Sensitivity analysis in the presence of hierarchical variables, a first attempt

E. Bartók [1], J. Pelamatti [2], V. Chabridon [2]

In collaboration with S. Da Veiga[3]

[1]CentraleSupélec, L2S [2]EDF R&D [3]ENSAI, CREST

GATSBII Kick-off meeting, Toulouse

30/01/2025

Context

Sensitivity analysis (SA)

Assessing how the uncertainties on the input variables of a phenomenon affect the uncertainty of the output of interest

It can serve various purposes:

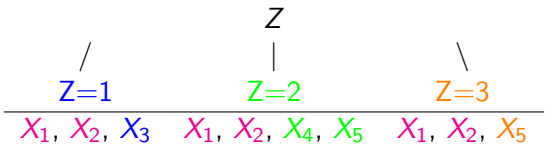
- Better comprehension of the phenomenon
- Model validation
- Model explainability
- Dimension reduction (screening)
- Uncertainty reduction

Context: hierarchical variables

Definition: A hierarchical variable Z is a random discrete variable presenting a hierarchical dependency with the other random inputs ($X := \{X_1, \ldots, X_d\} \in \mathbb{R}^d$).

Model: $Y = G(Z, X)$ or alternatively a **data set** $(Z, X, Y)_i$ for $i = 1, \ldots, n$

First intuition:



- Some variables are always *active* : X_1 et X_2.
- When a variable is not active, it presents no value (`NA`), and may usually make no physical sense

Illustrative example: housing heating cost

Type of housing Z	Windows isolation X_1	Surface X_2	Floor X_3	Wood density X_4	Roof isolation X_5	Heating cost Y
apartment	6/10	1010 ft^2	0	NA	NA	25€/month
apartment	8.5/10	753 ft^2	1	NA	NA	12€/month
apartment	4/10	322 ft^2	5	NA	NA	10€/month
wooden house	9.7/10	861 ft^2	NA	50 lb/ft^3	12 in	29€/month
wooden house	5.8/10	1200 ft^2	NA	26 lb/ft^3	6.5 in	50€/month
brick house	7.6/10	1399 ft^2	NA	NA	4 in	52€/month
brick house	4.2/10	750 ft^2	NA	NA	9 in	30€/month
⋮	⋮	⋮	⋮	⋮	⋮	⋮

$g : \{1, 2, 3\} \times \mathbb{R}^5 \to \mathbb{R}$

$g(Z, X) = g^{(1)}(X_1, X_2, X_3)\mathbb{1}_{Z=1} + g^{(2)}(X_1, X_2, X_4, X_5)\mathbb{1}_{Z=2} + g^{(3)}(X_1, X_2, X_5)\mathbb{1}_{Z=3} d$

Context: hierarchical variables

Other examples of hierarchical variables?

- System architectural choices
  - e.g., turbofan or turbopropeller engine on a plane?
- Demographical data
  - e.g., information depending on socio-economical class or location
- Conditional poll questions
  - e.g., *if you answered yes to the previous questions, could you tell us more?*
- Machine learning model architecture parameters
  - e.g., neural network layers and associated parameters and weights

Context: hierarchical variables

The literature is not consistent regarding nomenclature and definition

- Hierarchical variables
- Dimensional variables
- Branching & Nested factors
- Meta-variables
- Trigger variables

Objectives

Sensitivity analysis in the presence of hierarchical variables

- How to assess the influence of the hierarchical variables?
- How to assess the influence of the children variables?
- How to assess the influence of the interactions between the two?

- We consider a **given-data** framework: sensitivity to be performed on a simple data set.

- We focus here on **Sobol' indices** [Sob93].
  - Intuitive interpretation
  - Applicable to non-monotonous and non-linear models / phenomena
  - Input variables must be independent

Problem formalization

How to adapt Sobol' indices to the presence of hierarchical variables?

Hypotheses:

- The function of interest can be written as:

$$Y := g(Z, X) = \sum_{k=1}^{p} g^{(k)}(X) \mathbb{1}_{Z=k}$$

- $\forall k \in [\![1, p]\!]$, $g^{(k)} \in L^2(\mathbb{P}_X)$; $\quad X \in \mathbb{R}^d$
- The hierarchical variables are mutually independent, $\mathbb{P}_Z := \times_{i=1}^{\ell} \mathbb{P}_{Z_i}$;
- The children variables are mutually independent, $\mathbb{P}_X := \times_{i=1}^{d} \mathbb{P}_{X_i}$;
- Both hierarchical and children variables are mutually independent, $\mathbb{P}_{Z,X} := \mathbb{P}_Z \times \mathbb{P}_X$.
- Is there a better approach for this probabilistic modeling ?!

