Y more or less variable when input are fixed?

$\mathrm{Var}\,(Y|X_i = x_i)$, how to choose x_i? $\Rightarrow E\,[V\,(Y|X_i)]$

the smaller this quantity, (i.e. fixing X_i), the smaller is the variance of Y when fixing the ith input: variable X_i has a strong impact.

Theorem (Total variance)

$\mathrm{Var}(Y) = \mathrm{Var}\,[\mathbb{E}\,(Y|X_i)] + \mathbb{E}\,[\mathrm{Var}\,(Y|X_i)].$

Definition (First order Sobol' Index)

$i = 1, \ldots, d$

$$0 \leq S_i = \frac{V\,[E\,(Y|X_i)]}{\mathrm{Var}(Y)} \leq 1$$

ex. : linear output $Y = \sum_{i=1}^{d} \beta_i X_i$, we get $S_i = \frac{\beta_i^2 \mathrm{Var}(X_i)}{\mathrm{Var}(Y)} = \rho_i^2$, with ρ_i linear correlation coefficient.

Towards Sobol sensitivity indices

Is the output Y more or less variable when input are fixed?

$\text{Var}(Y|X_i = x_i)$, how to choose x_i? $\Rightarrow E[V(Y|X_i)]$

the smaller this quantity, (i.e. fixing X_i), the smaller is the variance of Y when fixing the ith input: variable X_i has a strong impact.

Theorem (Total variance)

$\text{Var}(Y) = \text{Var}[\mathbb{E}(Y|X_i)] + \mathbb{E}[\text{Var}(Y|X_i)]$.

Definition (First order Sobol' Index)

$i = 1, \ldots, d$

$$0 \leq S_i = \frac{V[E(Y|X_i)]}{\text{Var}(Y)} \leq 1$$

ex. : linear output $Y = \sum_{i=1}^{d} \beta_i X_i$, we get $S_i = \frac{\beta_i^2 \text{Var}(X_i)}{\text{Var}(Y)} = \rho_i^2$, with ρ_i linear correlation coefficient.

Towards Sobol sensitivity indices

Is the output Y more or less variable when input are fixed?

$\mathrm{Var}\,(Y|X_i = x_i)$, how to choose x_i? $\Rightarrow E\,[V\,(Y|X_i)]$

the smaller this quantity, (i.e. fixing X_i), the smaller is the variance of Y when fixing the ith input: variable X_i has a strong impact.

Theorem (Total variance)

$\mathrm{Var}(Y) = \mathrm{Var}\,[\mathbb{E}\,(Y|X_i)] + \mathbb{E}\,[\mathrm{Var}\,(Y|X_i)].$

Definition (First order Sobol' Index)

$i = 1, \ldots, d$

$$0 \leq S_i = \frac{V\,[E\,(Y|X_i)]}{\mathrm{Var}(Y)} \leq 1$$

ex. : linear output $Y = \sum_{i=1}^{d} \beta_i X_i$, we get $S_i = \frac{\beta_i^2 \mathrm{Var}(X_i)}{\mathrm{Var}(Y)} = \rho_i^2$, with ρ_i linear correlation coefficient.

Toy case:

$$Y = X_1^2 + X_2 \quad X_i \sim \mathcal{U}([0,1]) \quad X_1 \perp\!\!\!\perp X_2$$

$$\mathbb{E}(Y|X_1) = X_1^2 + \mathbb{E}(X_2) \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_1)] = \mathrm{Var}(X_1^2) = \frac{4}{45}$$

$$\mathbb{E}(Y|X_2) = \mathbb{E}(X_1^2) + X_2 \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_2)] = \mathrm{Var}(X_2) = \frac{1}{12}$$

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1^2) + \mathrm{Var}(X_2) = \frac{31}{180}$$

$$S_1 = \frac{16}{31} \approx 0,516\,,\ S_2 = \frac{15}{31} \approx 0,484$$

$S_1 + S_2 = 1$, additive model

I- Functional variance analysis

Toy case:

$$Y = X_1^2 + X_2 \qquad X_i \sim \mathcal{U}([0,1]) \quad X_1 \perp\!\!\!\perp X_2$$

$\mathbb{E}(Y|X_1) = X_1^2 + \mathbb{E}(X_2) \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_1)] = \mathrm{Var}(X_1^2) = \frac{4}{45}$

$\mathbb{E}(Y|X_2) = \mathbb{E}(X_1^2) + X_2 \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_2)] = \mathrm{Var}(X_2) = \frac{1}{12}$

$\mathrm{Var}(Y) = \mathrm{Var}(X_1^2) + \mathrm{Var}(X_2) = \frac{31}{180}$

$$S_1 = \frac{16}{31} \approx 0,516\,, \; S_2 = \frac{15}{31} \approx 0,484$$

$S_1 + S_2 = 1$, additive model

Toy case:

$$Y = X_1^2 + X_2 \quad X_i \sim \mathcal{U}([0,1]) \quad X_1 \perp\!\!\!\perp X_2$$

$\mathbb{E}(Y|X_1) = X_1^2 + \mathbb{E}(X_2) \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_1)] = \mathrm{Var}(X_1^2) = \frac{4}{45}$

$\mathbb{E}(Y|X_2) = \mathbb{E}(X_1^2) + X_2 \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_2)] = \mathrm{Var}(X_2) = \frac{1}{12}$

$\mathrm{Var}(Y) = \mathrm{Var}(X_1^2) + \mathrm{Var}(X_2) = \frac{31}{180}$

$$S_1 = \frac{16}{31} \approx 0,516\,, \; S_2 = \frac{15}{31} \approx 0,484$$

$S_1 + S_2 = 1$, additive model

Toy case:

$$Y = X_1^2 + X_2 \quad X_i \sim \mathcal{U}([0,1]) \quad X_1 \perp\!\!\!\perp X_2$$

$\mathbb{E}(Y|X_1) = X_1^2 + \mathbb{E}(X_2) \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_1)] = \mathrm{Var}(X_1^2) = \frac{4}{45}$

$\mathbb{E}(Y|X_2) = \mathbb{E}(X_1^2) + X_2 \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_2)] = \mathrm{Var}(X_2) = \frac{1}{12}$

$\mathrm{Var}(Y) = \mathrm{Var}(X_1^2) + \mathrm{Var}(X_2) = \frac{31}{180}$

$$S_1 = \frac{16}{31} \approx 0,516\,, \quad S_2 = \frac{15}{31} \approx 0,484$$

$S_1 + S_2 = 1$, additive model

Toy case:

$$Y = X_1^2 + X_2 \quad X_i \sim \mathcal{U}([0,1]) \quad X_1 \perp\!\!\!\perp X_2$$

$$\mathbb{E}(Y|X_1) = X_1^2 + \mathbb{E}(X_2) \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_1)] = \mathrm{Var}(X_1^2) = \frac{4}{45}$$

$$\mathbb{E}(Y|X_2) = \mathbb{E}(X_1^2) + X_2 \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_2)] = \mathrm{Var}(X_2) = \frac{1}{12}$$

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1^2) + \mathrm{Var}(X_2) = \frac{31}{180}$$

$$S_1 = \frac{16}{31} \approx 0,516\,,\ S_2 = \frac{15}{31} \approx 0,484$$

$S_1 + S_2 = 1$, additive model

Toy case:

$$Y = X_1^2 + X_2 \qquad X_i \sim \mathcal{U}([0,1]) \quad X_1 \perp\!\!\!\perp X_2$$

$$\mathbb{E}(Y|X_1) = X_1^2 + \mathbb{E}(X_2) \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_1)] = \mathrm{Var}(X_1^2) = \tfrac{4}{45}$$

$$\mathbb{E}(Y|X_2) = \mathbb{E}(X_1^2) + X_2 \Rightarrow \mathrm{Var}[\mathbb{E}(Y|X_2)] = \mathrm{Var}(X_2) = \tfrac{1}{12}$$

$$\mathrm{Var}(Y) = \mathrm{Var}(X_1^2) + \mathrm{Var}(X_2) = \tfrac{31}{180}$$

$$S_1 = \frac{16}{31} \approx 0,516\,,\; S_2 = \frac{15}{31} \approx 0,484$$

$S_1 + S_2 = 1$, additive model

More generally,

Theorem (Hoeffding decomposition)

$\mathcal{M} : [0,1]^d \to \mathbb{R}$, $\int_{[0,1]^d} \mathcal{M}^2(x)dx < \infty$

$\mathcal{M}$ has an unique decomposition

$\mathcal{M}_0 + \sum_{i=1}^d \mathcal{M}_i(x_i) + \sum_{1 \leq i < j \leq d} \mathcal{M}_{i,j}(x_i, x_j) + \ldots + \mathcal{M}_{1,\ldots,d}(x_1, \ldots, x_d)$

under the constraint

- *$\mathcal{M}_0$ constant,*
- $\forall\, 1 \leq s \leq d$, $\forall\, 1 \leq i_1 < \ldots < i_s \leq d$, $\forall\, 1 \leq p \leq s$

$$\int_0^1 \mathcal{M}_{i_1,\ldots,i_s}(x_{i_1}, \ldots, x_{i_s}) dx_{i_p} = 0$$

Consequences : $\mathcal{M}_0 = \int_{[0,1]^d} \mathcal{M}(x)dx$ and the terms of the decomposition are orthogonal.

