Global sensitivity analysis based on entropy

Benjamin Auder, Bertrand Iooss

CEA Cadarache

France

Context: sensitivity analysis of complex physical phenomena simulation

Examples: simulation of nuclear power plant severe accident

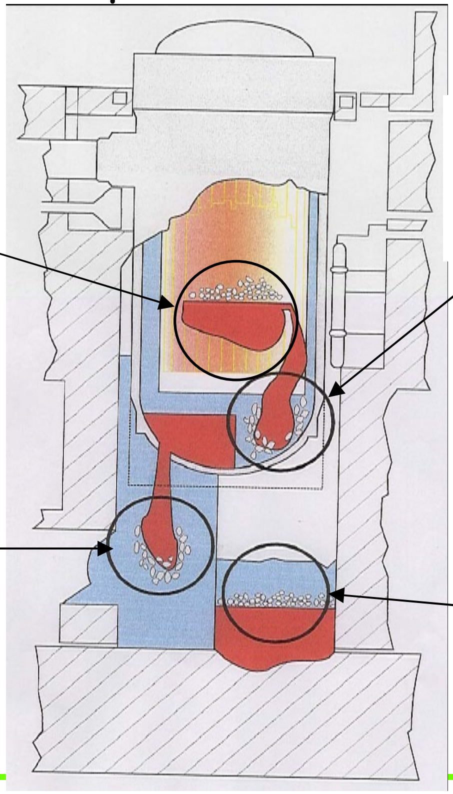
Complex and coupled numerical models which require sensitivity analysis to:

- Identify the main sources of uncertainties
- Understand the relations between uncertain inputs and interest outputs

Scenario :
Core degradation, corium transfer and interaction (inside in-vessel and ex-vessel)

23 Interest output variables :
Corium masses
Time of vessel of failure
Time of reactor pit failure

32 input random variables :
Water management, physical properties, …



Plan

- **Classical sensitivity indices**
- Entropy based sensitivity indices
    - Asymptotic properties and time complexity
- Analytic and industrial applications
- Conclusions and further prospects

Global sensitivity analysis (GSA)

•Opposed to local sensitivity analysis (LSA)
LSA studies the impact on the code output of a parameter variation around one value x_0, typically using tools of differential calculus.

Can be used to specify a GSA, which estimate the impact of the parameters on their whole range of variation → ranking input variables.

•GSA via variance-based indices (Sobol indices)
Model $y = f(x_1, \ldots, x_n)$
$x_1, \ldots, x_n$ → random variables ➔ $Y = f(X_1, \ldots, X_n)$, random variable.

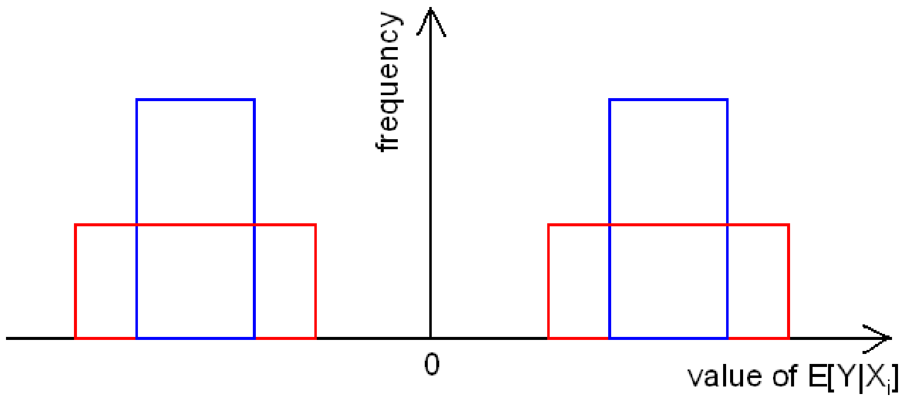
$$S_i = \frac{\mathrm{Var}(E[Y|X_i])}{\mathrm{Var}(Y)} \quad , \quad S_{T_i} = 1 - \frac{\mathrm{Var}(E[Y|X_{\sim i}])}{\mathrm{Var}(Y)}$$

Computation with Monte-Carlo or FAST method, … , or analytical derivations with a metamodel, or if explicit simple code function.