F-ANOVA decomposition

Sobol' indices rely on the Functional Analysis of Variance (FANOVA). We define the following projections:

- $\eta_{\varnothing} = \mathbb{E}\left[g(Z,X)\right]$
- $\eta_I(X_I) \ : \ \mathcal{X}_I \longrightarrow \mathbb{R} \ := \ \mathbb{E}\left[g(Z,X)|X_I\right] - \sum_{J \subsetneq I} \eta_J(X_J)$
- $\eta_Z(Z) \ : \ \mathcal{Z} \longrightarrow \mathbb{R} \ := \mathbb{E}\left[g(Z,X)|Z\right] - \eta_{\varnothing}$
- $\eta_{Z,I}(Z,X_I) \ : \ \mathcal{Z} \times \mathcal{X}_I \longrightarrow \mathbb{R} \ := \mathbb{E}\left[g(Z,X)|Z,X_I\right] - \sum_{J \subseteq I} \eta_J(X_J) - \sum_{J \subseteq I} \eta_{Z,J}(Z,X_J)$
- $\eta^k_{\varnothing} := \mathbb{E}\left[g^k(Z,X)\right]$
- $\eta^k_I(X_I) \ : \ \mathcal{X}_I \longrightarrow \mathbb{R} \ := \ \mathbb{E}\left[g_k(Z,X)|X_I\right] - \sum_{J \subsetneq I} \eta^k_J(X_J)$

F-ANOVA decomposition

We obtain the following orthogonal decomposition of the function of interest

$$g(Z, X) = \eta_{\varnothing} + \eta_Z(Z) + \sum_{I \subseteq \mathcal{P}} \eta_I(X_I) + \sum_{I \subseteq \mathcal{P}} \eta_{Z,I}(Z, X_I)$$

and

$$g^{(k)}(X) = \eta^k_{\varnothing} + \sum_{I \subseteq \mathcal{P}} \eta^k_I(X_I)$$

With $\mathcal{P}$ the power set of $\{1, \ldots, d\}$

F-ANOVA decomposition

We can obtain some straightforward expressions for some of the projections:

$$\eta_Z(Z) = \sum_{k=1}^{m} \mathbb{1}_{\{Z=k\}} \left(\mathbb{E}\left[g^{(k)}(X) \right] - \mathbb{E}\left[g(Z,X) \right] \right),$$

$$\eta_I(X_I) = \sum_{k=1}^{m} p_k \, \eta_I^k(X_I),$$

$$\eta_{Z,I}(Z,X_I) = \sum_{k=1}^{m} \mathbb{1}_{\{Z=k\}} \eta_I^k(X_I) - \eta_I(X_I),$$

which can be plugged into:

$$S_I := \frac{\mathrm{Var}\left(\eta_I(X_I)\right)}{\mathrm{Var}\left(Y\right)}.$$

Sobol' indices Adaptation

Interpretation of the proposed indices:

- S_z mesures the part of variance that depends on the weighted difference between the means of the submodels $g^{(k)}(X)$:

$$S_z = \frac{\sum_{k=1}^{p} p_k \left(\mathbb{E}\left[g^{(k)}(X) \right]^2 - \mathbb{E}\left[Y \right]^2 \right)}{\mathrm{Var}\left(Y \right)}.$$

Sobol' indices Adaptation

Interpretation of the proposed indices:

- S_z mesures the part of variance that depends on the weighted difference between the means of the submodels $g^{(k)}(X)$:

$$S_z = \frac{\sum_{k=1}^{p} p_k \left(\mathbb{E}\left[g^{(k)}(X)\right]^2 - \mathbb{E}\left[Y\right]^2 \right)}{\mathrm{Var}\left(Y\right)}.$$

- S_i quantifies the contribution of X_i to the variance of the submodels $g^{(k)}(X)$:

$$S_I = \frac{\mathrm{Var}\left(\sum_{k=1}^{p} p_k \eta_I^{(k)}(X_I) \right)}{\mathrm{Var}\left(Y\right)}.$$

Sobol' indices Adaptation

Interpretation of the proposed indices:

- S_z mesures the part of variance that depends on the weighted difference between the means of the submodels $g^{(k)}(X)$:

$$S_z = \frac{\sum_{k=1}^{p} p_k \left(\mathbb{E}\left[g^{(k)}(X) \right]^2 - \mathbb{E}\left[Y \right]^2 \right)}{\mathrm{Var}\left(Y \right)}.$$

