

# Sensitivity analysis

Anne DUTFOY

EDF R&D PERICLES. anne.dutfoy@edf.Fr

PRACE 2020



## Context

#### We consider

$$Y = f(\underline{X})$$

- f is a model (scientific simulation software, symbolic function ...)
- $\frac{X}{d} = (X_1, \dots, X_d)$  is the set of **uncertain parameters** modeled by a multivariate distribution of dimension
- Y is the feature of interest evaluated by the model, supposed here to be scalar.

#### Why sensitivity analyses?

The main objectives of sensitivity analyses are the following:

- remove some variables which are not influential on the feature of interest, within a context of high dimension,
- $oldsymbol{\circ}$  prioritize variables in order to prioritize modeling efforts : we need a *relative* quantification : we want that if  $S_i < S_j$  then  $\mathbb{P}\left(\hat{S}_j \leq \hat{S}_i\right) \leq \varepsilon$
- $\textbf{ 9 quantify the impact of a variable}: \text{ we need a } \textit{exact} \textit{ quantification}: \text{ we want that } \mathbb{P}\left(|\hat{S}_i S_i| > \eta\right) \leq \varepsilon$



# Sensitivity: several notions

Several features can quantify the dependence.

### Sensitivity = Dispersion = Variance

If we agree that the variance is a good way to quantify the dispersion, sensitivity analyses aim at determining the most important contributors to the variance of Y. We use the **conditional expectation**  $\mathbb{E}(Y|X_i) = Y_i^*$  which is the random variable function of  $X_i$  which approximates Y the best in the least square sense :

$$Y_i^* = argmin_g \mathbb{E}\left(\left[Y - g(X_i)\right]^2\right)$$

No constraint on the nature of the link between Y and  $X_i$ .

We want to compare  $Var(Y_i^*)$  to Var(Y):

- 1 in the case of independent variables: Sobol indices,
- 2 in the case of dependent variables : importance factors (Taylor decomposition variance), ANCOVA indices.

### Sensitivity = Distance to the independence

If Y and  $X_i$  are correlated, the copula of  $(Y, X_i)$  is not far away from the independent copula. The Csiszar divergence measures enable to quantify that distance.





# Sommaire

1 Independent Variables

Sobol Indices An example

Particular cases: historical measures

2 Dependent variables

Taylor decomposition ANCOVA Indices

3 Extensions

Indices based on the Csiszar divergence

4 References



## Sommaire

- Independent Variables
   Sobol Indices
   An example
   Particular cases: historical measures
- Dependent variables
- 3 Extensions
- 4 References



#### Variance decomposition

Generally, if Y = f(X) and X with **independent components**, then we can decompose the variance as follows:

$$Var(Y) = \sum_{i} Var(\mathbb{E}(Y|X_i)) + \sum_{i \neq j} Var(\mathbb{E}(Y|X_i, X_j)) + \dots + \underbrace{Var(\mathbb{E}(Y|X_1, \dots, X_n))}_{=0}$$
(1)

#### Sobol Indices

The Sobol indices of order k quantifies the part of the variance of Y explained by the variance of  $(X_{i_1}, \ldots, X_{i_k})$ :

$$S_{i_1,\ldots,i_k} = \frac{\operatorname{Var}\left(\mathbb{E}\left(Y|X_{i_1},\ldots,X_{i_k}\right)\right)}{\operatorname{Var}\left(Y\right)} \tag{2}$$

The **total Sobol indices of order** k quantifies the part of the variance of Y explained by the inputs  $(X_i, \ldots, X_{i_k}, \tilde{X})$ :

$$S_{i_1,\ldots,i_k}^T = \frac{\sum_{l} \operatorname{Var}(\mathbb{E}(Y|X_l))}{\operatorname{Var}(Y)}, \ \{i_1,\ldots,i_k\} \subset I \subset \{1,\ldots,n\}$$
(3)



# The Hoeffding decomposition

The decomposition (1) of the variance of Y comes from the functional Hoeffding decomposition.

