Uncertainty propagation and sensitivity analysis – Central dispersion

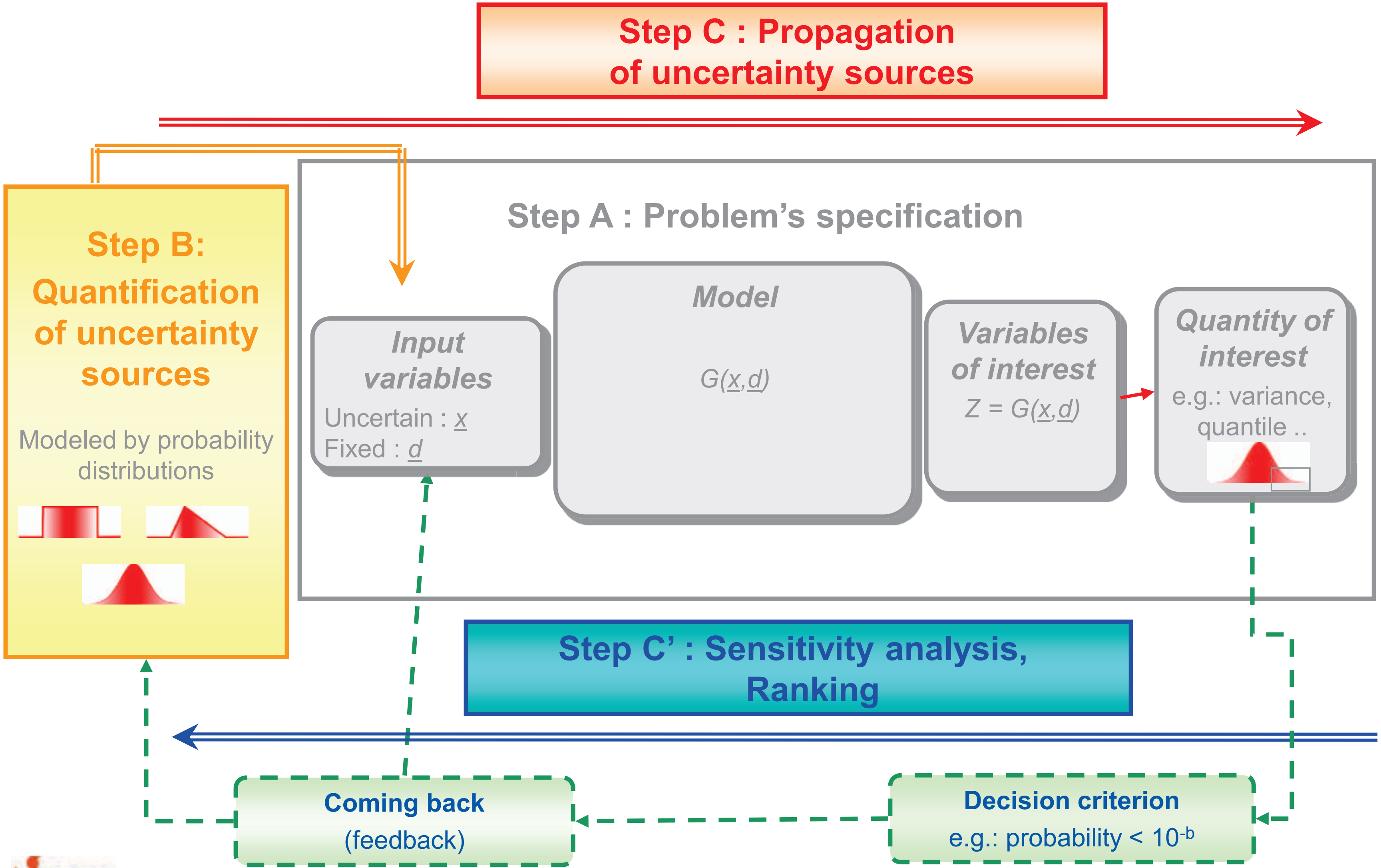
HPC and Uncertainty Treatment – Examples with OpenTURNS and Uranie
EDF – Phimeca – Airbus Groupe – CEA - IMACS

Thibault Delage, EDF R&D

PRACE Advanced Training Center

02 may 2017

UNCERTAINTY MANAGEMENT - THE GLOBAL METHODOLOGY



AGENDA

1. Taylor variance decomposition method

2. Random sampling method (Monte-Carlo)

3. Sensitivity Analysis

Uncertainty propagation

- **The quantity of interest is linked with decision stakes**
  - Two types of problems in uncertainty propagation :

    - Central trend (ex. mean) or dispersion (variance)
      – Metrology

      ➡ Analytical methods (sometimes) possible

    - Extreme quantile, « failure probability »
      – « Operator » point of view → justification of a safety criterion