- S_I quantifies the contribution of X_I to the variance of the submodels $g^{(k)}(X)$:

$$S_I = \frac{\mathrm{Var}\left(\sum_{k=1}^{p} p_k \eta_I^{(k)}(X_I) \right)}{\mathrm{Var}\left(Y \right)}.$$

- $S_{i \cup \{z\}}$, in the simple 2 submodels case:

$$S_{i \cup \{z\}} = \frac{p_1 p_2 \left(2\, \mathbb{E}\left[\left(\eta_I^1(X_I) - \eta_I^2(X_I) \right)^2 \right] \right)}{\mathrm{Var}\left(Y \right)}$$

Sobol' indices

Similarly, we can obtain the following total Sobol' indices

$$
\begin{aligned}
S_z^T &= \frac{\sum_{k=1}^{p}(p_k - p_k^2)\mathbb{E}\left[Y_k^2\right] - 2\sum_{k<l} p_k p_l \mathbb{E}\left[Y_k Y_l\right]}{\mathrm{Var}\left(Y\right)} \\
S_I^T &= \frac{\sum_{k=1}^{p} p_k \mathbb{E}\left[\mathrm{Var}\left(Y_k | I^c\right)\right]}{\mathrm{Var}\left(Y\right)} \\
S_{\{Z\}\cup I}^T &= \frac{\sum_{k=1}^{p} p_k^2 \mathbb{E}\left[\mathrm{Var}\left(Y_k | I^c\right)\right]}{\mathrm{Var}\left(Y\right)} \\
&\quad + \frac{\sum_{k=1}^{p} \mathbb{E}\left[Y_k^2\right](p_k - p_k^2) - 2\sum_{k<l} p_k p_l \mathbb{E}\left[\mathbb{E}\left[Y_k | I^c\right]\mathbb{E}\left[Y_l | I^c\right]\right]}{\mathrm{Var}\left(Y\right)}
\end{aligned}
$$

Given-data estimators

We suppose only having access to an i.i.d. sample of inputs and outputs: $(Z^{(i)}, X^{(i)}, Y^{(i)})_{1 \leq i \leq n}$.

In practice, estimating the Sobol' indices boils down to estimating

- T1 $= \mathrm{Var}\left(\mathbb{E}\left[Y_k | X_I\right]\right)$
- T2 $= \mathbb{E}\left[\mathbb{E}\left[Y_k | X_I\right] \mathbb{E}\left[Y_h | X_I\right]\right]$.

We propose two estimators for the quantity T1, inspired by the work of Broto [Bro20]:

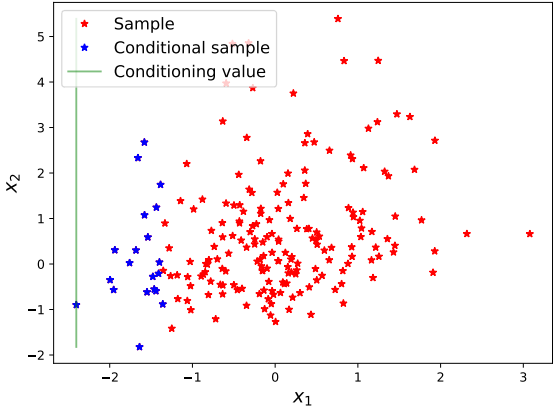
- **Double Monte Carlo** (MC): $\mathbb{E}\left[\mathrm{Var}\left(Y | X_I\right)\right] = \mathrm{Var}\left(Y\right) - \mathrm{Var}\left(\mathbb{E}\left[Y | X_I\right]\right)$;
- **Pick-and-Freeze** (P&F): $\mathbb{E}\left[\mathbb{E}\left[Y | X_I\right]^2\right] - \mathbb{E}\left[Y\right]^2 = \mathrm{Var}\left(\mathbb{E}\left[Y | X_I\right]\right)$.