The computation of each term in the decomposition writes:

- $\mathcal{M}_i(x_i) = \int_{[0,1]^{d-1}} \mathcal{M}(x)\,\Pi_{p\neq i}dx_p - \mathcal{M}_0$
- $i \neq j$ $\mathcal{M}_{i,j}(x_i, x_j) = \int_{[0,1]^{d-2}} \mathcal{M}(x)\,\Pi_{p\neq i,j}dx_p - \mathcal{M}_0 - \mathcal{M}_i(x_i) - \mathcal{M}_j(x_j)$
- ...

$\Rightarrow$ computation of multiple integrals.

Consequences : $\mathcal{M}_0 = \int_{[0,1]^d} \mathcal{M}(x)dx$ and the terms of the decomposition are orthogonal.

The computation of each term in the decomposition writes:

- $\mathcal{M}_i(x_i) = \int_{[0,1]^{d-1}} \mathcal{M}(x) \Pi_{p \neq i} dx_p - \mathcal{M}_0$
- $i \neq j$ $\mathcal{M}_{i,j}(x_i, x_j) = \int_{[0,1]^{d-2}} \mathcal{M}(x) \Pi_{p \neq i,j} dx_p - \mathcal{M}_0 - \mathcal{M}_i(x_i) - \mathcal{M}_j(x_j)$
- ...

$\Rightarrow$ computation of multiple integrals.

Consequences : $\mathcal{M}_0 = \int_{[0,1]^d} \mathcal{M}(x) dx$ and the terms of the decomposition are orthogonal.

The computation of each term in the decomposition writes:

- $\mathcal{M}_i(x_i) = \int_{[0,1]^{d-1}} \mathcal{M}(x) \Pi_{p \neq i} dx_p - \mathcal{M}_0$
- $i \neq j$ $\mathcal{M}_{i,j}(x_i, x_j) = \int_{[0,1]^{d-2}} \mathcal{M}(x) \Pi_{p \neq i,j} dx_p - \mathcal{M}_0 - \mathcal{M}_i(x_i) - \mathcal{M}_j(x_j)$
- $\ldots$

$\Rightarrow$ computation of multiple integrals.

Consequences : $\mathcal{M}_0 = \int_{[0,1]^d} \mathcal{M}(x) dx$ and the terms of the decomposition are orthogonal.

The computation of each term in the decomposition writes:

- $\mathcal{M}_i(x_i) = \int_{[0,1]^{d-1}} \mathcal{M}(x) \Pi_{p \neq i} dx_p - \mathcal{M}_0$
- $i \neq j$ $\mathcal{M}_{i,j}(x_i, x_j) = \int_{[0,1]^{d-2}} \mathcal{M}(x) \Pi_{p \neq i,j} dx_p - \mathcal{M}_0 - \mathcal{M}_i(x_i) - \mathcal{M}_j(x_j)$
- $\ldots$

$\Rightarrow$ computation of multiple integrals.

Variance decomposition : $X_1, \ldots, X_d$ i.i.d. $\sim \mathcal{U}([0,1])$

$$Y = \mathcal{M}(X) = \mathcal{M}_0 + \sum_{i=1}^d \mathcal{M}_i(X_i) + \ldots + \mathcal{M}_{1,\ldots,d}(X_1, \ldots, X_d)$$

- $\mathcal{M}_0 = \mathbb{E}(Y)$,
- $\mathcal{M}_i(X_i) = \mathbb{E}(Y|X_i) - \mathbb{E}(Y)$,
- $i \neq j$ $\mathcal{M}_{i,j}(X_i, X_j) = \mathbb{E}(Y|X_i, X_j) - \mathbb{E}(Y|X_i) - \mathbb{E}(Y|X_j) + \mathbb{E}(Y)$,
- $\ldots$

$$\mathrm{Var}(Y) = \sum_{i=1}^d \mathrm{Var}(\mathcal{M}_i(X_i)) + \ldots + \mathrm{Var}(\mathcal{M}_{1,\ldots,d}(X_1, \ldots, X_d))$$

Variance decomposition : $X_1, \ldots, X_d$ i.i.d. $\sim \mathcal{U}([0,1])$

$Y = \mathcal{M}(X) = \mathcal{M}_0 + \sum_{i=1}^{d} \mathcal{M}_i(X_i) + \ldots + \mathcal{M}_{1,\ldots,d}(X_1, \ldots, X_d)$

- $\mathcal{M}_0 = \mathbb{E}(Y)$,
- $\mathcal{M}_i(X_i) = \mathbb{E}(Y|X_i) - \mathbb{E}(Y)$,
- $i \neq j$ $\mathcal{M}_{i,j}(X_i, X_j) = \mathbb{E}(Y|X_i, X_j) - \mathbb{E}(Y|X_i) - \mathbb{E}(Y|X_j) + \mathbb{E}(Y)$,
- $\ldots$

$\mathrm{Var}(Y) = \sum_{i=1}^{d} \mathrm{Var}(\mathcal{M}_i(X_i)) + \ldots + \mathrm{Var}(\mathcal{M}_{1,\ldots,d}(X_1, \ldots, X_d))$

Definition (Sobol' indices)

$$\forall\, i = 1, \ldots, d \;\; S_i = \frac{\mathrm{Var}\left(\mathcal{M}_i(X_i)\right)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right]}{\mathrm{Var}(Y)}$$

$\forall\, i \neq j$

$$S_{i,j} = \frac{\mathrm{Var}\left(\mathcal{M}_{i,j}(X_i, X_j)\right)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i, X_j\right)\right] - \mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right] - \mathrm{Var}\left[\mathbb{E}\left(Y|X_j\right)\right]}{\mathrm{Var}(Y)}$$

$\ldots$

$$1 = \sum_{i=1}^{d} S_i + \sum_{i \neq j} S_{i,j} + \ldots + S_{1,\ldots,d}$$

Definition (Total indices)

$$i = 1, \ldots, d \qquad S_{T_i} = \sum_{\mathbf{u} \subseteq \{1,\ldots,d\}\,,\; \mathbf{u} \neq \emptyset\,,\; i \in \mathbf{u}} S_{\mathbf{u}}\,.$$

Definition (Sobol' indices)

$$\forall\, i = 1, \ldots, d \;\; S_i = \frac{\mathrm{Var}\left(\mathcal{M}_i(X_i)\right)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right]}{\mathrm{Var}(Y)}$$

$\forall\, i \neq j$

$$S_{i,j} = \frac{\mathrm{Var}\left(\mathcal{M}_{i,j}(X_i, X_j)\right)}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_i, X_j\right)\right] - \mathrm{Var}\left[\mathbb{E}\left(Y|X_i\right)\right] - \mathrm{Var}\left[\mathbb{E}\left(Y|X_j\right)\right]}{\mathrm{Var}(Y)}$$

$\ldots$

$$1 = \sum_{i=1}^{d} S_i + \sum_{i \neq j} S_{i,j} + \ldots + S_{1,\ldots,d}$$

Definition (Total indices)

$$i = 1, \ldots, d \qquad S_{T_i} = \sum_{\mathbf{u} \subseteq \{1,\ldots,d\}\,,\; \mathbf{u} \neq \emptyset\,,\; i \in \mathbf{u}} S_{\mathbf{u}}\;.$$

Sobol' indices :

Definition (Total indices)

$$i = 1, \ldots, d \qquad S_{T_i} = \sum_{\mathbf{u} \subseteq \{1,\ldots,d\}\,,\ \mathbf{u} \neq \emptyset\,,\ i \in \mathbf{u}} S_{\mathbf{u}}\ .$$

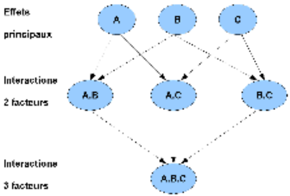
$X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$

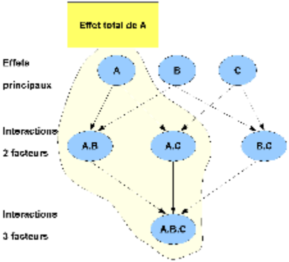
Using the theorem of the total variance,

$$S_{T_i} = \frac{\mathbb{E}\left[\mathrm{Var}\left(Y|X_{-i}\right)\right]}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}\left[\mathbb{E}\left(Y|X_{-i}\right)\right]}{\mathrm{Var}(Y)}\ .$$

I- Functional variance analysis

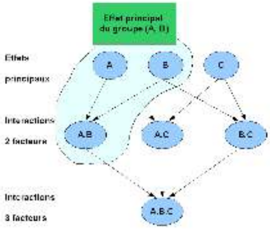
Indices with factor:

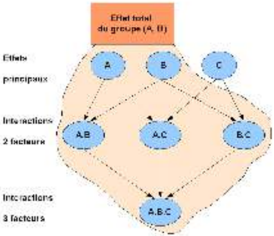




I- Functional variance analysis

Indices with groupe of factors:





Fact : Analytical expressions of Sobol' indices, with integrals in high dimensional spaces, are rarely available.

We present different approaches

II.1- Monte Carlo estimators (hypothesis $\mathbb{L}^2$ with the model);

II.2- Given data estimators (under mild regularity assumptions on the model);

II.3- Spectral estimators (additional hypotheses of regularity);

II.4- Conclusion on Sobol' index inference.