Limitations of Sobol indices

– Variance based indices ➔ only sensitive to average variations around a mean value, not much to the **shape** of the output distribution.
– Some cases where a shape intuitively gives more information :

•Geometrico-intuitive considerations ➔ same variance for red and blue.
•But intuition says blue gives more Information ➔ Y more sensitive to blue because smaller range.



Example : Bimodal distributions

•**Information** = key word. Interesting measure in information theory :
Entropy, with definitions (resp. continue and discrete) :

$$H(X) = -\int_{t \in D(X)} f(t)\ln(f(t))dt \quad , \quad H(X) = -\sum_{i=1}^{n} P(x_i)\ln(P(x_i))$$

Plan

- Classical sensitivity indices
- **Entropy based sensitivity indices**
  - Asymptotic properties and time complexity
- Analytic and industrial applications
- Conclusions and further prospects

Entropy based indices : definition

- Aim at ranking inputs, as variance does, but in a different way.
- Based on the measure of entropy ➔ different properties than variance, so at least interesting as complementary tool.

- *Idea* : As the entropy of the conditional distribution $Y|X_i$ increases, the *less* X_i affects the output Y, because entropy is :
  - **Maximal** for uniform distribution
  - **Minimal** for deterministic distribution

➔ natural definition : $\eta_i = 1 - \dfrac{H(Y|X_i)}{H(Y)}$ (Krzykacz-Hausmann, 01)

Where $H(Y|X_i) = H(E[Y|X_i])$.

Advantages :

- Adequation with intuition
- Easy to compute (linear in n for $x_1, \ldots, x_n$)

Entropy based indices : alternative definition

- Based on relative entropy or Kullback-Leibler distance between probability measures :

$$D(p:q)=\int_X p(x)\ln\frac{p(x)}{q(x)}dx$$

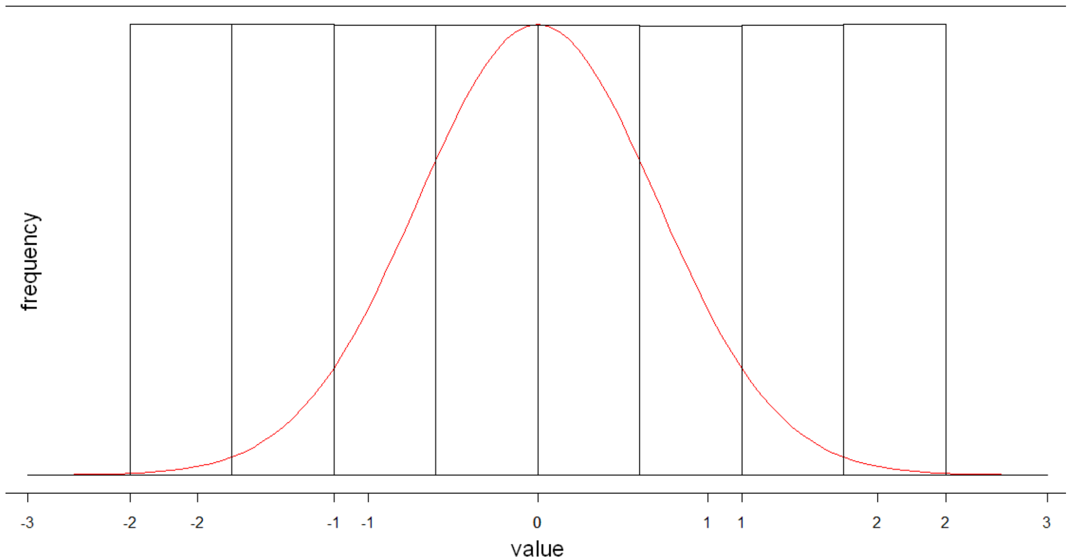
Definition by Liu & al. , 06 :

$$KL_i(p_1|p_0)=\int_{-\infty}^{+\infty} p_1(y(x_1,\dots,\overline{x}_i,\dots,x_n))\ln\frac{p_1(y(x_1,\dots,\overline{x}_i,\dots,x_n))}{p_0(y(x_1,\dots,x_i,\dots,x_n))}$$