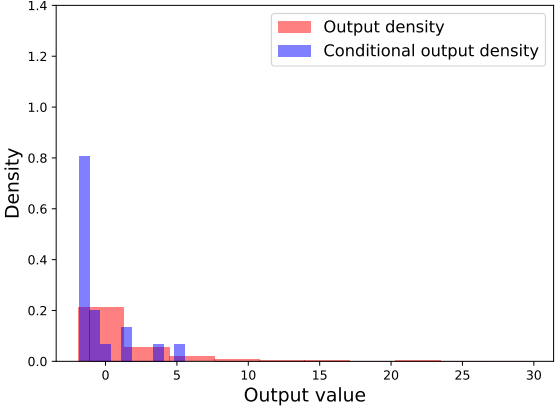
These estimators are based on a k-nearest neighbors (knn) approach

Given-data estimators

Core concept:

- The distribution of X_I is replaced by its empirical discrete distribution
- replacing conditional sampling of $Y|X_I = x_I$ with the sample containing the k nearest neighbors of x_i





Given-data estimators

Estimation of T1 $:= \mathrm{Var}\left(\mathbb{E}\left[Y_k|X_I\right]\right)$

- **Double Monte Carlo** (MC):

$$\mathbb{E}\left[\widehat{\mathrm{Var}\left(Y|X_I\right)}\right] = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(\frac{1}{K-1}\sum_{n=1}^{K}\left[g\left(X_I^{\left(knn^I_{k,k}(i,n)\right)}\right) - \frac{1}{K}\sum_{h=1}^{K} g\left(X_I^{\left(knn^I_{k,k}(i,h)\right)}\right)\right]^2\right)$$

- **Pick-and-Freeze** (P&F):

$$\widehat{\mathrm{Var}\left(\mathbb{E}\left[Y|X_I\right]\right)} = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(g\left(X^{(i,k)}\right) g\left(X_I^{\left(knn^I_{k,k}(i,2)\right)}\right)\right) - \left(\frac{1}{N_k}\sum_{i=1}^{N_k} g\left(X^{(i,k)}\right)\right)^2$$

Given-data estimators

Estimation of T2 $:= \mathbb{E}\left[\mathbb{E}\left[Y_k|X_l\right]\mathbb{E}\left[Y_h|X_l\right]\right]$

$$\widehat{T2} = \frac{1}{N_k + N_l}\left(\sum_{n=1}^{N_k} g^{(k)}\left(X^{(n,k)}\right) g^{(l)}\left(X^{(knn_{l,k}^{l\cap l_k \cap l_l}(n,1))}\right) + \sum_{n=1}^{N_l} g^{(k)}\left(X^{(knn_{k,l}^{l\cap l_k \cap l_l}(n,1))}\right) g^{(l)}\left(X^{(n,l)}\right)\right)$$

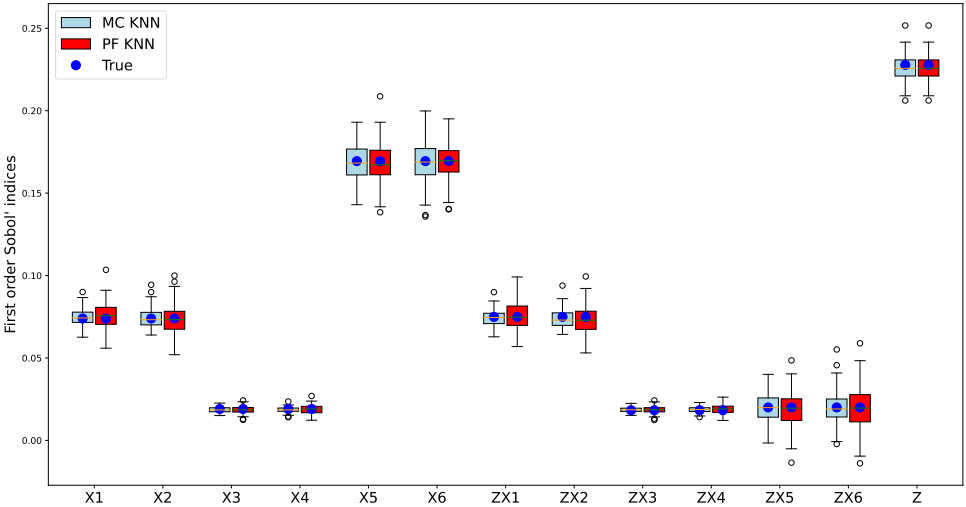
A few numerical results: first toy case

$$\begin{cases} Y_1 := g^{(1)}(X) = X_1 + X_2 + X_5 + X_6 - 1 \\ Y_2 := g^{(2)}(X) = X_3 X_4 + X_5 X_6 \\ g(Z, X) = g^{(1)}(X)\mathbb{1}_{Z=1} + g^{(2)}(X)\mathbb{1}_{Z=2} \end{cases}$$