### Hoeffding decomposition of a function integrable on $[0,1]^n$

If f is integrable on  $[0,1]^n$ , it admits an unique decomposition which writes :

$$f(x_1,\ldots,x_n)=f_0+\sum_{i=1}^{n-1}f_i(x_i)+\sum_{1\leq i< j\leq n}f_{i,j}(x_i,x_j)+\cdots+f_{1,\ldots,n}(x_1,\ldots,x_n)$$
 (4)

where  $f_0 = cst$  and the other functions are mutually orthogonal with respect to the Lebesgue measure on  $[0,1]^n$ :

$$\int_0^1 f_{i_1,\ldots,i_s}(x_{i_1},\ldots,x_{i_s})f_{j_1,\ldots,j_k}(x_{j_1},\ldots,x_{j_k})d\underline{x}=0$$
 (5)

as soon as  $(i_1,\ldots,i_s)\neq (j_1,\ldots,j_k)$ .



How can we use this result for  $Y = f(\underline{X})$  with  $\underline{X}$  a random vector?

#### How can we use this result

We would like to decompose f according to Hoeffding decomposition ...but :

- **1** The inputs of f are not in  $[0,1]^n$ : generally,  $Y=f(\underline{X})$  where  $\underline{X}$  is defined on  $\mathbb{R}$ .
- $\Longrightarrow$  If we note

$$\underline{U} = (F_1(X_1), \dots, F_n(X_n))^t) = \phi^{-1}(\underline{X})$$
(6)

then  $\underline{U}$  has uniform marginals and its copula is the same as  $\underline{X}$ , then we can use the Hoeffding decomposition on  $f \circ \phi$ .

- **1** Are the Sobol indices w.r.t. the  $U_i$  the same as those w.r.t. the  $X_i$ ?
- $\implies$  If  $\underline{U} = \psi(\underline{X})$  where  $\psi$  is a diffeomorphism and  $Y = f(\underline{X})$  then :

$$\mathbb{E}\left(Y|\underline{U}\right) = \mathbb{E}\left(Y|\underline{X}\right) \tag{7}$$

As a matter of fact :  $\mathbb{E}(Y|\underline{U}) = \mathbb{E}(Y|\psi(X))$  is the orthogonal projection in a  $L_2$  sense of Y on the space generated by  $\psi(X)$ , which is the same as the one generated by X, thus the equality of the random variables (7). As the transformation  $\phi$  (6) acts component by component,  $(U_i \leftrightarrow X_i)$  then we have :

$$\operatorname{Var}\left(\mathbb{E}\left(Y|U_{i_1},\ldots,U_{i_k}\right)\right) = \operatorname{Var}\left(\mathbb{E}\left(Y|X_{i_1},\ldots,X_{i_k}\right)\right) \tag{8}$$

then the equality of the Sobol indices w.r.t. the  $U_i$  and to the  $X_i$ .



#### Probabilistic interpretation of the Hoeffding decomposition

Let's suppose, without loss of generality, that the  $X_i$  are in [0,1]. Then, using the Hoeffding decomposition (4), we have :

$$Y = f(\underline{X}) = f_0 + \sum_{i=1}^{n-1} f_i(X_i) + \sum_{1 \le i < j \le n} f_{i,j}(X_i, X_j) + \dots + f_{1,\dots,n}(X_1, \dots, X_n)$$
(9)

The orthogonal condition (5) of the  $f_{i_1,...,i_k}$  w.r.t. the Lebesgue measure on  $[0,1]^n$  can be interpreted as an expectation calculus if the  $X_i$  are independent.

 $\Longrightarrow$  We suppose now that the  $X_i$  are independent.

Conclusion: Y can be decomposed as:

$$Y = f(\underline{X}) = Z_0 + \sum_{i=1}^{n} Z_i + \sum_{1 \le i < j \le n} Z_{i,j} + \dots + Z_1, \dots, n$$
 (10)

where  $Z_0 = cst$  et  $Z_{i_1,\ldots,i_s} \perp Z_{j_1,\ldots,j_k}$  (ie  $\mathbb{E}\left(Z_{i_1,\ldots,i_s},Z_{j_1,\ldots,j_k}\right) = 0$ ).