      ➡ Numerical methods (optimization, Monte Carlo sampling)

Taylor variance decomposition method

Concept

Idea : if the input variations around nominal values (e.g. mean) are « not too important », the output can be approximated by a Taylor decomposition

Data : - mean values of the inputs $\mathrm{E}\left[\underline{X}\right] \equiv \underline{\mu}$

- covariance matrix of the inputs

$$\mathrm{Cov}\left[X_i, X_j\right] = \mathrm{E}\left[\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)\right] \equiv C_{ij}$$

or correlation matrix

$$\rho_{\mathrm{ij}} = \mathrm{E}\left[\left(\frac{X_i - \mu_i}{\sigma_i}\right)\left(\frac{X_j - \mu_j}{\sigma_j}\right)\right]$$

Taylor response expansion

$$Z(\underline{X}) = Z(\underline{\mu}) + \sum_{i=1}^{p} \left.\frac{\partial Z}{\partial X_i}\right|_{\underline{X}=\underline{\mu}} (X_i - \mu_i)$$

$$+\frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} \left.\frac{\partial^2 Z}{\partial X_i \partial X_j}\right|_{\underline{X}=\underline{\mu}} (X_i - \mu_i)(X_j - \mu_j)$$

$$+\, o\left(\left\|\underline{X} - \underline{\mu}\right\|^2\right)$$

Mean estimation

By taking the expectation of the previous formula :

First order :

$$E\left[Z(\underline{X})\right] = Z(\underline{\mu})$$

The mean of the response is equal, at the first order, to the response assessed for mean values of the inputs

Second order

$$E\left[Z(\underline{X})\right] = Z(\underline{\mu}) + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p} C_{ij} \left.\frac{\partial^2 Z}{\partial X_i \partial X_j}\right|_{\underline{X}=\underline{\mu}}$$

Variance estimation

First order

$$Var[Z] = E\left[(Z - E[Z])^2\right] = E\left[\left(\sum_{i=1}^{p} \frac{\partial Z}{\partial X_i}\bigg|_{\underline{X}=\underline{\mu}} (X_i - \mu_i)\right)^2\right]$$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial Z}{\partial X_i}\bigg|_{\underline{X}=\underline{\mu}} \frac{\partial Z}{\partial X_j}\bigg|_{\underline{X}=\underline{\mu}} E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

$$\boxed{Var[Z] = \sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial Z}{\partial X_i}\bigg|_{\underline{X}=\underline{\mu}} \frac{\partial Z}{\partial X_j}\bigg|_{\underline{X}=\underline{\mu}} \rho_{ij}\sigma_i\sigma_j}$$

Special case of independant variables

Mean :

$$E\left[Z(\underline{X})\right] = Z(\underline{\mu}) + \frac{1}{2}\sum_{j=1}^{p} \left.\frac{\partial^2 Z}{\partial X_i^2}\right|_{\underline{X}=\underline{\mu}} \sigma_i^2$$

Variance : Summation in quadrature formula

$$\boxed{Var\left[Z\right] = \sum_{i=1}^{p}\sum_{j=1}^{p}\left(\left.\frac{\partial Z}{\partial X_i}\right|_{\underline{X}=\underline{\mu}_{ij}}\right)^2 \sigma_i^2}$$

Deterministic Sensitivity (gradient)

Variability of the i[th] input

Input ranking (independant variables)

By normalizing the previous formula :

$$1 = \sum_{i=1}^{p} \left(\left. \frac{\partial Z}{\partial X_i} \right|_{\underline{X}=\underline{\mu}} \right)^2 \left(\frac{\sigma_i}{\sigma_Z} \right)^2$$

Importance factor of the i^th input :

$$\boxed{\eta_i^2 = \left(\left. \frac{\partial Z}{\partial X_i} \right|_{\underline{X}=\underline{\mu}} \right)^2 \left(\frac{\sigma_i}{\sigma_Z} \right)^2}$$

$$\sum_{i=1}^{p} \eta_i^2 = 1$$

It is also possible to define them when variables are dependant but the interpretation is difficult

Conclusions

- ➢ The Taylor method is based on a Taylor expansion of the model response: use of partial derivatives of the model
- ➢ It uses only the first two moments of the input random variables and their correlation coefficient