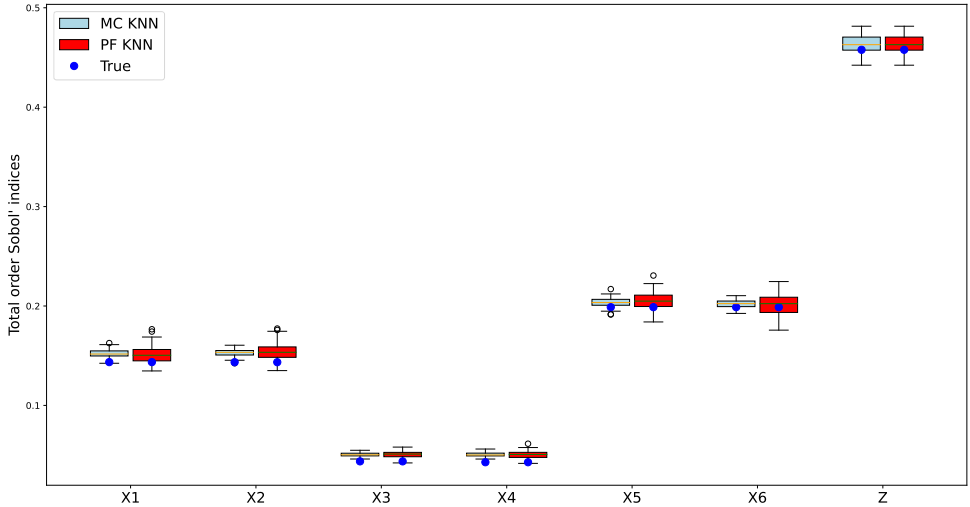
With:

- $X = (X_1, \cdots, X_6)$, $X_i \sim \mathcal{U}([0,1])$ i.i.d.
- Z : $\mathbb{P}(Z = 1) = 0.5$, $\mathbb{P}(Z = 2) = 0.5$

A few numerical results: first toy case



A few numerical results: first toy case

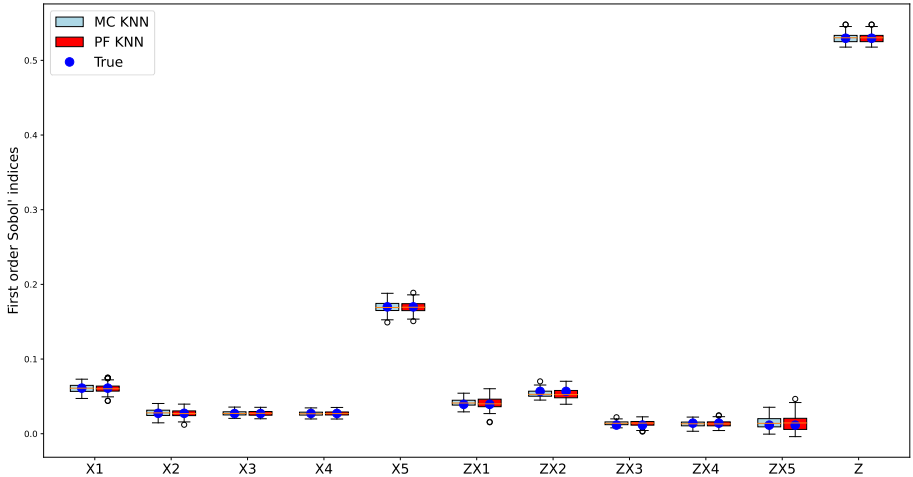


Additional test-case

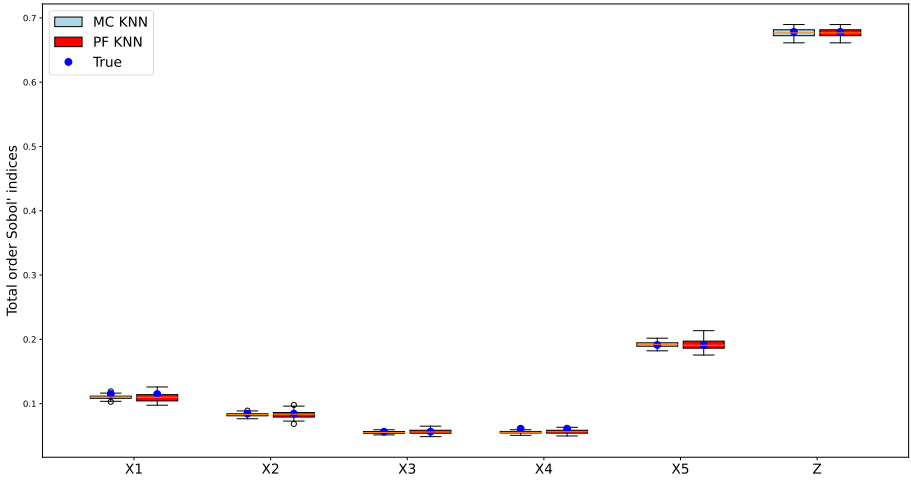
$$\begin{cases} Y_1 := G^{(1)}(X) = X_1 + X_2 + X_5 \\ Y_2 := G^{(2)}(X) = X_3 X_4 + X_5 \\ Y_3 := G^{(3)}(X) = X_3 X_4 + X_1 X_5 \\ G(Z, X) = G^{(1)}(X)\mathbb{1}_{Z=1} + G^{(2)}(X)\mathbb{1}_{Z=2} + G^{(3)}(X)\mathbb{1}_{Z=3} \end{cases}$$