#### Calculus of the Sobol indices

From the probabilistic decomposition (10), we calculate  $\mathbb{E}(Y)$  and  $\mathsf{Var}(Y)$  :

$$\left\{ \begin{array}{l} \mathbb{E}\left(Y\right) = Z_0 + \sum_{i=1}^{i=n} \mathbb{E}\left(Z_i\right) + \sum_{1 \leq i < j \leq n} \mathbb{E}\left(Z_{i,j}\right) + \dots + \underbrace{\mathbb{E}\left(Z_1, \dots, n\right)}_{=0 \text{ since } \bot Z_0} \\ \mathbb{E}\left(Y^2\right) = \sum_{l \neq J} \mathbb{E}\left(Z_l Z_J\right) + \sum_{l} \mathbb{E}\left(Z_l^2\right) \sum_{l} \mathbb{E}\left(Z_l^2\right) \end{array} \right. = 0 \text{ since } \bot Z_0$$

$$\Rightarrow \operatorname{Var}(Y) = \sum_{i=1}^{i=n} V_i + \sum_{1 \le i < j \le n} V_{i,j} + \dots + V_{1,\dots,n}$$
(11)

where  $V_{i_1,\ldots,i_k} = \mathsf{Var}\left(Z_{i_1,\ldots,i_k}\right) = \mathsf{Var}\left(f_{i_1,\ldots,i_k}(X_{i_1},\ldots,X_{i_k})\right)$ .



## An example

### Data base analysis of aerodynamical coefficients

#### Data

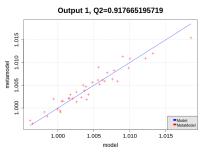
- We focus on a black box from  $\mathbb{R}^{24}$  into  $\mathbb{R}^{12}$
- We only know that function through a data base of size n = 377
- · We have no information on the distribution followed by the input vector
- · The objective is to identify, for each output component, the most influential inputs
- We only show the analysis on the first component.

#### How to proceed?

- $\bullet$  We tested the independence hypothesis of the input using the Spearman coefficients : we can't reject the hypothesis with a level 95%
- We built a meta model between the output and the inputs, using the penalized chaos polynomial expansion: the model is built from 90% of the data base and tested on the remaining 10%
- We exploit the model to calculate the Sobol indices (total and of order 1).



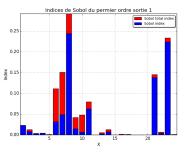
# Quality of the meta-model



Model validation



12 / 28



Input contributions to the variance of the output

We notice that it seems important to keep the inputs 6, 7, 8, 11, 21 et 23, and it is very likely that we can remove the inputs 3, 4, 5, 12, 13, 15, 16, 17, 18, 19, 20, 22 et 24 from the study. Doing that, we divided by 2 at least the input dimension.



### Historical measures

Sobol indices were introduced by Sobol in 2001 ([5]). But sensitivity indices were already existing!:

- SRC. SRRC indices
- Pearson, Spearman, PCC, PRCC indices
- importance factors from the Taylor decomposition

When the components  $X_i$  are independent, these indices are exactly particular cases of Sobol indices.

### If the model f is linear w.r.t. the $X_i$ : SRC

If  $Y = \alpha_0 + \sum_i \alpha_i X_i$ , with  $X_i$  independent, then we define the Standard Regression Coefficient (SRC) :

$$SRC_i = \frac{\alpha_i^2 \text{Var}(X_i)}{\text{Var}(Y)}$$
 (12)

Then SRC is the Sobol index of order 1 of  $X_i$ :  $SRC(Y/X_i) = S_i(Y/X_i)$ .