→ we can not quantify the impact of the shape of the distribution

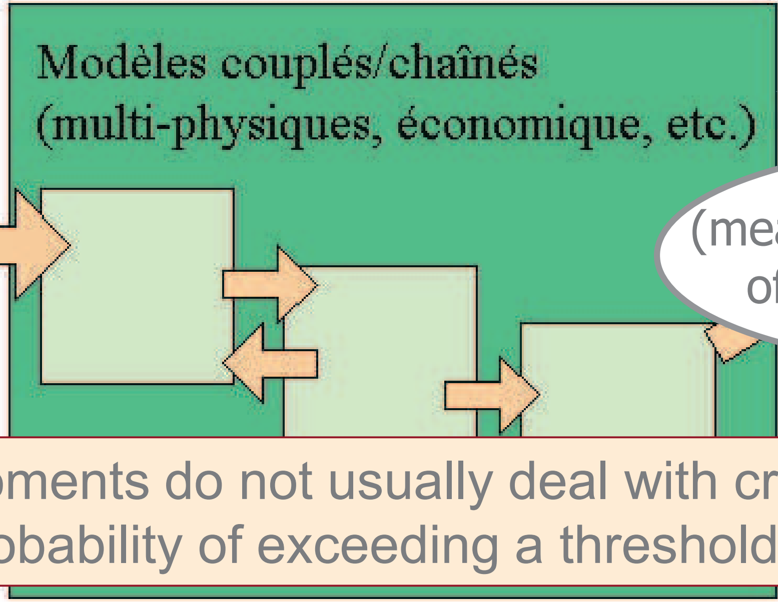
→ Monte Carlo Sampling

- ➢ Correction with order 2 is necessary if the variability of the inputs is important.

Conclusions

First order approximation is not sufficient for highly nonlinear models in the range of variation of uncertain variables

mean, variance
+ correlation



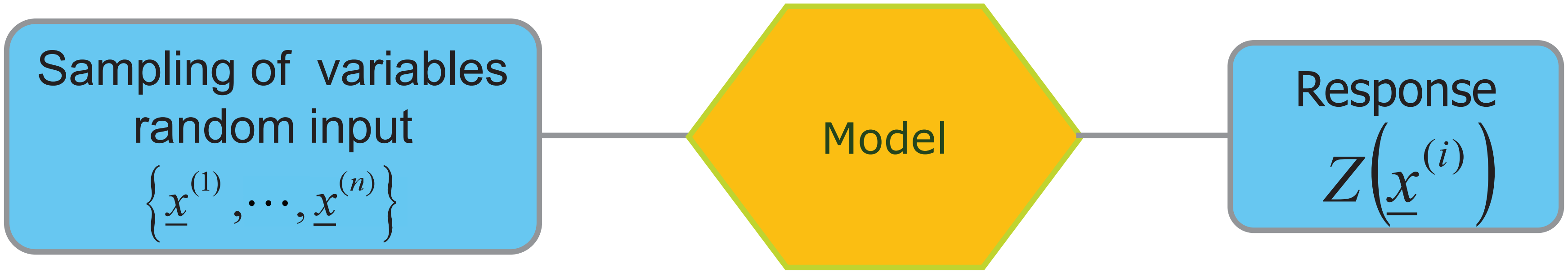
(mean, standard deviation)
of the output variables

The first two moments do not usually deal with criteria such as "probability of exceeding a threshold"

Variance decomposition if independent variables

Random Sampling method (Monte Carlo)

Concept



Distribution law of the response : estimation of the main characteristics

$$\hat{\mu}_Z = \frac{1}{n}\sum_{i=1}^{n} Z\left(\underline{x}^{(i)}\right)$$

$$\hat{\sigma}_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[Z\left(\underline{x}^{(i)}\right)-\hat{\mu}_Z\right]^2$$

$$\hat{P}_f = \frac{1}{n}\sum_{j=1}^{n} 1_{\left\{Z\left(\underline{x}^{(i)}\right)>threshold\right\}}$$

How to generate a random number?

- Simulation of a uniform r.v. $X \sim U[a,b]$

There is a generator of pseudo-random numbers $U \sim U[0,1]$. (ex : Mersenne-Twister – 1998)

We process to the transformation $X = (b-a)U + a$

- Simulation of a gaussian r.v. : Box and Müller method

$$\begin{cases} X = \sqrt{-2\ln U}\cos 2\pi V \\ Y = \sqrt{-2\ln U}\sin 2\pi V \end{cases}$$

U,V independant U [0,1]

X,Y independant N(0,1)

- Simulation of a r.v. X whose cumulative distribution fonction is F
  - Inverse method inverse : F^{-1} can be calculated (analytically or numerically)

  $U=F(X) \sim U[0,1]$ $\Rightarrow$ $X= F^{-1}(U)$

  Example : exponential law $$F^{-1}(x) = -\frac{1}{\lambda}\ln(1-x)$$

  - Acceptance-rejection method: f(x) is finite and has a bounded support

  $f(x) <= m$ and $x <= x_{max}$; we sample $U \sim U[0,x_{max}]$ and $V \sim U[0,m]$