Where $X = (X_1, X_2, X_3, X_4, X_5)$, $X_i \sim \mathit{Unif}([0,1])$, $i = 1, \ldots, 5$

A few numerical results: second toy case



A few numerical results: second toy case

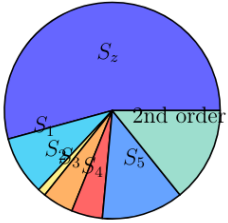


A few numerical results: second toy case

$\mathbb{P}(Z = 1) = 0.2,$

$\mathbb{P}(Z = 2) = 0.1$

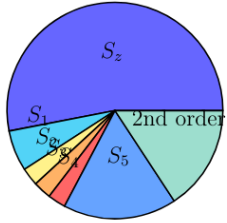
$\mathbb{P}(Z = 3) = 0.7$



$\mathbb{P}(Z = 1) = 0.33,$

$\mathbb{P}(Z = 2) = 0.33$

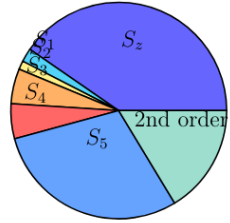
$\mathbb{P}(Z = 3) = 0.33$



$\mathbb{P}(Z = 1) = 0.2,$

$\mathbb{P}(Z = 2) = 0.7$

$\mathbb{P}(Z = 3) = 0.1$



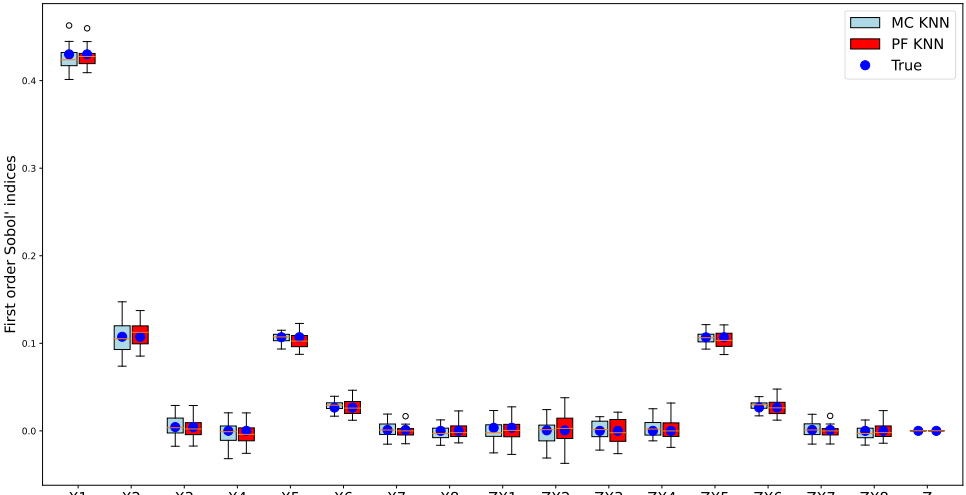
A few numerical results: the G-Sobol function

$$\begin{cases} Y_1 := g^{(1)}(X) = \prod_{i=1}^{4} \frac{|4x_i-2|+a_i}{1+a_i}, \text{ where } (a_i)_i = (0,1,9,99) \\ Y_2 := g^{(2)}(X) = \prod_{i=1}^{8} \frac{|4x_i-2|+b_i}{1+b_i}, \text{ where } (b_i)_i = (0,1,9,99,0,1,9,99) \\ g(Z,X) := g^{(1)}(X)\mathbb{1}_{Z=1} + g^{(2)}(X)\mathbb{1}_{Z=2} \end{cases}$$