If the model is too costly to assess, we fit a metamodel before applying these techniques.

ex.: parametric and non-parametric regressions, Gaussian metamodel. . .

Fact : Analytical expressions of Sobol' indices, with integrals in high dimensional spaces, are rarely available.

We present different approaches

- II.1- Monte Carlo estimators (hypothesis $\mathbb{L}^2$ with the model);
- II.2- Given data estimators (under mild regularity assumptions on the model);
- II.3- Spectral estimators (additional hypotheses of regularity);
- II.4- Conclusion on Sobol' index inference.

If the model is too costly to assess, we fit a metamodel before applying these techniques.

ex.: parametric and non-parametric regressions, Gaussian metamodel...

Fact : Analytical expressions of Sobol' indices, with integrals in high dimensional spaces, are rarely available.

We present different approaches

II.1- Monte Carlo estimators (hypothesis $\mathbb{L}^2$ with the model);

II.2- Given data estimators (under mild regularity assumptions on the model);

II.3- Spectral estimators (additional hypotheses of regularity);

II.4- Conclusion on Sobol' index inference.

If the model is too costly to assess, we fit a metamodel before applying these techniques.

ex.: parametric and non-parametric regressions, Gaussian metamodel...

II.1- Monte Carlo based Sobol' index inference

Monte-Carlo type Approaches : (Sobol' 93, Saltelli 02, Mauntz, ...)

Idea : X'_{-i} indep. copy of X_{-i}, $Y = \mathcal{M}(X_i, X_{-i})$, $Y^i = \mathcal{M}(X_i, X'_{-i})$

We have $S_i = \frac{\mathrm{Cov}(Y, Y^i)}{\mathrm{Var}(Y)}$, the idea is based on empirical formulas.

Two independent samples A and B (Monte-Carlo, LHS)

$$A = \begin{pmatrix} x^A_{1,1} & \cdots & x^A_{d,1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x^A_{1,n} & \ldots & x^A_{d,n} \end{pmatrix} \qquad B = \begin{pmatrix} x^B_{1,1} & \cdots & x^B_{d,1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x^B_{1,n} & \ldots & x^B_{d,n} \end{pmatrix}$$

From A and of B, we create d sampling matrices C_i, $i = 1, \ldots, d$.

$$C_i = \begin{pmatrix} x^A_{1,1} & \ldots & x^B_{i,1} & \ldots & x^A_{d,1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ x^A_{1,n} & \ldots & x^B_{i,n} & \ldots & x^A_{d,n} \end{pmatrix}$$

We compute $(1 + d) \times n$ the model $\mathcal{M}$:

$$y^B = \begin{pmatrix} y_1^B \\ \cdot \\ \cdot \\ \cdot \\ y_n^B \end{pmatrix} \quad \text{and} \quad \forall\, 1 \leq i \leq d \quad y^{C_i} = \begin{pmatrix} y_1^{C_i} \\ \cdot \\ \cdot \\ \cdot \\ y_n^{C_i} \end{pmatrix}$$

`sobolEff()` (Janon *et al.*, 2014 & 2016)

- $\hat{V}_i = \frac{1}{n}\sum_{k=1}^{n} y_k^B y_k^{C_i} - \left(\frac{1}{n}\sum_{k=1}^{n} \frac{y_k^B + y_k^{C_i}}{2}\right)^2$ numerator of the first-order index
- $\hat{V} = \frac{1}{n}\sum_{k=1}^{n} \frac{(y_k^B)^2 + (y_k^{C_i})^2}{2} - \left(\frac{1}{n}\sum_{k=1}^{n} \frac{y_k^B + y_k^{C_i}}{2}\right)^2$ denominator

This type of estimators is known as pick-freeze estimators.

Remarks:

Pick-freeze estimators can be defined for any subset $\mathbf{u} \subseteq \{1, \ldots, d\}$.

In practice, we can replace MC or LHS samplings by QMC (hyp. of regular variations).

What about the statistical properties of pick-freeze estimators?

- Is it consistent? yes, proof by using the Strong Law of Large Numbers.
- If yes, at which rate of convergence? yes, CLT (cv in $\sqrt{n}$).
- Is it asymptotically efficient? yes.
- Is it possible to measure its performance for a fixed n?
  yes, Berry-Esseen and/or concentration inequalities.

see, Janon *et al.* (2014,2016) or Gamboa *et al.* (2014)

As an example, let us state in the next slide a central limit theorem. From such a CLT, one can also deduce asymptotic confidence intervals or hypothesis testing, e.g., on the nullity of Sobol' index associated to $\mathbf{u} \subseteq \{1, \ldots, d\}$.

II.1- Monte Carlo based Sobol' index inference

$$\widehat{S_{\mathbf{u}}^{\text{clo}}} = \frac{\frac{1}{n}\sum_{k=1}^{n} Y_k^B Y_k^{C_{\mathbf{u}}} - \left(\frac{1}{n}\sum_{k=1}^{n} \frac{Y_k^B + Y_k^{C_{\mathbf{u}}}}{2}\right)^2}{\frac{1}{n}\sum_{k=1}^{n} \frac{(Y_k^B)^2 + (Y_k^{C_{\mathbf{u}}})^2}{2} - \left(\frac{1}{n}\sum_{k=1}^{n} \frac{Y_k^B + Y_k^{C_{\mathbf{u}}}}{2}\right)^2}, \quad S_{\mathbf{u}}^{\text{clo}} = \frac{\operatorname{Var}\left[\mathbb{E}\left(Y|\mathbf{X}_{\mathbf{u}}\right)\right]}{\operatorname{Var}[Y]}.$$

Theorem (Janon *et al.*, 2014)

1. *One has* $\widehat{S_{\mathbf{u}}^{\text{clo}}} \xrightarrow[n\to\infty]{a.s.} S_{\mathbf{u}}^{\text{clo}}$.
2. *If* $\mathbb{E}\left(Y^4\right) < \infty$, *then*

$$\sqrt{n}\left(\widehat{S}_{\mathbf{u}}^{\text{clo}} - S_{\mathbf{u}}^{\text{clo}}\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma_{\mathbf{u}}^2\right)$$

with $\sigma_{\mathbf{u}}^2 = \dfrac{\operatorname{Var}\left[(Y-\mathbb{E}(Y))(Y_{\mathbf{u}}-\mathbb{E}(Y)) - \frac{S_{\mathbf{u}}^{\text{clo}}}{2}\left((Y-\mathbb{E}(Y))^2 + (Y_{\mathbf{u}}-\mathbb{E}(Y))^2\right)\right]}{(\operatorname{Var}[Y])^2}$.

Using Bennett's concentration inequality, one gets for fixed sample size n:

Proposition (Janon *et al.*, 2016; Gamboa *et al.*, 2014)

Let $\mathbf{u}$ *be a subset of* $\{1, \ldots, d\}$*. Let* $b > 0$ *and* $t > 0$*. Let* $Y \in [-b, b]$*. Then,*

$$\mathbb{P}\left(\widehat{S_{\mathbf{u}}^{\text{clo}}} \geqslant S_{\mathbf{u}}^{\text{clo}} + t\right) \leqslant \exp\left(-\frac{n\mathrm{Var}[Y]^2}{128}\left(1 - \frac{1}{n}\right)^2\left(\frac{t}{1+t}\right)^2\right).$$

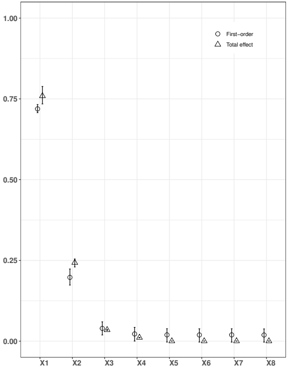
Assume further that $\frac{9}{8n} \leq t \leq 1$*, then*

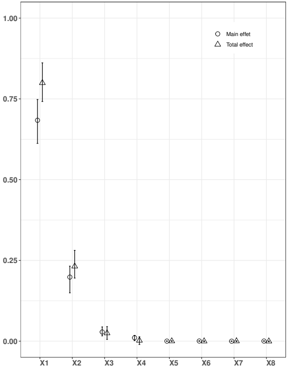
$$\mathbb{P}\left(\widehat{S_{\mathbf{u}}^{\text{clo}}} \leqslant S_{\mathbf{u}}^{\text{clo}} - t\right) \leqslant \exp\left(-\frac{n\mathrm{Var}[Y]^2}{128}\left(t - \frac{9}{8n}\right)^2\right).$$

II.1- Monte Carlo based Sobol' index inference

The Sobol' g-function: $f(x) = \prod_{i=1}^{d} f_i(x_i)$ with $f_i(x_i) = \frac{|4x_i - 2| + a_i}{1 + a_i}$,

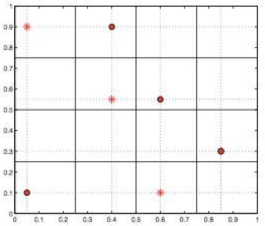
- $d = 8$,
- $a_1 = 0$, $a_2 = 1$, $a_3 = 4.5$, $a_4 = 9$, $a_i = 99$ for $5 \leq i \leq 8$,
- $n = 5000$, $b = 100$, $IC(0.95)$.