- With p_1 and p_0 respectively the probability distributions of the model output, depending if X_i become known or not, and $\overline{x}_i=E[X_i]$
- Measure the gap between those two probability distributions
➔ $KL_i(p_1|p_0)$ = global influence of X_i on the output distribution.
- Easy to interpret but should be integrated over all $\overline{x}_i$ values to be global ➔ **LSA method**.
- full study not completed

Entropy based indices : computation

- *Go back to η_i indices now.*
- Subdivise U range in $I_1, \ldots, I_m$ partitioning intervals, and empirically compute $p_1, \ldots, p_m$ probabilities that U be in $I_1, \ldots, I_m$ respectively.



- Discrete formulation : $\eta_i = \left(\sum_{j,k} p_{jk}^{iY} \ln \frac{p_{jk}^{iY}}{p_j^i p_k^Y} \right) / H(Y)$
- *Problem* : couple density estimations difficult if too few points.

Note : formulation ➔ easy to generalize to multi-dimensional output Y.

Entropy based indices : properties, limitations

- *Convergence* :

- Equipartition $I_1, \ldots, I_m$ for the range of X_i : $\mathbf{H_m(X_i) \sim ln(m)}$

For a clear convergence of indices η_i : $n \geq m^2$

(n = number of sampling points)

- *Drawbacks* :

- Slow numerical convergence (logarithmic) due to the nature of entropy.
- Logarithmic nature ➔ bad repartition of indices in [0,1]

- *Which entropy to choose ?*
- If U and V are two random variables with the same compact support, and with almost **same discrete entropy**, they also have **same continue entropy**, and reversely ➔ choose discrete form in the algorithms.

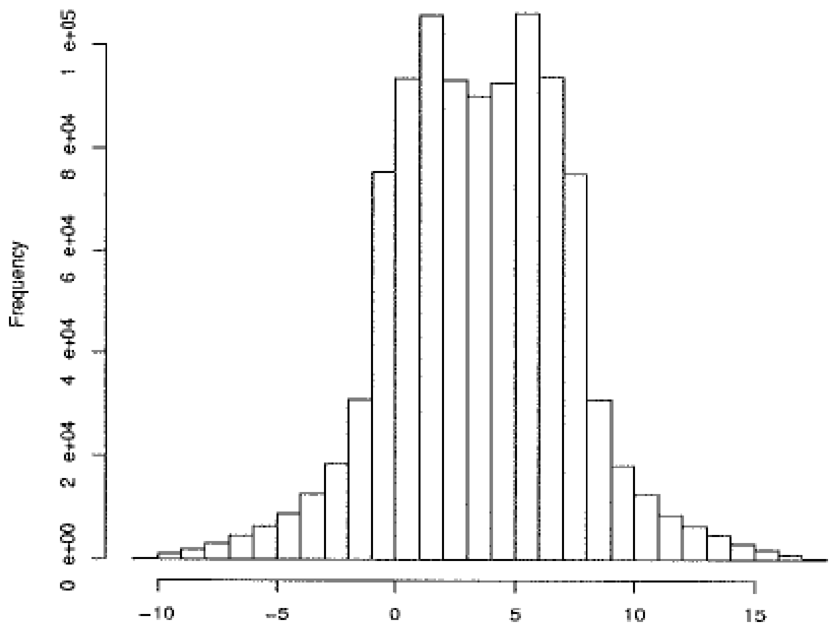
Plan

- Classical sensitivity indices
- Entropy based sensitivity indices
  - Asymptotic properties and time complexity
- **Analytic and industrial applications**
- Conclusions and further prospects

Ishigami function

- Classical benchmark in (global) sensitivity analysis.
- y : **R**³ → **R**

  $(x_1,x_2,x_3) = \sin(x_1) + a \sin^2(x_2) + b\, x_3^4 \sin(x_1)$
- Histogram of Y if X_1, X_2, X_3 uniform random variables on [-1,1] :



Bimodal distribution (here a=7 b=0.1)

= One case where entropy could be interesting.