### If the model f is linear w.r.t. the $X_i$ : Pearson

If  $Y = \alpha_0 + \sum_i \alpha_i X_i$ , then we define the Pearson correlation between Y and  $X_i$  as:

$$\rho(Y, X_i) = \frac{\operatorname{cov}[Y, X_i]}{\sqrt{\operatorname{Var}(X_i)\operatorname{Var}(Y)}} = \frac{\mathbb{E}\left([Y - \mathbb{E}(Y)][X_i - \mathbb{E}(X_i)]\right)}{\sqrt{\operatorname{Var}(X_i)\operatorname{Var}(Y)}}$$
(13)

Moreover, if the  $X_i$  are independent, we show that :

$$\rho(Y, X_i) = \frac{\alpha_i \text{Var}(X_i)}{\sqrt{\text{Var}(X_i) \text{Var}(Y)}} \Longrightarrow (\rho(Y, X_i))^2 = SRC_i = S_i(Y/X_i)$$

Anne Dutfov Sensitivity analysis 14 / 28

## Historical measures

#### If the model rank(f) is linear w.r.t. the $rank(X_i)$ : SRRC

If  $Y = f(\underline{X})$  with  $X_i$  independent, with  $\underline{U} = (F_1(X_1), \dots, F_n(X_n))^t) = \phi^{-1}(\underline{X})$ , we have  $Z = F_Y(Y) = F_Y \circ f \circ \phi(U)$ .

If we assume in addition that

$$Z = \alpha_0 + \sum_i \alpha_i U_i \tag{14}$$

then we define the Standard Rank Regression Coefficient (SRRC):

$$SRRC(Y/X_i) = SRC(Z/U_i) = \frac{\alpha_i^2 Var(U_i)}{Var(Z)} = S_i(Z/U_i)$$

Then **SRRC** is a **Sobol** index of order 1 calculated on the ranks of  $X_i$  and Y.

### If the model rang(f) is linear w.r.t. the ranks $rang(X_i)$ : Spearman

If we assume that (14), we define the rank Spearman correlation between Y and  $X_i$  as :

$$\rho_S(Y, X_i) = \rho(F_Y(Y), F_i(X_i))$$

As previously, we show that in the case of independent variables :

$$(\rho_S(Y,X_i))^2 = SRRC(Y/X_i) = SRC(Z/U_i) = S(Z/U_i)$$



## Historical measures

Importance factors from the Taylor decomposition have been defined in metrology first where:

- Y = f(X)
- $\underline{X}$  is a gaussian vector with independent components with low variation coefficient  $(\sigma/\mu\ll 1)$
- $\implies$  f is linearized at  $\mathbb{E}(X)$

#### Taylor approximation of order 1 at $\mathbb{E}(X)$

Y = f(X) is approximated by its Taylor approximation of order 1 at  $\mathbb{E}(X)$ :

$$Y = f[\mathbb{E}(\underline{X})] + \langle \underline{\nabla} f[\mathbb{E}(\underline{X})], \underline{X} - \mathbb{E}(\underline{X}) \rangle = f[\mathbb{E}(\underline{X})] + \sum_{i} [X_{i} - \mathbb{E}(\underline{X}_{i})] \left. \frac{\partial f}{\partial X_{i}} \right|_{\mathbb{E}(\underline{X})}$$
(15)

Under the assumption of a linear model at  $\mathbb{E}(\underline{X})$ , and independent  $X_i$ , we have :

$$\operatorname{Var}(Y) = \sum_{i} \left( \frac{\partial f}{\partial X_{j}} \Big|_{\mathbb{E}(\underline{X})} \right)^{2} \operatorname{Var}(X_{i})$$
 (16)

We define the importance factor of  $X_i$ :

$$FI(X_i) = \left(\frac{\partial f}{\partial X_j}\Big|_{\mathbb{E}(X)}\right)^2 \frac{\text{Var}(X_i)}{\text{Var}(Y)} = SRC(Y/X_i) = S_i(Y/X_i)$$

The FI are Sobol indices of order 1.