  If $V < f(U)$ we accept U ; else we reject it

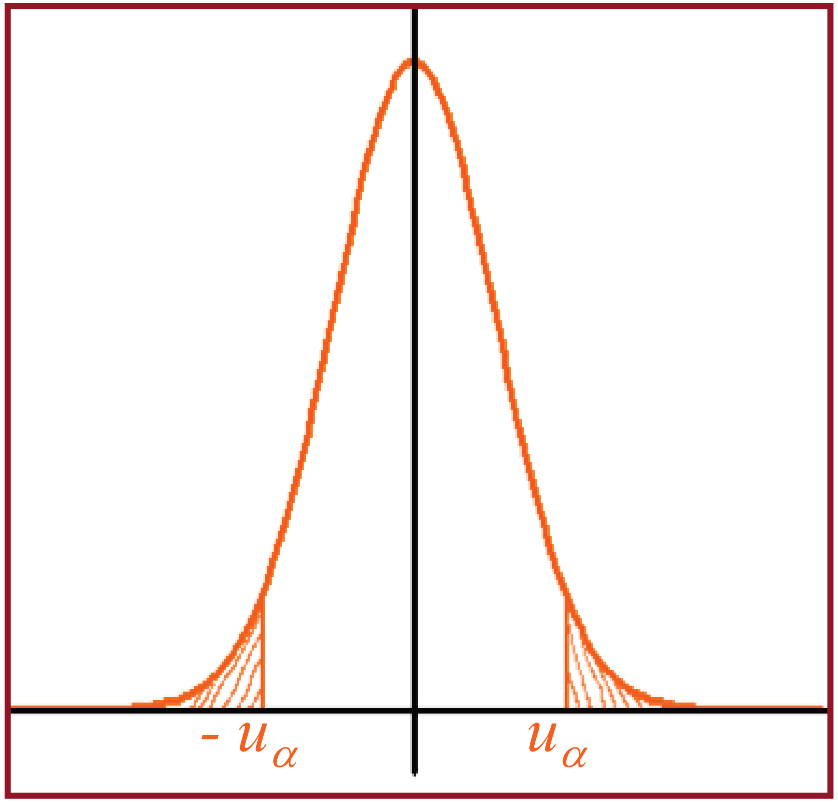
Confidence intervals

- By the central limit theorem, the previous estimators follow a **Gaussian distribution** when n→ ∞ (in practice, when **n>30**)

- When n is important, the variable $T_{n-1} = \sqrt{n-1}\,\dfrac{\hat{\mu}_Z - \mu_Z}{\hat{\sigma}_Z}$ follows approximately a centered reduced gaussian law

- We can then get a **confidence interval** for the estimator of the mean

Confidence intervals



If α is the chosen confidence level

We obtain :

$$u_\alpha : P\left(\left|T_{n-1}\right| \le u_\alpha\right) = 1 - \alpha$$

α	u_α
10%	1,645
5%	1,960
2%	2,326
1%	2,576

Confidence intervals

And the confidence interval at a 1-α level:

$$\left| \sqrt{n-1}\,\frac{\hat{\mu}_Z - \mu_Z}{\hat{\sigma}_Z} \right| \le u_\alpha$$

i.e :

$$\boxed{\hat{\mu}_Z - u_\alpha \hat{\sigma}_Z / \sqrt{n-1} \le \mu_Z \le \hat{\mu}_Z + u_\alpha \hat{\sigma}_Z / \sqrt{n-1}}$$

for a 1-α confidence level

Usual case : 1-$\alpha = 95\%$ → $u_\alpha \approx$ 2

$$\boxed{\hat{\mu}_Z - 2\hat{\sigma}_Z / \sqrt{n-1} \le \mu_Z \le \hat{\mu}_Z + 2\hat{\sigma}_Z / \sqrt{n-1}}$$

With a 95% confidence level

Confidence intervals

Confidence interval at a 1-α level of a threshold exceedance probability P_f :

$$\boxed{\hat{P}_f - u_\alpha \sqrt{\hat{P}_f / n-1} \leq P_f \leq \hat{P}_f + u_\alpha \sqrt{\hat{P}_f / n-1}}$$

Heuristically: to obtain an estimator of $P_f = 10^{-r}$ with a variation coefficient of 10%, $N = 10^{r+2}$ simulations are required

$\Rightarrow$ Use of advanced methods for rare events (FORM-SORM, accelerated Monte-Carlo)