`sobolEff` `sobol2007`

Replicated latin hypercubes: (Tissot *et al.*)



Definition (Replicated Latin Hypercube Sampling)

$k = 1, \dots, n$

$$\mathbf{x}_k = \left(\frac{\pi_1(k) - U_{1,\pi_1(k)}}{n}, \dots, \frac{\pi_d(k) - U_{d,\pi_d(k)}}{n} \right)$$

$$\tilde{\mathbf{x}}_k = \left(\frac{\tilde{\pi}_1(k) - U_{1,\tilde{\pi}_1(k)}}{n}, \dots, \frac{\tilde{\pi}_d(k) - U_{d,\tilde{\pi}_d(k)}}{n} \right)$$

We have two matrices B and $\widetilde{B}$ at our disposal

$$B = \begin{pmatrix} x_{1,1} & \dots & x_{d,1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ x_{1,n} & \dots & x_{d,n} \end{pmatrix} \qquad \widetilde{B} = \begin{pmatrix} \tilde{x}_{1,1} & \dots & \tilde{x}_{d,1} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \tilde{x}_{1,n} & \dots & \tilde{x}_{d,n} \end{pmatrix}$$

We compute the model $\mathcal{M}$ 2n times (on the n lines of B and the n lines of $\widetilde{B}$).

Permutation of lines:

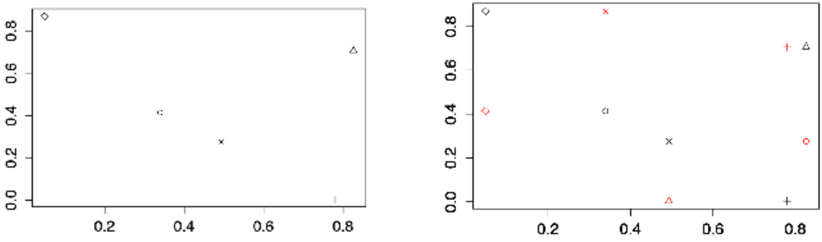
$$\begin{cases} \widetilde{B} = (\tilde{x}_{k,l})_{1 \le k \le n, 1 \le l \le d} & \to \quad \widetilde{B}_i = (\tilde{x}^i_{k,l})_{1 \le k \le n, 1 \le l \le d} \\ L_k & \mapsto \quad L_{\tilde{\pi}_i^{-1} \circ \pi_i(k)}, \; k = 1, \dots, n \end{cases}$$

Then, $\tilde{x}^i_{k,i} = \tilde{x}_{\tilde{\pi}_i^{-1} \circ \pi_i(k), i} = x_{k,i}$, $k = 1, \dots, n$.

To estimate S_i, we replace C_i with $\widetilde{B}_i$ (same column number i).

Caption:
point 1 ○ point 2 △ point 3 + point 4 × point 5 ◇

Design B (left), B and $\widetilde{B}$ (right)

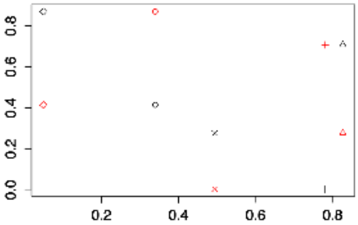


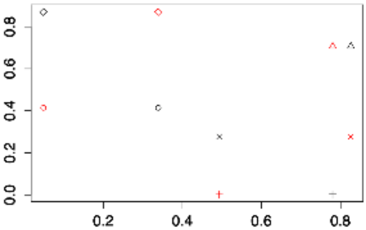
II- Monte Carlo based Sobol' index inference

Caption:

point 1 ○   point 2 △   point 3 +   point 4 ×   point 5 ◇

Design B and $\widetilde{B}_1$ (left), B and $\widetilde{B}_2$ (right)





Asymptotic confidence intervals with variance smaller than for MC.

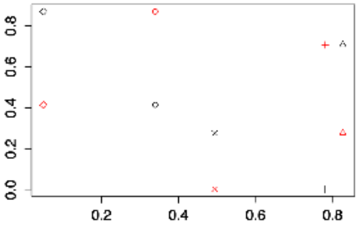
Possible extension to indices of order two (via orthogonal arrays of strength 2).

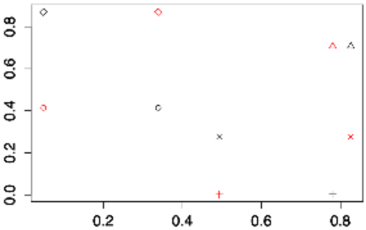
II- Monte Carlo based Sobol' index inference

Caption:

point 1 $\circ$ point 2 $\triangle$ point 3 $+$ point 4 $\times$ point 5 $\diamond$

Design B and $\widetilde{B}_1$ (left), B and $\widetilde{B}_2$ (right)





Asymptotic confidence intervals with variance smaller than for MC.

Possible extension to indices of order two (via orthogonal arrays of strength 2).

The trick cannot be used to estimate total Sobol' indices due to constraints inherent to the construction of OAs of strength $d-1$. If one wants to estimate total Sobol' indices, the best is to use Saltelli's trick (see Saltelli, 2002):

For $A \subseteq \{1, \ldots, d\}$, let us use the notation $U_A = \mathrm{Var}\left(\mathbb{E}(Y|\mathbf{X}_A)\right) + \mathbb{E}^2(Y)$ and $\mathbf{x}^{\sim A} = \left(\mathbf{x}_A, \mathbf{x}'_{-A}\right)$.

	$\mathbf{x}'$	$\mathbf{x}^{\sim 1}$	$\mathbf{x}^{\sim 2}$	$\mathbf{x}^{\sim 3}$	$\mathbf{x}^{\sim 4}$	$\mathbf{x}^{\sim \{2,3,4\}}$	$\mathbf{x}^{\sim \{1,3,4\}}$	$\mathbf{x}^{\sim \{1,2,4\}}$	$\mathbf{x}^{\sim \{1,2,3\}}$	$\mathbf{x}^{\sim \{1,2,3,4\}}$
$\mathbf{x}'$	V									
$\mathbf{x}^{\sim 1}$	U_{-1}	V								
$\mathbf{x}^{\sim 2}$	U_{-2}	U_{-12}	V							
$\mathbf{x}^{\sim 3}$	U_{-3}	U_{-13}	U_{-23}	V						
$\mathbf{x}^{\sim 4}$	U_{-4}	U_{-14}	U_{-24}	U_{-34}	V					
$\mathbf{x}^{\sim \{2,3,4\}}$	U_1		U_{12}	U_{13}	U_{14}	V				
$\mathbf{x}^{\sim \{1,3,4\}}$	U_2	U_{12}		U_{23}	U_{24}	U_{-12}	V			
$\mathbf{x}^{\sim \{1,2,4\}}$	U_3	U_{13}	U_{23}		U_{34}	U_{-13}	U_{-23}	V		
$\mathbf{x}^{\sim \{1,2,3\}}$	U_4	U_{14}	U_{24}	U_{34}		U_{-14}	U_{-24}	U_{-34}	V	
$\mathbf{x}^{\sim \{1,2,3,4\}}$	$\mathbb{E}^2\, Y$	U_1	U_2	U_3	U_4	U_{-1}	U_{-2}	U_{-3}	U_{-4}	V

Table: The table gives for each cell the term that can be estimated by evaluating the model on the corresponding input vectors, $d = 4$. For example, U_{-12} can be estimated from the evaluation of the model on two n-samples $\mathbf{x}^{\sim 2,(i)}$ and $\mathbf{x}^{\sim 1,(i)}$, $i = 1, \ldots, n$.

In conclusion, Saltelli's trick lead to the estimation of all first-order and total indices at a cost of $n(d+2)$ model evaluations and to the estimation of all first-order, second-order and total indices at a cost of $n(2d+2)$ model evaluations.

Conclusions about Monte Carlo type inference :

We recommend the following (see Gilquin *et al.*, 2019):

First and second order Sobol' indices: R package `sensitivity`, function `sobolrep` with `total=FALSE`.
The cost is $2n$ with $n = q^2$, q a prime number.

First, second order and total Sobol' indices: R package `sensitivity`, function `sobolrep` with `total=TRUE`.
The cost is $n(d+2)$ with $n = q^2$, q a prime number (see Gilquin *et al.*, 2019).

II.2- Given data Sobol' index inference

Pick-freeze estimator is based on a specific design of experiments 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, that is a given data estimator.*

Let us present rank estimator of S_1 from in Gamboa *et al.* (2021) .

Let's consider a n-sample of the input/output pair (X_1, Y) *given by* $(X_{1,1}, Y_1), \ldots, (X_{1,n}, Y_n)$.

The pairs $(X_{1,(1)}, Y_{(1)}), \ldots, (X_{1,(n)}, Y_{(n)})$ *are rearranged in such a way that* $X_{1,(1)} < \ldots < X_{1,(n)}$.

Example:

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

II.2- Given data Sobol' index inference

Pick-freeze estimator is based on a specific design of experiments 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, that is a given data estimator.*

Let us present rank estimator of S_1 from in Gamboa *et al.* (2021) .

Let's consider a n-sample of the input/output pair (X_1, Y) *given by* $(X_{1,1}, Y_1), \ldots, (X_{1,n}, Y_n)$.

The pairs $(X_{1,(1)}, Y_{(1)}), \ldots, (X_{1,(n)}, Y_{(n)})$ *are rearranged in such a way that* $X_{1,(1)} < \ldots < X_{1,(n)}$.