Variance also good because bimodality is not clearly marked.

Sensitivity indices for Ishigami function

•Indices S_i and S_{Ti} : (theoretical values)

First order indices $\mathbf{S_i}$	$\mathbf{S_1 = 0.32}$	$\mathbf{S_2 = 0.44}$	$\mathbf{S_3 = 0}$
Total indices $\mathbf{S_{ti}}$	$\mathbf{S_{T1} = 0.56}$ **(45%)**	$\mathbf{S_{T2} = 0.44}$ **(35%)**	$\mathbf{S_{T3} = 0.24}$ **(19%)**

•Value for third index : 0 for 1st order Sobol indices, >0 for entropy-based and total indices :
•First order do not take interactions into account, whereas entropy based indices compute a global impact of a parameter

➔ better compare to total indices S_{Ti}.

•Indices η_i (after convergence) :

$\mathbf{\eta_1 = 0.092}$ **(34%)**	$\mathbf{\eta_2 = 0.12}$ **(45%)**	$\mathbf{\eta_3 = 0.058}$ **(21%)**

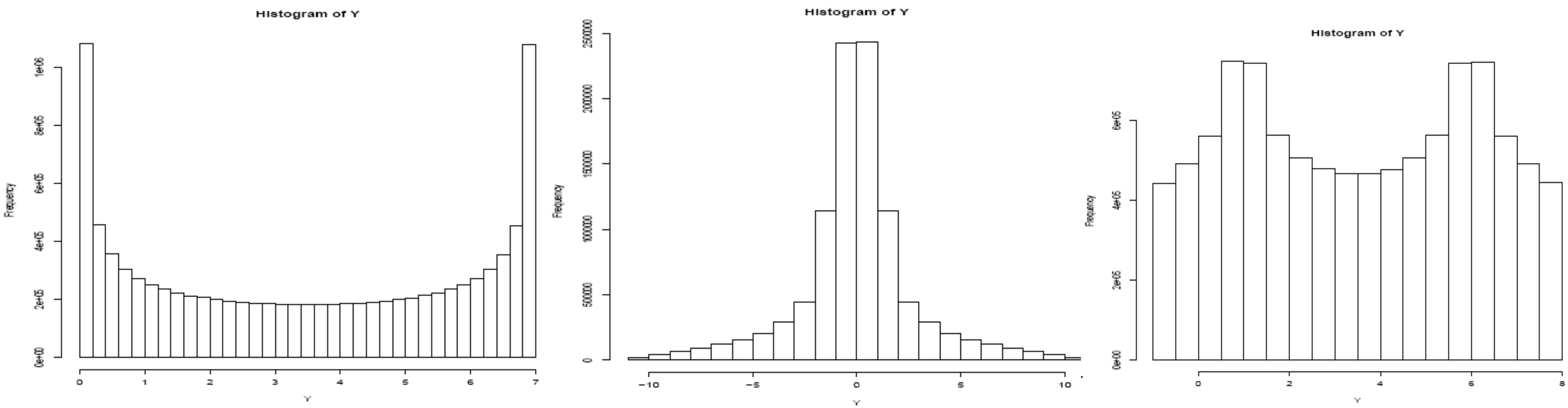
➔ **Similar results for both methods**

Kullback-Leibler based indices

•The results are quite different for the 3 sensitivity indices :

KL_1 = 2.92 (55%)	**KL_2 = 0.65 (12%)**	**KL_3 = 1.73 (33%)**

•*Explanation* : the 3 histograms of conditional outputs $Y|X_i$ (=0)