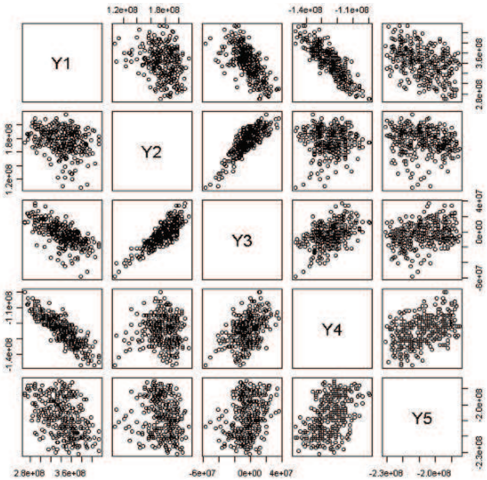
Input ranking / Sensitivity analysis

- For each input r.v. X_i and each output S, we have samples :

$$\{x_{i,1},\ldots,x_{i,n}\}, \{z_1,\ldots,z_n\}$$

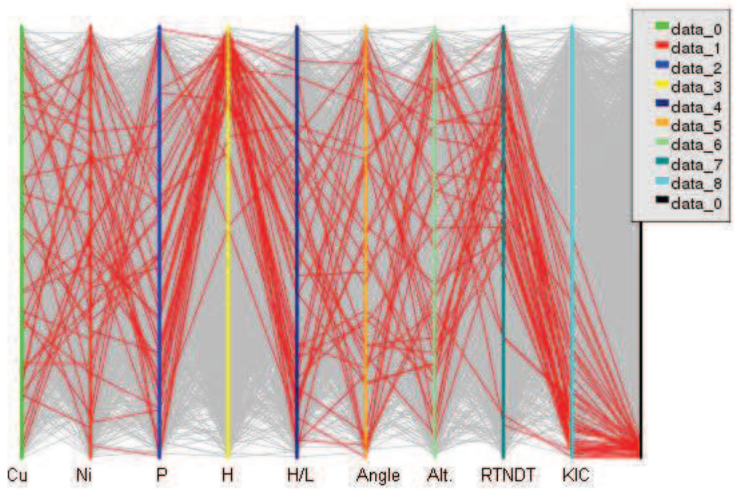
How to detect a relationship in the cloud $\{X_i, Z\}$?

- First step (<u>essential!</u>) : graphical tools



Scatter plot

(individual effect)



Cobweb plot (interaction effect)

CONCEPTS AND GOALS

Two notions :

- **sensitivity** (impatc of the model shape) $\partial Z / \partial X_i$
- **contribution** = Sensitivity x importance (the variability of the input) $\dfrac{\partial Z}{\partial X_i}\sigma(X_i)$

Objectives :