Example:

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

We define $Y_{(n+1)} = Y_{(1)}$. We then introduce

$$\widehat{S_1}^{\text{rank}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_{(i)} Y_{(i+1)} - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2}.$$

Theorem (Gamboa *et al.*, 2021, see also Chatterjee, 2020)

1. *Assume that* $X_i \sim \mathcal{U}[0,1])$*,* $i = 1, \ldots, n$ *and that* $\mathcal{M}$ *is bounded. One has* $\widehat{S_1}^{\text{rank}} \xrightarrow[n\to\infty]{a.s.} S_1$.
2. *Assume that* $X_i \sim \mathcal{U}[0,1])$*,* $i = 1, \ldots, n$ *and that* $\mathcal{M}$ *is twice differentiable wrt its first coordinate with bounded first derivatives. Then*

$$\sqrt{n}\left(\widehat{S_1}^{\text{rank}} - S_1\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \sigma^2_{\text{rank}}\right).$$

Rank estimator is limited to first-order Sobol' index estimation.

In Broto *et al.* (2020), the authors propose a given data estimator based on **nearest neighbors**. This estimator can be defined for any order of interaction.

Consistency is proved under regularity assumptions on the model. No CLT is proved.

Ishigami toy model: $\mathcal{M}(x) = \sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4\sin(x_1)$,
$X_i \sim \mathcal{U}([-\pi, \pi])$, $i = 1, 2, 3$.

We compare `sobolrank` with `sobolrep` with
$2 \times n = 2 \times 19^2 = 2 \times 361 = 722$ model evaluations, $n_{\text{rep}} = 100$. Root mean square errors are computed with 100 samples.

`sobolrank`	0.03635195	0.03440188	0.04715759
`sobolrep`	0.04199731	0.04436713	0.07468821

For the same number of model evaluatons, `sobolrep` also provides second-order Sobol' indices. However it requires a pick-freeze design based on replicated OAs of strength 2.

II.3- Spectral Sobol' index inference

For sake of clarity in the presentation, we consider the case $d = 2$.

$Y = \sum_{\mathbf{k}=(k_1,k_2)\in\mathbb{Z}^2} c_{\mathbf{k}}(\mathcal{M})\Phi_{1,k_1}(X_1)\Phi_{2,k_2}(X_2)$

with , for all $i = 1, 2$, $(\Phi_{i,k})_{k\in\mathbb{Z}}$ is an orthonormal basis of $\mathbb{L}^2([0,1])$ and $\Phi_{i,0} \equiv 1$.

$\mathcal{M}_0 = c_0(\mathcal{M})$,

$\mathcal{M}_1(X_1) = \sum_{k_1\in\mathbb{Z}^*} c_{k_1,0}(\mathcal{M})\Phi_{1,k_1}(X_1)$, $\mathcal{M}_2(X_2) = \sum_{k_2\in\mathbb{Z}^*} c_{0,k_2}(\mathcal{M})\Phi_{2,k_2}(X_2)$,

$\mathcal{M}_{1,2}(X_1, X_2) = \sum_{k_1\in\mathbb{Z}^*, k_2\in\mathbb{Z}^*} c_{k_1,k_2}(\mathcal{M})\Phi_{1,k_1}(X_1)\Phi_{2,k_2}(X_2)$.

We have with Parseval identity:

- $\mathrm{Var}(\mathcal{M}_1(X_1)) = \sigma_1^2 = \sum_{k_1\in\mathbb{Z}^*} |c_{k_1,0}(\mathcal{M})|^2$, (idem for σ_2^2),
- $\mathrm{Var}(\mathcal{M}_{1,2}(X_1, X_2)) = \sigma_{1,2}^2 = \sum_{k_1\in\mathbb{Z}^*, k_2\in\mathbb{Z}^*} |c_{k_1,k_2}(\mathcal{M})|^2$,
- $\mathrm{Var}(Y) = \sigma^2 = \sum_{(k_1,k_2)\in\mathbb{Z}\times\mathbb{Z}, (k_1,k_2)\neq(0,0)} |c_{k_1,k_2}(\mathcal{M})|^2$.

ex. : orthogonal polynomials, wavelet basis, Fourier basis.

II.3- Spectral Sobol' index inference

For sake of clarity in the presentation, we consider the case $d = 2$.

$Y = \sum_{\mathbf{k}=(k_1,k_2)\in\mathbb{Z}^2} c_{\mathbf{k}}(\mathcal{M})\Phi_{1,k_1}(X_1)\Phi_{2,k_2}(X_2)$

with , for all $i = 1, 2$, $(\Phi_{i,k})_{k\in\mathbb{Z}}$ is an orthonormal basis of $\mathbb{L}^2([0,1])$ and $\Phi_{i,0} \equiv 1$.

$\mathcal{M}_0 = c_0(\mathcal{M})$,

$\mathcal{M}_1(X_1) = \sum_{k_1\in\mathbb{Z}^*} c_{k_1,0}(\mathcal{M})\Phi_{1,k_1}(X_1)$, $\mathcal{M}_2(X_2) = \sum_{k_2\in\mathbb{Z}^*} c_{0,k_2}(\mathcal{M})\Phi_{2,k_2}(X_2)$,

$\mathcal{M}_{1,2}(X_1, X_2) = \sum_{k_1\in\mathbb{Z}^*,k_2\in\mathbb{Z}^*} c_{k_1,k_2}(\mathcal{M})\Phi_{1,k_1}(X_1)\Phi_{2,k_2}(X_2)$.

We have with Parseval identity:

- $\mathrm{Var}\left(\mathcal{M}_1(X_1)\right) = \sigma_1^2 = \sum_{k_1\in\mathbb{Z}^*} |c_{k_1,0}(\mathcal{M})|^2$, (idem for σ_2^2),
- $\mathrm{Var}\left(\mathcal{M}_{1,2}(X_1, X_2)\right) = \sigma_{1,2}^2 = \sum_{k_1\in\mathbb{Z}^*,k_2\in\mathbb{Z}^*} |c_{k_1,k_2}(\mathcal{M})|^2$,
- $\mathrm{Var}\left(Y\right) = \sigma^2 = \sum_{(k_1,k_2)\in\mathbb{Z}\times\mathbb{Z},(k_1,k_2)\neq(0,0)} |c_{k_1,k_2}(\mathcal{M})|^2$.

ex. : orthogonal polynomials, wavelet basis, Fourier basis.

For sake of clarity in the presentation, we consider the case $d = 2$.

$$Y = \sum_{\mathbf{k}=(k_1,k_2)\in\mathbb{Z}^2} c_{\mathbf{k}}(\mathcal{M})\Phi_{1,k_1}(X_1)\Phi_{2,k_2}(X_2)$$

with , for all $i = 1, 2$, $(\Phi_{i,k})_{k\in\mathbb{Z}}$ is an orthonormal basis of $\mathbb{L}^2([0,1])$ and $\Phi_{i,0} \equiv 1$.

$\mathcal{M}_0 = c_0(\mathcal{M})$,

$\mathcal{M}_1(X_1) = \sum_{k_1\in\mathbb{Z}^*} c_{k_1,0}(\mathcal{M})\Phi_{1,k_1}(X_1)$, $\mathcal{M}_2(X_2) = \sum_{k_2\in\mathbb{Z}^*} c_{0,k_2}(\mathcal{M})\Phi_{2,k_2}(X_2)$,

$\mathcal{M}_{1,2}(X_1, X_2) = \sum_{k_1\in\mathbb{Z}^*,k_2\in\mathbb{Z}^*} c_{k_1,k_2}(\mathcal{M})\Phi_{1,k_1}(X_1)\Phi_{2,k_2}(X_2)$.

We have with Parseval identity:

- $\mathrm{Var}\,(\mathcal{M}_1(X_1)) = \sigma_1^2 = \sum_{k_1\in\mathbb{Z}^*} |c_{k_1,0}(\mathcal{M})|^2$, (idem for σ_2^2),
- $\mathrm{Var}\,(\mathcal{M}_{1,2}(X_1, X_2)) = \sigma_{1,2}^2 = \sum_{k_1\in\mathbb{Z}^*,k_2\in\mathbb{Z}^*} |c_{k_1,k_2}(\mathcal{M})|^2$,
- $\mathrm{Var}\,(Y) = \sigma^2 = \sum_{(k_1,k_2)\in\mathbb{Z}\times\mathbb{Z},(k_1,k_2)\neq(0,0)} |c_{k_1,k_2}(\mathcal{M})|^2$.

ex. : orthogonal polynomials, wavelet basis, Fourier basis.

II.3- Spectral Sobol' index inference

For sake of clarity in the presentation, we consider the case $d = 2$.

$Y = \sum_{\mathbf{k}=(k_1,k_2)\in\mathbb{Z}^2} c_{\mathbf{k}}(\mathcal{M})\Phi_{1,k_1}(X_1)\Phi_{2,k_2}(X_2)$

with , for all $i = 1, 2$, $(\Phi_{i,k})_{k\in\mathbb{Z}}$ is an orthonormal basis of $\mathbb{L}^2([0,1])$ and $\Phi_{i,0} \equiv 1$.