Anne Dutfov Sensitivity analysis

## Sommaire

- 1 Independent Variables
- 2 Dependent variables Taylor decomposition ANCOVA Indices
- 3 Extensions
- Reference



17 / 28

# Taylor decomposition

In the case of dependent  $X_i$ , we take into account the covariance matrix only in order to calculate:

- the importance factors from the Taylor decomposition
- the ANCOVA indices

#### Taylor decomposition

Y = f(X) is approximated by its Taylor approximation of order 1 at  $\mathbb{E}(X)$ :

$$Y = f[\mathbb{E}(\underline{X})] + \langle \underline{\nabla} f[\mathbb{E}(\underline{X})], \underline{X} - \mathbb{E}(\underline{X}) \rangle = f[\mathbb{E}(\underline{X})] + \sum_{i} [X_{i} - \mathbb{E}(X_{i})] \left. \frac{\partial f}{\partial X_{j}} \right|_{\mathbb{E}(\underline{X})}$$
(17)

Under the assumption of a linear model at  $\mathbb{E}(\underline{X})$ , we have :

$$\operatorname{Var}(Y) = {}^{t}\underline{\nabla} f[\mathbb{E}(\underline{X})].\underline{\underline{\operatorname{Cov}}}[\underline{X}].\underline{\nabla} f[\mathbb{E}(\underline{X})] = \sum_{i,i} \left. \frac{\partial f}{\partial X_{i}} \right|_{\mathbb{E}(\underline{X})} \operatorname{Cov}[X_{i}, X_{j}]. \left. \frac{\partial f}{\partial X_{i}} \right|_{\mathbb{E}(\underline{X})}$$
(18)

We define the importance factor of  $X_i$  as :

$$FI(X_{i}) = \frac{\left(\sum_{j} \frac{\partial f}{\partial X_{j}} \Big|_{\mathbb{E}(\underline{X})} \operatorname{Cov}[X_{i}, X_{j}]\right) \frac{\partial f}{\partial X_{i}} \Big|_{\mathbb{E}(\underline{X})}}{\operatorname{Var}(Y)}$$
(19)



## **ANCOVA** indices

The ANCOVA (ANalysis of COVAriance) method, is a variance-based method generalizing the ANOVA (ANalysis Of VAriance) decomposition for models with correlated input parameters (see [2]). It is based on the Hoeffding decomposition of f that writes :

$$Y = f(x_1, ..., x_n) = f_0 + \sum_{U \subset \{1, n\}} f_U(\underline{X}_U)$$
 (20)

where U is a non empty set of indices in  $\{1, n\}$ . Thus  $f_U(\underline{X}_U)$  is the combined contribution of  $X_U$  to Y.

#### Definition

The total part of variance of Y due to  $X_{II}$  writes :

$$S_U = \frac{\mathsf{Cov}\left(Y, f_U(\underline{X}_U)\right)}{\mathsf{Var}(Y)} = S_U^1 + S_U^2$$

where

$$\begin{cases} S_U^1 & = & \frac{\mathsf{Var}\left(f_U(\underline{X}_U)\right)}{\mathsf{Var}\left(Y\right)} \\ \\ S_U^2 & = & \frac{\mathsf{Cov}\left(f_U(\underline{X}_U), \sum_{V|V\cap U=\emptyset} f_V(\underline{X}_V)\right)}{\mathsf{Var}\left(Y\right)} \end{cases}$$

 $S_U^1$  is the contribution to Var(Y) of  $X_U$ .

 $S_{II}^{1}$  is the contribution to Var (Y) of  $X_{II}$  through its correlation to the other variables.

er

## Sommaire

- 1 Independent Variables
- 2 Dependent variables
- 3 Extensions
  Indices based on the Csiszar divergence
- 4 References



# Csiszar Divergence

Principle: The sensitivity of Y w.r.t.  $X_i$  is no more defined as the part of the variance of Y due to the variance of  $X_i$ . We use a notion of distance between the real dependence between Y and  $X_i$ , and the independence. We assume that Y and  $X_i$  are scalar to ease the notations of this presentation.

#### Indices based on the Csiszar divergence

In [1] and [4], the authors compare the distribution of  $(X_i, Y)$ , with pdf  $p_{X_i, Y}$  to the product distribution of  $X_i$  and Y (which assumes the independence), with pdf  $p_Y \otimes p_{X_i}$ .