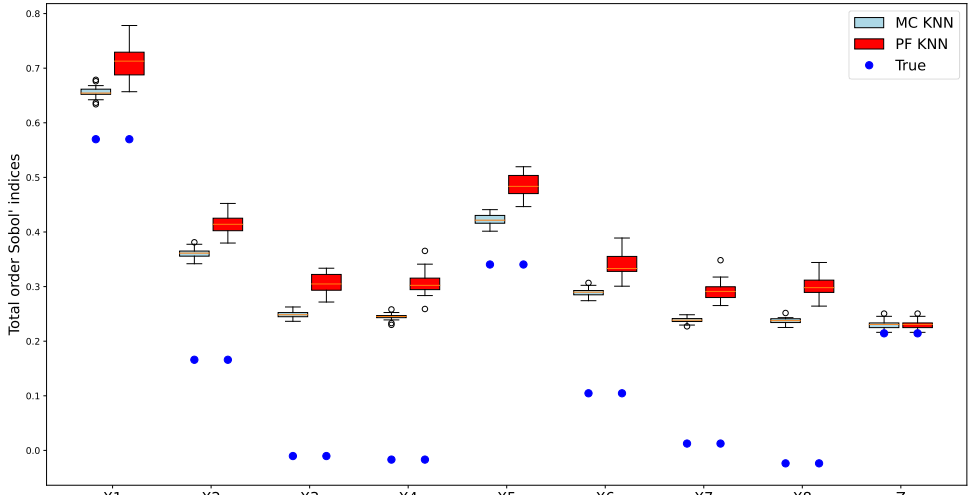
With:

- $X = (X_1, \cdots, X_8)$, $X_i \sim \mathcal{U}([0,1])$ i.i.d.
- Z : $\mathbb{P}(Z=1) = 0.5$, $\mathbb{P}(Z=2) = 0.5$

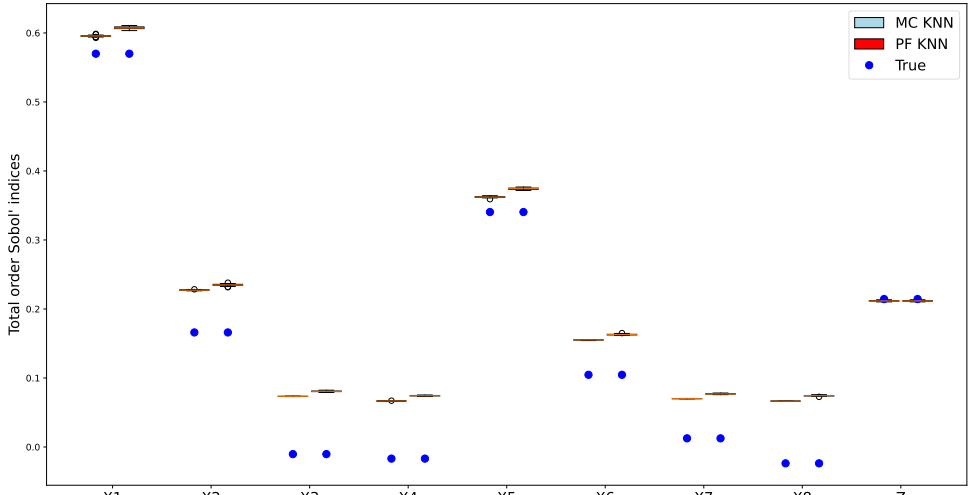
A few numerical results: modified G-Sobol function, 10000 samples