$\mathcal{M}_0 = c_0(\mathcal{M})$,

$\mathcal{M}_1(X_1) = \sum_{k_1\in\mathbb{Z}^*} c_{k_1,0}(\mathcal{M})\Phi_{1,k_1}(X_1)$, $\mathcal{M}_2(X_2) = \sum_{k_2\in\mathbb{Z}^*} c_{0,k_2}(\mathcal{M})\Phi_{2,k_2}(X_2)$,

$\mathcal{M}_{1,2}(X_1, X_2) = \sum_{k_1\in\mathbb{Z}^*,k_2\in\mathbb{Z}^*} c_{k_1,k_2}(\mathcal{M})\Phi_{1,k_1}(X_1)\Phi_{2,k_2}(X_2)$.

We have with Parseval identity:

- $\mathrm{Var}(\mathcal{M}_1(X_1)) = \sigma_1^2 = \sum_{k_1\in\mathbb{Z}^*} |c_{k_1,0}(\mathcal{M})|^2$, (idem for σ_2^2),
- $\mathrm{Var}(\mathcal{M}_{1,2}(X_1, X_2)) = \sigma_{1,2}^2 = \sum_{k_1\in\mathbb{Z}^*,k_2\in\mathbb{Z}^*} |c_{k_1,k_2}(\mathcal{M})|^2$,
- $\mathrm{Var}(Y) = \sigma^2 = \sum_{(k_1,k_2)\in\mathbb{Z}\times\mathbb{Z},(k_1,k_2)\neq(0,0)} |c_{k_1,k_2}(\mathcal{M})|^2$.

ex. : orthogonal polynomials, wavelet basis, Fourier basis.

II.3- Spectral Sobol' index inference

Inference scheme:

If D is an experimental design with $[0,1]^2$, we propose the quadrature formula:

$$\hat{c}_{k_1,k_2}(\mathcal{M}, D) = \frac{1}{\mathrm{card} D} \sum_{\mathbf{x}=(x_1,x_2)\in D} \mathcal{M}(x) e^{-2i\pi(k_1 x_1 + k_2 x_2)} .$$

We then infer each part of variance with a truncation:

- $\hat{\sigma}_1^2(\mathcal{M}, K_1, D) = \sum_{k_1 \in K_1} |\hat{c}_{k_1,0}(\mathcal{M}, D)|^2$, with $K_1 \subset \mathbb{Z}^*$ of finite cardinal, (idem for $\hat{\sigma}_2^2$),
- $\hat{\sigma}_{1,2}^2(\mathcal{M}, K_{1,2}, D) = \sum_{(k_1,k_2)\in K_{1,2}} |\hat{c}_{k_1,k_2}(\mathcal{M}, D)|^2$, with $K_{1,2} \subset \mathbb{Z}^* \times \mathbb{Z}^*$ of finite cardinal.

We infer the total variance with $\hat{\sigma}^2(\mathcal{M}, D) = \hat{c}_{0,0}(\mathcal{M}^2, D) - \hat{c}_{0,0}(\mathcal{M}, D)^2$.

The estimators of Sobol' indices can be written as:

$$\hat{S}_i = \frac{\hat{\sigma}_i^2}{\hat{\sigma}^2}, \; i = 1, 2, \quad S_{1,2} = \frac{\hat{\sigma}_{1,2}^2}{\hat{\sigma}^2} .$$

II.3- Spectral Sobol' index inference

Inference scheme:

If D is an experimental design with $[0,1]^2$, we propose the quadrature formula:

$$\hat{c}_{k_1,k_2}(\mathcal{M}, D) = \frac{1}{\mathrm{card} D} \sum_{\mathbf{x}=(x_1,x_2)\in D} \mathcal{M}(x) e^{-2i\pi(k_1 x_1 + k_2 x_2)} .$$

We then infer each part of variance with a truncation:

- $\hat{\sigma}_1^2(\mathcal{M}, K_1, D) = \sum_{k_1 \in K_1} |\hat{c}_{k_1,0}(\mathcal{M}, D)|^2$, with $K_1 \subset \mathbb{Z}^*$ of finite cardinal, (idem for $\hat{\sigma}_2^2$),
- $\hat{\sigma}_{1,2}^2(\mathcal{M}, K_{1,2}, D) = \sum_{(k_1,k_2)\in K_{1,2}} |\hat{c}_{k_1,k_2}(\mathcal{M}, D)|^2$, with $K_{1,2} \subset \mathbb{Z}^* \times \mathbb{Z}^*$ of finite cardinal.

We infer the total variance with $\hat{\sigma}^2(\mathcal{M}, D) = \hat{c}_{0,0}(\mathcal{M}^2, D) - \hat{c}_{0,0}(\mathcal{M}, D)^2$.

The estimators of Sobol' indices can be written as:

$$\hat{S}_i = \frac{\hat{\sigma}_i^2}{\hat{\sigma}^2},\ i = 1, 2, \quad S_{1,2} = \frac{\hat{\sigma}_{1,2}^2}{\hat{\sigma}^2} .$$

II.3- Spectral Sobol' index inference

Inference scheme:

If D is an experimental design with $[0,1]^2$, we propose the quadrature formula:

$$\hat{c}_{k_1,k_2}(\mathcal{M}, D) = \frac{1}{\text{card} D} \sum_{\mathbf{x}=(x_1,x_2)\in D} \mathcal{M}(x) e^{-2i\pi(k_1 x_1 + k_2 x_2)} .$$

We then infer each part of variance with a truncation:

- $\hat{\sigma}_1^2(\mathcal{M}, K_1, D) = \sum_{k_1 \in K_1} |\hat{c}_{k_1,0}(\mathcal{M}, D)|^2$, with $K_1 \subset \mathbb{Z}^*$ of finite cardinal, (idem for $\hat{\sigma}_2^2$),
- $\hat{\sigma}_{1,2}^2(\mathcal{M}, K_{1,2}, D) = \sum_{(k_1,k_2)\in K_{1,2}} |\hat{c}_{k_1,k_2}(\mathcal{M}, D)|^2$, with $K_{1,2} \subset \mathbb{Z}^* \times \mathbb{Z}^*$ of finite cardinal.

We infer the total variance with $\hat{\sigma}^2(\mathcal{M}, D) = \hat{c}_{0,0}(\mathcal{M}^2, D) - \hat{c}_{0,0}(\mathcal{M}, D)^2$.

The estimators of Sobol' indices can be written as:

$$\hat{S}_i = \frac{\hat{\sigma}_i^2}{\hat{\sigma}^2}, \; i = 1,2, \quad S_{1,2} = \frac{\hat{\sigma}_{1,2}^2}{\hat{\sigma}^2} .$$

II.3- Spectral Sobol' index inference

Inference scheme:

If D is an experimental design with $[0,1]^2$, we propose the quadrature formula:

$$\hat{c}_{k_1,k_2}(\mathcal{M}, D) = \frac{1}{\text{card} D} \sum_{\mathbf{x}=(x_1,x_2)\in D} \mathcal{M}(x) e^{-2i\pi(k_1 x_1 + k_2 x_2)} .$$

We then infer each part of variance with a truncation:

- $\hat{\sigma}_1^2(\mathcal{M}, K_1, D) = \sum_{k_1 \in K_1} |\hat{c}_{k_1,0}(\mathcal{M}, D)|^2$, with $K_1 \subset \mathbb{Z}^*$ of finite cardinal, (idem for $\hat{\sigma}_2^2$),
- $\hat{\sigma}_{1,2}^2(\mathcal{M}, K_{1,2}, D) = \sum_{(k_1,k_2)\in K_{1,2}} |\hat{c}_{k_1,k_2}(\mathcal{M}, D)|^2$, with $K_{1,2} \subset \mathbb{Z}^* \times \mathbb{Z}^*$ of finite cardinal.

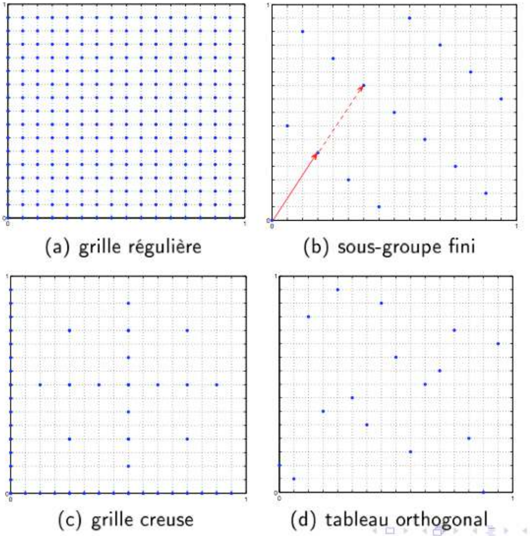
We infer the total variance with $\hat{\sigma}^2(\mathcal{M}, D) = \hat{c}_{0,0}(\mathcal{M}^2, D) - \hat{c}_{0,0}(\mathcal{M}, D)^2$.

The estimators of Sobol' indices can be written as:

$$\hat{S}_i = \frac{\hat{\sigma}_i^2}{\hat{\sigma}^2}, \; i = 1, 2, \quad S_{1,2} = \frac{\hat{\sigma}_{1,2}^2}{\hat{\sigma}^2} .$$

II.3- Spectral Sobol' index inference

Classical designs:



The performance of previous estimators is linked to the decreasing speed of Fourier spectrum (regularity) of $\mathcal{M}$. The techniques FAST and RBD are two particular cases of such approaches (after model regularisation). See Tissot & Prieur (2012) or Prieur & Tarantola (2017) for a review.