They define some sensitivity indices based on the Csiszar divergence  $D_f$  as:

$$S_i^f = D_f(p_{Y \otimes X_i} || p_{(Y,X_i)})$$

We show that this index :

- depends on the whole distribution and not on its first moments only
- is independent of the margins (and then of the scale of the components)

This index depends on the copula only as :

$$S_i^f = D_f(\Pi \| c_{(Y,X_i)})$$

<u>Recall</u>: The copula of (X, Y) is the same as the copula of (f(X), g(Y)) where f and g increasing functions. In particular, we can consider the uniform margins with  $f = F_X$  et  $g = F_Y$ .



# CsiszarDivergence

#### Définition

([3]) Let P and Q be two probability measures defined on the space  $\Omega$  and f a convex positive function defined at least on  $\mathbb{R}^+$  such that f(1) = 0.

The f-Csiszar divergence of Q w.r.t. P is defined as :

• If P and Q are absolutely continuous w.r.t. the Lebesgue measure dx, with pdf p and q, and if  $P \ll Q$ , then .

$$D_f(P||Q) = \int_{\Omega} f\left(\frac{p(x)}{q(x)}\right) q(x) dx \quad \in [0, +\infty]$$
 (21)

• If P and Q are absolutely continuous w.r.t. the counting measure defined on the  $(x_k)_{k\in\mathbb{N}}$  (Dirac) and if  $P \ll Q$ . then:

$$D_f(P||Q) = \sum_{k=0}^{\infty} f\left(\frac{p(x_k)}{q(x_k)}\right) q(x_k)$$
 (22)

Recall :  $P \ll Q$  means  $q(x) = 0 \Longrightarrow p(x) = 0$ 



Sensitivity analysis Anne Dutfov

## Examples

| Name               | Formula   | Generator $f(u)$            | $f(0)+f^*(0)$ |
|--------------------|---|-----------------------------|---------------|
| Total Variation    | $\frac{1}{2}\int  p(x)-q(x) dx$   | $\frac{1}{2} u-1 $          | 1             |
| Kullback-Liebler   | $\int p(x) \log \frac{p(x)}{q(x)} dx$   | $-\log u$                   | $\infty$      |
| Hellinger (square) | $\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$ $\int \frac{(p(x) - q(x))^2}{p(x)} dx$ | $\left(\sqrt{u}-1\right)^2$ | 2             |
| Chi-2 Pearson      | $\int \frac{(p(x) - q(x))^2}{p(x)} dx$  | $(u-1)^2$                   | $\infty$      |

where  $f^*: u \mapsto uf(1/u)$  the function \*-conjugate of f

## **Properties**

- Uniqueness :  $\forall (P,Q), D_{f_1}(P||Q) = D_{f_2}(P||Q) \Leftrightarrow \exists c \in \mathbb{R}, f_1(u) f_2(u) = c(u-1)$ 
  - The divergences  $D_{f_1}$  and  $D_{f_2}$  quantify the gaps between the distributions exactly the same way when  $f_1$  and  $f_2$  differ from a linear function of (u-1)
  - The divergences based on Kullback-Liebler and Hellinger are different
- Symmetry:  $\forall (P,Q), D_f(P||Q) = D_{f*}(Q||P)$  and  $\forall (P,Q), D_{f^*}(P||Q) = D_f(P||Q) \Leftrightarrow \exists c \in \mathbb{R}, f^*(u) - f(u) = c(u-1)$
- Range:  $0 = f(1) < D_f(P||Q) < f(0) + f^*(0)$
- Convexity:  $\forall \lambda \in [0,1], \quad D_f(\lambda P_1 + (1-\lambda)P_2||\lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D_f(P_1||Q_1) + (1-\lambda)D_f(P_2||Q_2)$

Anne Dutfov Sensitivity analysis

## Csiszar Divergence

The sensitivity index writes :

$$S_{i}^{f} = D_{f}(\Pi \| c_{(Y,X_{i})}) = \int_{[0,1]^{2}} f\left(\frac{1}{c_{X_{i},Y}(u,v)}\right) c_{X_{i},Y}(u,v) du dv = \int_{[0,1]^{2}} f^{*}\left(c_{X_{i},Y}(u,v)\right) du dv$$

#### How to interpret these indices?

- If Y \(\preceq X\_i\) then S<sub>i</sub><sup>f</sup> = 0 (equivalence if f is strictly convex). In that case, X<sub>i</sub> can be removed from the study since it has no impact on Y.
- If  $Y = f(X_i)$  then  $S_i^f = f(0) + f^*(0)$ .