A few numerical results: modified G-Sobol function, 10000 samples

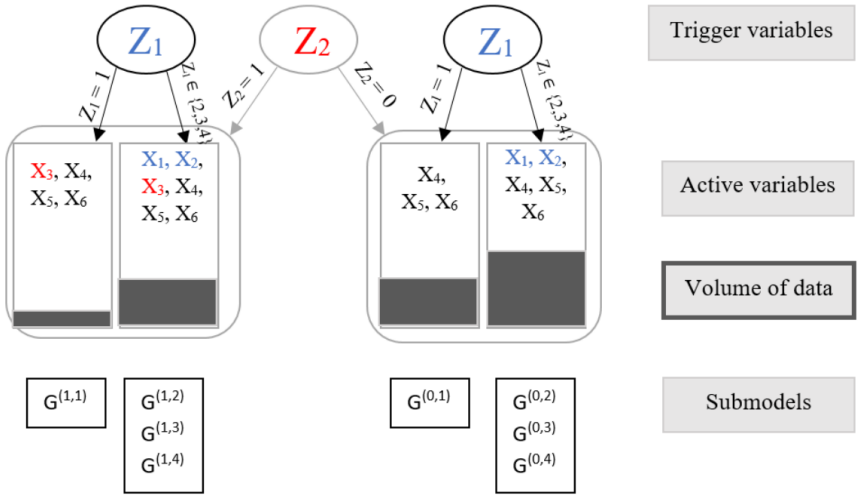


A few numerical results: modified G-Sobol function, 1000000 samples

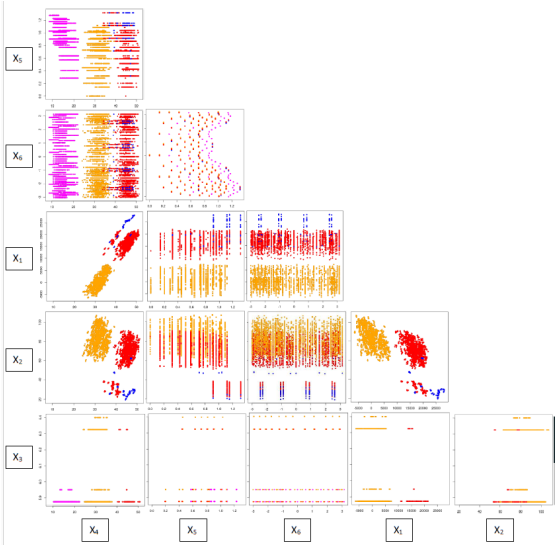


Limitations when dealing with real world data

Measurements on fissile material conditions between exploitation cycles



Limitations when dealing with real world data



Conclusions

Contribution highlights:

- Formalizing the concept of hierarchical variables and their role within the scope of sensitivity analysis
- Specifying the hypotheses necessary for the adaptation of Sobol' indices
- Definition of Sobol' indices in the presence of hierarchical variables, and identification of intuitive expressions
- Proposition of knn estimators allowing to estimate these new Sobol' indices on an i.i.d. sample
- Testing on several numerical test-cases

Limitations & perspectives

Some limitation

- The inputs are considered as **independent**, which is usually not true
- The **independence between hierarchical variables and associated children variables** is a particularly limiting hypothesis
- Closed form 'interpretable' formulas have only been obtained for the case with a single hierarchical variable
- The knn estimation of the Sobol' indices, especially for the total ones, presents a bias and converges slowly in the presence of strong interactions

Limitations & perspectives

Now what? Some perspectives

- Exploiting the connection between Sobol' indices and MDA (based on random forests) in order to avoid knn related estimation issues [BDVS22]
  - The tree-like structure may offer a natural framework for hierarchical variables
- Extending the proposed approach to Shapley indices, allowing to handle dependent inputs
- Shapley or other weight allocation schemes ? (e.g., PMVD)
- Kernel-based sensitivity indices (e.g., HSIC) based on a specifically designed covariance kernel allowing to deal with hierarchical dependencies [Sav+24; Pel+21]
  - Necessary to validate the necessary theoretical properties for these kernels in the presence of hierarchical variables


"That's all Folks!"

Références I

[Sob93] I M Sobol. "On sensitivity estimation for nonlinear mathematical models". In: *Mathematical Modeling and Computational Experiment* 1.4 (1993), pp. 407–414.

[Bro20] B. Broto. "Sensitivity Analysis with dependent random variables : Estimation of the Shapley Effects for unknown input distribution and linear Gaussian models". PhD thesis. Université Paris-Saclay, 2020.

[BDVS22] Clément Bénard, Sébastien Da Veiga, and Erwan Scornet. "Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA". In: *Biometrika* 109.4 (2022), pp. 881–900.

[Sav+24] Paul Saves et al. "SMT 2.0: A Surrogate Modeling Toolbox with a focus on hierarchical and mixed variables Gaussian processes". In: *Advances in Engineering Software* 188 (2024), p. 103571.

[Pel+21] Julien Pelamatti et al. "Bayesian optimization of variable-size design space problems". In: *Optimization and Engineering* 22 (2021), pp. 387–447.

[DV+21] S. Da Veiga et al. *Basics and Trends in Sensitivity Analysis: Theory and Practice in R*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2021. DOI: 10.1137/1.9781611976694.

[DRDT08] E. De Rocquigny, N. Devictor, and S. Tarantola. *Uncertainty in Industrial Practice: A Guide to Quantitative Uncertainty Management*. Wiley, 2008. ISBN: 978-0-470-77074-0. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470770733.