FAST: (Cukier *et al.*, 78) *Fourier Amplitude Sensitivity Test*

• we fix $K_{\mathbf{u}}$ an ensemble of a priori non negligible frequencies;

• we chose D cyclic group (design (b)) in order to control the quadrature error.

Remarks:

- if $\mathcal{M}$ regular, we can obtain a speed of convergence $>> \sqrt{n}$;
- for the total indices `fast99()` (no confidence intervals in the function) Saltelli *et al.*, 99.

II.3- Spectral Sobol' index inference

The performance of previous estimators is linked to the decreasing speed of Fourier spectrum (regularity) of $\mathcal{M}$. The techniques FAST and RBD are two particular cases of such approaches (after model regularisation). See Tissot & Prieur (2012) or Prieur & Tarantola (2017) for a review.

FAST: (Cukier *et al.*, 78) *Fourier Amplitude Sensitivity Test*

• we fix $K_{\mathbf{u}}$ an ensemble of a priori non negligible frequencies;

• we chose D cyclic group (design (b)) in order to control the quadrature error.

Remarks:

- if $\mathcal{M}$ regular, we can obtain a speed of convergence $>> \sqrt{n}$;
- for the total indices `fast99()` (no confidence intervals in the function) Saltelli *et al.*, 99.

The performance of previous estimators is linked to the decreasing speed of Fourier spectrum (regularity) of $\mathcal{M}$. The techniques FAST and RBD are two particular cases of such approaches (after model regularisation). See Tissot & Prieur (2012) or Prieur & Tarantola (2017) for a review.

FAST: (Cukier *et al.*, 78) *Fourier Amplitude Sensitivity Test*

- we fix $K_{\mathbf{u}}$ an ensemble of a priori non negligible frequencies;
- we chose D cyclic group (design (b)) in order to control the quadrature error.

Remarks:

- if $\mathcal{M}$ regular, we can obtain a speed of convergence $>> \sqrt{n}$;
- for the total indices `fast99()` (no confidence intervals in the function) Saltelli *et al.*, 99.

The performance of previous estimators is linked to the decreasing speed of Fourier spectrum (regularity) of $\mathcal{M}$. The techniques FAST and RBD are two particular cases of such approaches (after model regularisation). See Tissot & Prieur (2012) or Prieur & Tarantola (2017) for a review.

FAST: (Cukier *et al.*, 78) *Fourier Amplitude Sensitivity Test*

- we fix $K_{\mathbf{u}}$ an ensemble of a priori non negligible frequencies;
- we chose D cyclic group (design (b)) in order to control the quadrature error.

Remarks:

- if $\mathcal{M}$ regular, we can obtain a speed of convergence $>> \sqrt{n}$;
- for the total indices `fast99()` (no confidence intervals in the function) Saltelli *et al.*, 99.

<u>RBD:</u> (Tarantola *et al.*, 06) *Random Balance Designs*

• we choose D an orthogonal array of strength 1 (design (d)), randomized by a random permutation ($D(\pi)$));

• $K_{\mathbf{u}}$ choice of a priori non negligible frequencies.

<u>Remarks:</u>

- these estimators are known to be biased;
- we can correct a part of this biais (Tissot *et al.*, 2012);
- if the function is not regular enough, the bias remains important.

RBD: (Tarantola *et al.*, 06) *Random Balance Designs*

• we choose D an orthogonal array of strength 1 (design (d)), randomized by a random permutation ($D(\pi)$));

• $K_{\mathbf{u}}$ choice of a priori non negligible frequencies.

Remarks:

- these estimators are known to be biased;
- we can correct a part of this biais (Tissot *et al.*, 2012);
- if the function is not regular enough, the bias remains important.

RBD: (Tarantola *et al.*, 06) *Random Balance Designs*

• we choose D an orthogonal array of strength 1 (design (d)), randomized by a random permutation ($D(\pi)$));

• $K_{\mathbf{u}}$ choice of a priori non negligible frequencies.

Remarks:

- these estimators are known to be biased;
- we can correct a part of this biais (Tissot *et al.*, 2012);
- if the function is not regular enough, the bias remains important.

II.4- Conclusion on Sobol' index inference

see Jupyter notebook Premiers–Pas and GSA–COVID19.

III- Sensitivity indices based on the Cramér-von-Mises distance

Let $Y = \mathcal{M}(X_1, \ldots, X_d) \in \mathbb{R}^p$ be the code output and F be its cumulative distribution function defined as

$$F(t) = \mathbb{P}\left(Y \leqslant t\right) = \mathbb{E}\left[\mathbb{1}_{\{Y \leqslant t\}}\right] = \mathbb{E}\left[Z(t)\right],\ t = (t_1, \ldots, t_p) \in \mathbb{R}^p.$$

Let $F^{\mathbf{u}}(t)$ be the conditional cumulative distribution function of Y conditionally on $X_{\mathbf{u}}$ defined as

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

We perform the Hoeffding decomposition of $Z(t)$:

$$\begin{aligned} Z(t) = \mathbb{1}_{\{Y \leqslant t\}} = & \underbrace{\mathbb{E}\left[Z(t)\right]}_{\text{Mean effect}} \\ & + \underbrace{\left(\mathbb{E}\left[Z(t)|X_{\mathbf{u}}\right] - \mathbb{E}\left[Z(t)\right]\right) + \left(\mathbb{E}\left[Z(t)|X_{-\mathbf{u}}\right] - \mathbb{E}\left[Z(t)\right]\right)}_{\text{First order effects}} \\ & + \underbrace{R(t, \mathbf{u}).}_{\text{Remainder term: higher order effects}} \end{aligned}$$

We then compute the variance of both sides of the previous equation:

$$\mathrm{Var}\left[Z(t)\right] = \mathbb{E}\left[\left(F^{\mathbf{u}}(t) - F(t)\right)^2\right] + \mathbb{E}\left[\left(F^{-\mathbf{u}}(t) - F(t)\right)^2\right] + \mathrm{Var}\left[R(t, \mathbf{u})\right]$$

using orthogonality in the Hoeffding decomposition.
Finally by integrating with respect to the distribution of $Z(t)$ and by normalizing we get:

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

involving the Cramér-von Mises distance between the distribution of $Z(t)$ and the one of $Z(t)|X_{\mathbf{u}}$.

Properties of the Cramér-von Mises indices:

1. the different contributions sum to 1;
2. invariance by any translation and by any nondegenerated scaling of the components of Y.

Cramér-von Mises indices have no clear dual formulation, however they can be estimated with a Pick-Freeze scheme.

Other estimation procedures such as U-statistics or rank-based inference (only for scalar inputs and $\mathbf{u}$ a singleton) are also interesting alternatives (see Gamboa *et al.*, 2018) .

IV- Towards general metric space indices

Let us consider the more general case where $Y = \mathcal{M}(X_1, \ldots, X_d)$ valued in $\mathcal{Y}$, a general metric space. Let $m \in \mathbb{N}^*$ and $a = (a_i)_{i=1,\ldots,m} \in \mathcal{Y}^m$. We consider the family of test functions

$$\left\{ \begin{array}{ccc} \mathcal{Y}^m \times \mathcal{Y} & \to & \mathbb{R} \\ (a, y) & \mapsto & T_a(y). \end{array} \right.$$

We assume that $T_a(\cdot) \in L^2(\mathbb{P}^{\otimes m} \otimes \mathbb{P})$ with $\mathbb{P}$ the probability distribution of Y.

The general metric space sensitivity index with respect to **u**, introduced in Fort *et al.* (2021), is defined as

$$S_{2,GMS}^{\boldsymbol{u}} := \frac{\int_{\mathcal{Y}^m} \mathbb{E}_{X_{\boldsymbol{u}}} \left[\left(\mathbb{E}_Y[T_a(Y)] - \mathbb{E}_Y[T_a(Y)|X_{\boldsymbol{u}}] \right)^2 \right] d\mathbb{P}^{\otimes m}(a)}{\int_{\mathcal{Y}^m} \mathrm{Var}(T_a(Y)) d\mathbb{P}^{\otimes m}(a)}$$

$$= \frac{\int_{\mathcal{Y}^m} \mathrm{Var}\left[\mathbb{E}\left(T_a(Y)|X_{\boldsymbol{u}}\right)\right] d\mathbb{P}^{\otimes m}(a)}{\int_{\mathcal{Y}^m} \mathrm{Var}(T_a(Y)) d\mathbb{P}^{\otimes m}(a)}.$$

Particular examples:

1. for $\mathcal{Y} = \mathbb{R}$, $m = 0$ and $T_a(y) = y$, one recovers Sobol' indices;
2. for $\mathcal{Y} = \mathbb{R}^k$, $m = 1$ and $T_a(y) = \mathbb{1}_{\{y \leqslant a\}}$, one recovers the index based on the Cramér-von-Mises distance;
3. for $\mathcal{Y} = \mathcal{M}$ a manifold, $m = 2$ and

$$T_a(y) = \mathbb{1}_{y \in \widetilde{B}(a_1, a_2)} = \mathbb{1}_{\|y-(a_1+a_2)/2\| \leqslant \|a_1-a_2\|/2},$$

   where $\widetilde{B}(a_1, a_2)$ is the ball in $\mathcal{M}$ of diameter $\overline{a_1 a_2}$, one recovers the indices introduced in Fraiman *et al.* (2021).