The characterization of the range of the indices is the main result for the dependence analysis.

#### Methodology and numeric issues

Works are in progress on the following challenges:

- how to interpret the value of the indice?  $S_i^f = 0.8$ : if Sobol index, it means that 80% of the variance of Y is explained by the variance of  $X_i$ ... but what if Csiszar divergence?
- which f to consider? If  $S_i^f > S_j^f$ , do we still have  $S_i^g > S_j^g$ ? Answer: no... thus, the hierarchization depends on f. We have to  $adapt\ f$  to the needs of the study. For example, if  $c(x_i, y)$  is low in some particular zones, we take a f which increases the gaps to 1 in that zone. We need to build a know-how!
- how to estimate a copula density  $c_{(Y,X_i)}: \hat{S}_i^f = S_i^f(\hat{c}) \Longrightarrow$  use of the Bernstein copula?
- how to create independence tests based on an estimation of  $S_i^f$ , according to f?: under the independence assumption, which confidence interval do we have on the values of  $\hat{S}_i^f$ ?

# Csiszar Divergence - Independence Test

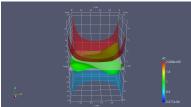
#### Proposition I

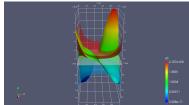
#### How to proceed:

- **1** On the sample k of  $(x_i, y)$  of size n, generated under the independence assumption between  $x_i$  and y, we build the copula density  $\hat{c}_k(x_i, y)$  of  $(x_i, y)$  thanks to the Bernstein copula;
- We repeat Step 1 N times: we draw, at any point  $(x_i, y)$ , the percentile 5% and 95% of the values of  $\hat{c}_k(x_i, y)$ , 1 < k < N;
- 3 We build 90% confidence domain point by point.

From the new sample to be tested, we build the copula density: if it goes out of the confidence domaine, then we reject the independence assumption.

Example: Copula of  $(X_{19}, Y_1)$  (left) and of  $(X_8, Y_1)$  (right)





These graphs show that we can't reject the assumption that  $Y_1$  is independent from  $X_{19}$ , while Y is clearly highly dependent on  $X_8$ .

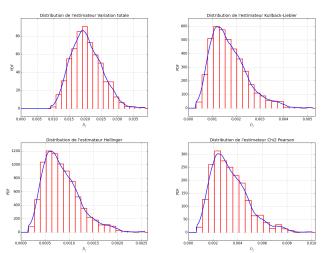


# Csiszar Divergence - Independence Test

#### Proposition II

According to the previous procedure, we calculate  $\hat{S}_i^f = S_i^j(\hat{c})$  for each f and we determine a distribution of  $\hat{S}_i^f$  and a confidence interval under the independence assumption.

Example : Estimation of sensitivity indices, n = 1000,  $N = 10^4$ .





26 / 28

# Sommaire

- 1 Independent Variables
- 2 Dependent variables
- 3 Extensions
- 4 References



## Références I



Emanuele Borgonovo and Elmar Plischke.

Sensitivity analysis : A review of recent advances.

European Journal of Operational Research, 248(3):869-887, 2016.



Yann Caniou.

Global sensitivity analysis for nested and multiscale modelling.

Theses, Université Blaise Pascal - Clermont-Ferrand II, November 2012.



Imen Csiszár.

Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der egodizität von markoffschen ketten.

Publ. Math. Inst. Hungar. Acad. Sci., 8:85-107, 1963.



Sébastien Da Veiga.

Global Sensitivity Analysis with Dependence Measures.

working paper or preprint, November 2013.



I. M. Sobolá.

Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Math. Comput. Simul.*, 55(1-3):271–280, February 2001.