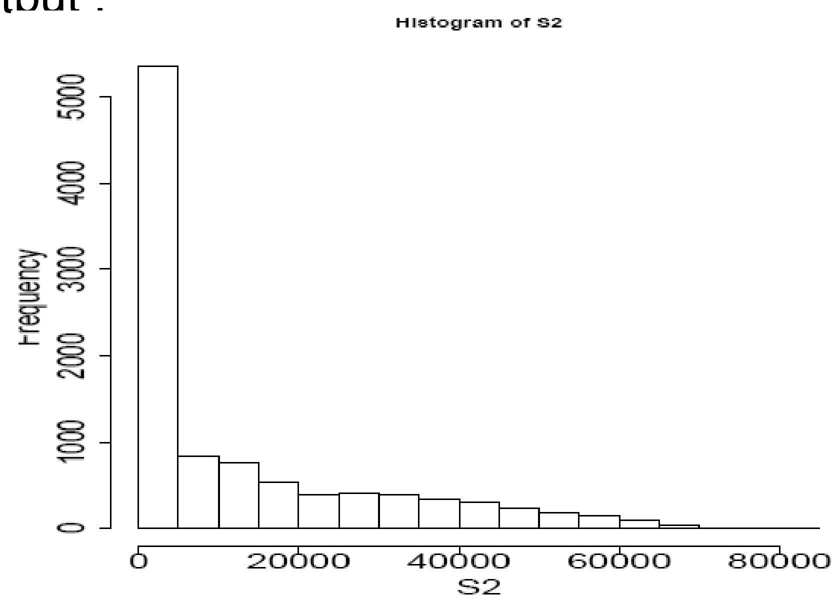


•Visually, $Y|X_1$ most different from Y (reverse) and then $Y|X_3$ (too « uniform »).

Industrial case : LEONAR physical code

•**32** inputs parameters : too many ➔ first Morris screening to select the 6 most important variables, renamed from X_1 to X_6.

•**1** studied output :



•Y : fast decreasing with heavy tail.

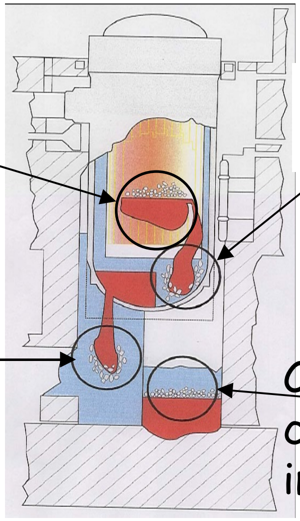
➔ Interesting to compare variance and entropy in this case.

LEONAR sensitivity indices

•*As expected, the results are different from entropy to variance-based :*

Y - SA	X_1	X_2	X_3	X_4	X_5	X_6
S_{Ti}	0.82	0.04	0.39	0.03	0.02	0.35
η_i	0.24	0.13	0.14	0.13	0.13	0.33

Y: corium mass in the vessel bottom
X_1: water arrival time in the reactor pit
X_3 : water flow rate in the pit



•In blue : logical high values shared for the two methods.

•In red : entropy-based method does not give as much importance to the variable than the Sobol indice ➔ the corresponding input does not control much the distribution tail, but affects output variablility.

•*Number of sampling points* **≈10⁵** for both methods.

Conclusions and further prospects

•*Entropy* = new point of view, complementary information added to variance. In some situations entropy can rank variables more accurately.

•Entropy-based indices would complete or precede an analysis using variance. **Study still in progress.**

Positive aspects :

- entropy naturally deals with **multi-output sensitivity analysis**.
- Deterministic sampling for entropy based indices allows fast enough calculation to get precise indices.

…*Negative* :

- Choosing the right discretisation step for input and output variables, conditioning the (slow) convergence speed.
- Slow convergence → greedey in sampling points ; solutions :
  - Metamodel to fasten the computer code
  - R & D to develop efficient algorithms

•*Next* : Do more tests to estimate the efficiency of the new indices depending on the number of simulations run.

Thanks for your attention,

Any questions ?