General metric space indices can be estimated with either a pick-freeze scheme or U-statistics. For scalar inputs and first-order indices, a rank-based inference procedure is also an alternative.

Principle:

- multiple Monte-Carlo estimation procedure (one to handle the integration part, one to handle the pick-freeze part);
- cost to estimate all first-order indices: $N(m + d + 1)$;
- non trivial proof of the CLT using Donsker theorem and the functional delta method (see Fort *et al.*, 2021).

Design of experiments:

- a classical pick-freeze N-sample, that is two N-samples of Y: $(y^{(k)}, y^{\mathbf{u},(k)})$, $1 \leqslant k \leqslant N$;
- m other N-samples of Y independent of $(Y^{(k)}, Y^{\mathbf{u},(k)})_{1 \leqslant k \leqslant N}$, namely $w_i^{(k)}$, $1 \leqslant i \leqslant m$, $1 \leqslant k \leqslant N$.

V- Pick-freeze estimation procedure for Cramér-von Mises indices

The estimator of the numerator of $S_{2,\mathrm{GMS}}^{\mathbf{u}}$ is then given by

$$\frac{1}{N^m}\sum_{1\leqslant i_1,\dots,i_m\leqslant N}\Bigg\{\left[\frac{1}{N}\sum_{k=1}^{N} T_{w_1^{(i_1)},\cdots,w_m^{(i_m)}}(y^{(k)})\,T_{w_1^{(i_1)},\cdots,w_m^{(i_m)}}(y^{\mathbf{u},(k)})\right] - \left[\frac{1}{2N}\sum_{k=1}^{N}\left(T_{w_1^{(i_1)},\cdots,w_m^{(i_m)}}(y^{(k)}) + T_{w_1^{(i_1)},\cdots,w_m^{(i_m)}}(y^{\mathbf{u},(k)})\right)\right]^2\Bigg\}$$

while the one of the denominator is

$$\frac{1}{N^m}\sum_{1\leqslant i_1,\dots,i_m\leqslant N}\Bigg\{\frac{1}{2N}\sum_{k=1}^{N}\left[\left(T_{w_1^{(i_1)},\cdots,w_m^{(i_m)}}(y^{(k)})\right)^2 + \left(T_{w_1^{(i_1)},\cdots,w_m^{(i_m)}}(y^{\mathbf{u},(k)})\right)^2\right] - \left[\frac{1}{2N}\sum_{k=1}^{N}\left(T_{w_1^{(i_1)},\cdots,w_m^{(i_m)}}(y^{(k)}) + T_{w_1^{(i_1)},\cdots,w_m^{(i_m)}}(y^{\mathbf{u},(k)})\right)\right]^2\Bigg\}.$$

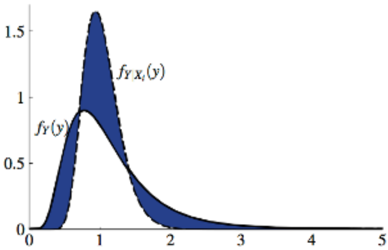
VI- Indices “à la Borgonovo”

In Borgonovo *et al.* (2007), the following index is introduced:

$$\delta_i = \frac{1}{2}\mathbb{E}_{X_i}\left(S_i(X_i)\right) \text{ with } S_i(X_i) = \int \left|p_Y(y) - p_{Y|X_i}(y)\right| dy.$$

Note that $S_i(X_i)$ is the total variation distance between $\mathbb{P}_Y$ and $\mathbb{P}_{Y|X_i}$.

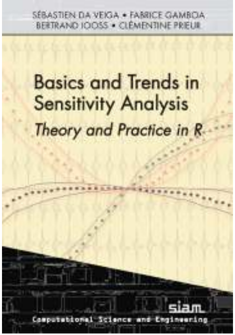
The definition can be generalized as: $S_i(X_i) = \int_{\mathbb{R}} f\left(\frac{p_Y(y)}{p_{Y|X_i}(y)}\right) p_{Y|X_i}(y)dy$ for f any convex function with $f(1) = 0$. E.g., for $f(t) = -\ln(t)$ or $f(t) = t\ln(t)$ one recovers the Kullback-Leibler divergence.



VII- Kernel based ANOVA decomposition

See the lecture by Sébastien Da Veiga based on Da Veiga (2021) .

Also, to appear this week:



A short bibliography I

[BBD20] B. Broto, F. Bachoc, and M. Depecker. Variance Reduction for Estimation of Shapley Effects and Adaptation to Unknown Input Distribution. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):693–716, 2020.

[Bor07] E. Borgonovo. A new uncertainty importance measure. *Reliability Engineering and System Safety*, 92(6):771–784, 2007.

[CFS+73] R. I. Cukier, C. M. Fortuin, K. E. Shuler, A. G. Petschek, and J. H. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients: Theory. *Journal of Chemical Physics*, 59:3873–3878, 1973.

[Cha20] S. Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 0(0):1–21, 2020.

[CLS78] R. I. Cukier, H. B. Levine, and K. E. Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics*, 26:1–42, 1978.

[CSS75] R. I. Cukier, J. H. Schaibly, and K. E. Shuler. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients: Analysis of the approximations. *Journal of Chemical Physics*, 63:1140–1149, 1975.

[DV21] S. Da Veiga. Kernel-based anova decomposition and shapley effects–application to global sensitivity analysis. *arXiv preprint arXiv:2101.05487*, 2021.

[dVGI+21] S. da Veiga, F. Gamboa, B. Iooss, C. Prieur, Society for Industrial, and Applied Mathematics. *Basics and Trends in Sensitivity Analysis: Theory and Practice in R*. Computational science and engineering. Society for Industrial and Applied Mathematics, 2021.

[FGM20] R. Fraiman, F. Gamboa, and L. Moreno. Sensitivity indices for output on a riemannian manifold. *International Journal for Uncertainty Quantification*, 10(4), 2020.

A short bibliography II

[FKL21] J.-C. Fort, T. Klein, and A. Lagnoux. Global sensitivity analysis and wasserstein spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):880–921, 2021.

[GAPJ19] L. Gilquin, E. Arnaud, C. Prieur, and A. Janon. Making the best use of permutations to compute sensitivity indices with replicated orthogonal arrays. *Reliability Engineering & System Safety*, 187:28–39, 2019.

[GGKL] F. Gamboa, P. Gremaud, T. Klein, and journal=arXiv preprint arXiv:2003.01772 year=2020 Lagnoux, A. Global sensitivity analysis: a new generation of mighty estimators based on rank statistics.

[GJK+16] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for sobol pick-freeze monte carlo method. *Statistics*, 50(4):881–902, 2016.

[GJKL14] F. Gamboa, A. Janon, T. Klein, and A. Lagnoux. Sensitivity analysis for multidimensional and functional outputs. *Electronic Journal of Statistics*, 8(1):575–603, 2014.

[GKL18] F. Gamboa, T. Klein, and A. Lagnoux. Sensitivity analysis based on cramér–von mises distance. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):522–548, 2018.

[Hoe48] W. F. Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Mathematical Statistics*, 19:293–325, 1948.

[Ioo11] B. Iooss. Revue sur l'analyse de sensibilité globale de modèles numériques. *Journal de la Société Française de Statistique*, 152(1):3–25, 2011.

[JKL+14] 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, 2014.

[LIPG13] M. Lamboni, B. Iooss, A.-L. Popelin, and F. Gamboa. Derivative-based global sensitivity measures: general links with Sobol' indices and numerical tests. *Mathematics and Computers in Simulation*, 87:45–54, 2013.

A short bibliography III

[Mau02] W. Mauntz. Global sensitivity analysis of general nonlinear systems. *Master's Thesis, Imperial College. Supervisors: C. Pantelides and S. Kucherenko*, 2002.

[Mor91] M. D. Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

[PT17] Clémentine Prieur and Stefano Tarantola. Variance-based sensitivity analysis: Theory and estimation algorithms. *Handbook of uncertainty quantification*, pages 1217–1239, 2017.

[Sal02] A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145:280–297, 2002.

[SCS00] A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity Analysis*. John Wiley & Sons, 2000.

[SG95] I. M. Sobol' and A Gresham. On an alternative global sensitivity estimators. *Proceedings of SAMO, Belgirate*, pages 40–42, 1995.

[SK09] I. M. Sobol' and S. Kucherenko. Derivative based global sensitivity measures and the link with global sensitivity indices. *Mathematics and Computers in Simulation*, 79:3009–3017, 2009.

[Sob93] I. M. Sobol'. Sensitivity analysis for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414, 1993.

[TGM06] S. Tarantola, D. Gatelli, and T. A. Mara. Random balance designs for the estimation of first-order global sensitivity indices. *Reliability Engineering and System Safety*, 91:717–727, 2006.

[TP12a] J. Y. Tissot and C. Prieur. Bias correction for the estimation of sensitivity indices based on random balance designs. *Reliability Engineering and System Safety*, 107:205–213, 2012.

[TP12b] J. Y. Tissot and C. Prieur. Variance-based sensitivity analysis using harmonic analysis. `http://hal.archives-ouvertes.fr/docs/00/68/07/25/PDF`, 2012.

[TP15] J. Y. Tissot and C. Prieur. A randomized orthogonal array-based procedure for the estimation of first- and second-order sobol' indices. *Journal of Statistical Computation and Simulation*, 85:1358–1381, 2015.