- Reduction of output uncertainty, by ranking uncertainty sources

- R&D priorization

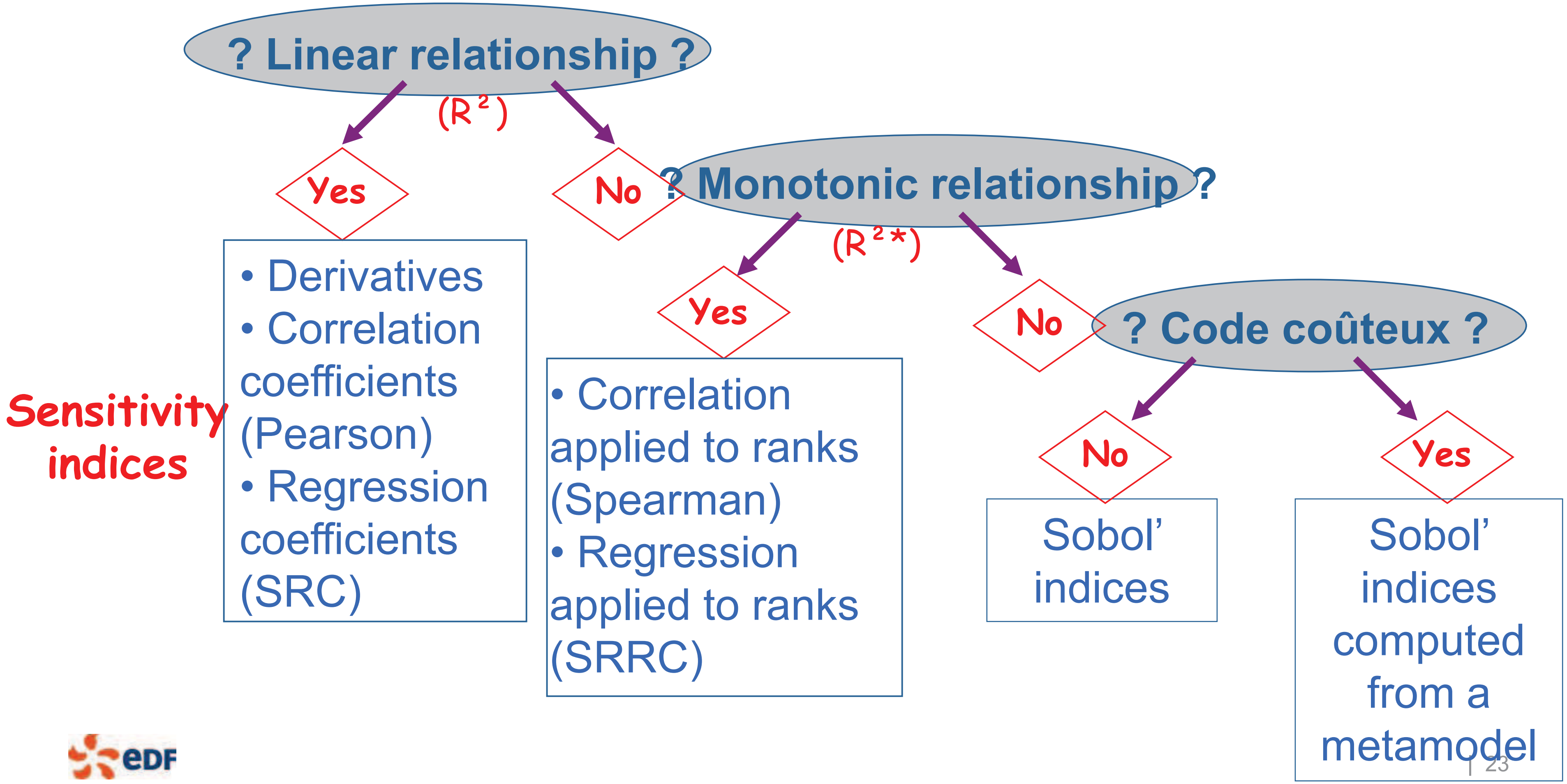
- Simplication of the model

- Set the non-influent variables to decrease the problem input dimension (e.g. useful for fitting a metamodel)

Global idea for choosing the appropriate sensitivity indices

Sample ($\boldsymbol{X}$, $Z(X)$) of size n > p , preferentially n >> p

Sensitivity analysis between inputs and output



PEARSON CORRELATION COEFFICIENT

Pearson correlation coefficient between X et Y is defined by

$$\boxed{\rho_P(X,Y)=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}=\frac{E\big((X-E(X))(Y-E(Y))\big)}{\sigma_X\sigma_Y}}$$

If equal to 1 or -1, it exists a linear relationship between X and Y

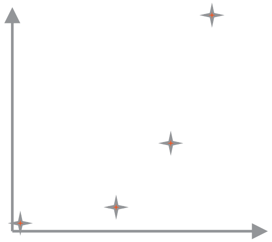
If equal to zero, X and Y are not correlated.

- **Warning : Independance of X and Y ➔ non correlation between X and Y. But the inverse can be wrong !**

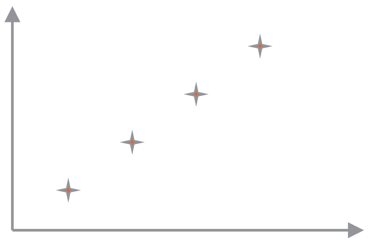
SPEARMAN CORRELATION COEFFICIENT

Instead of studying directly Xi, we focus on the rank of Xi in the sample

Exemple :



$$\begin{pmatrix} x & z \\ 4 & 4 \\ 0 & 0 \\ 5 & 10 \\ 3 & 1 \end{pmatrix}$$

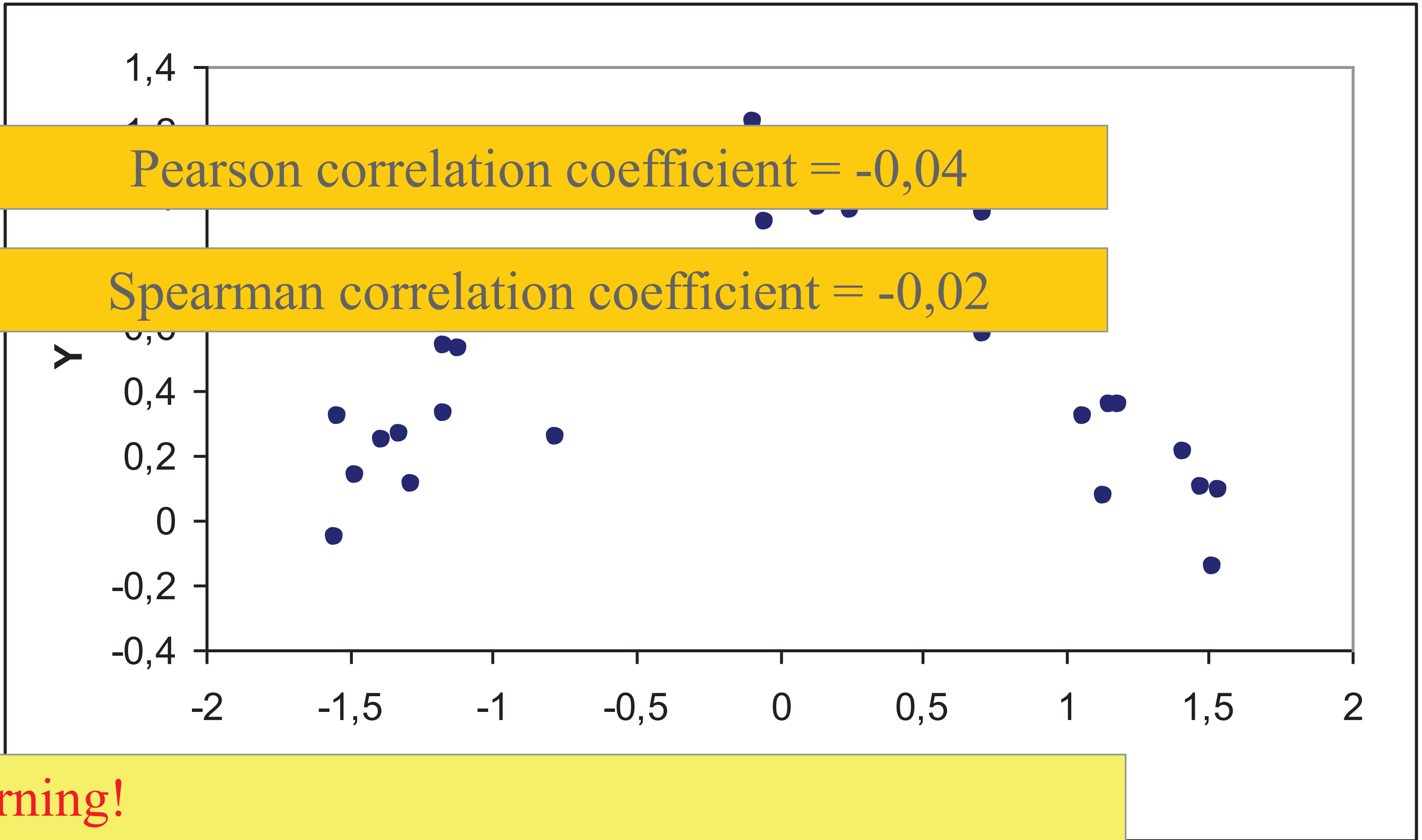


$$\begin{pmatrix} r_x & r_z \\ 3 & 3 \\ 1 & 1 \\ 4 & 4 \\ 2 & 2 \end{pmatrix}$$

$$\rho_S = \frac{\operatorname{cov}(R_X, R_Z)}{\sigma_{R_X} \sigma_{R_Z}}$$

Spearman coefficient measures the monotonicity of the relationship between X and Z

Dependances vs. Correlation



Pearson correlation coefficient = -0,04

Spearman correlation coefficient = -0,02

Warning!

Independance $\Rightarrow$ correlation coefficient equal to zero

correlation coefficient equal to zero $\cancel{\Rightarrow}$ independance

STANDARD REGRESSION COEFFICIENTS

Model : Z= β_0**+** β_1**X_1+** β_2**X_2+** β_3**X_3+** β_4**X_4**

Parameters estimated $\beta=(\beta_0,\ldots,\beta_4)$ **by least squares**

Graphical analysis – R^2 validation

Standard regression coefficient : $\mathrm{SRC}_i = \beta_i \dfrac{\sigma(X_i)}{\sigma(Z)}$

SOBOL' INDICES (NO HYPOTHESIS ON THE MODEL)

Functional ANOVA *[Efron & Stein 81]* (X_i independants) :

$$\mathrm{Var}(Y) = \sum_{i=1}^{p} V_i(Y) + \sum_{i<j}^{p} V_{ij}(Y) + \cdots + V_{12\cdots p}(Y)$$

$$\text{where } V_i(Y) = \mathrm{Var}[\mathrm{E}(Y|X_i)] \;; V_{ij} = \mathrm{Var}\left[\mathrm{E}(Y|X_i X_j)\right] - V_i - V_j \;, \ldots$$

Definition of Sobol' indices :

- Sensitivity index of order 1 : $S_i = \dfrac{V_i}{\mathrm{Var}(Y)}$
- Sensitivity index of order 2 : $S_{ij} = \dfrac{V_{ij}}{\mathrm{Var}(Y)}$
- ...

Rem. : If the model is linear, $S_i = \mathrm{SRC}^2(X_i)$

Some properties of Sobol' indices

$$1 = \sum_{i=1}^{p} S_i + \sum_i \sum_j S_{ij} + \sum_i \sum_j \sum_k S_{ijk} \ldots + S_{1,2,\ldots,k}$$

$\sum_i S_i \leq 1$ Always

$\sum_i S_i = 1$ When the model is purely additive

$1 - \sum_i S_i$ Measures the degree of interactions

Total Sensitivity Index:

[Homma & Saltelli 1996]

$$S_{Ti} = S_i + \sum_j S_{ij} + \sum_{j,k} S_{ijk} + \ldots = 1 - S_{\sim i}$$

Conclusions

- The Monte Carlo simulation is a universal tool to propagate uncertainties and perform sensitivity analyzes

- It is based on the simulation of samples (random numbers)

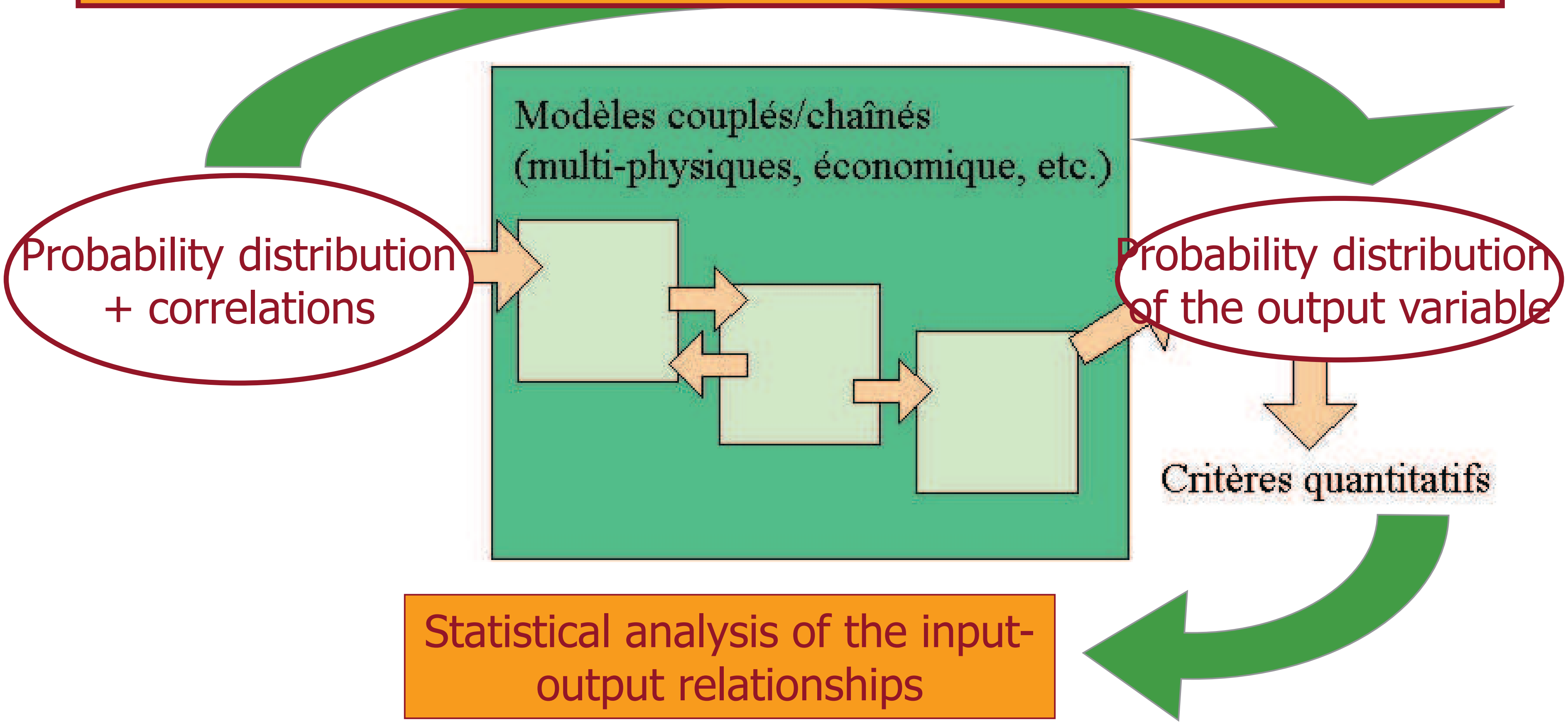
- It requires to specify the distributions of each variable and their dependance relationship

- The convergence is $1/\sqrt{n}$, where n is the number of values

- It allows to obtain a confidence interval on the result

Conclusions sur la simulation

Number of simulations very important to understand rare events! ⇒ advanced propagation methods: (FORM-SORM, accelerated Monte Carlo, metamodel...)



THANK YOU