DE LA RECHERCHE À L'INDUSTRIE

SENSITIVITY ANALYSIS BASED ON HSIC DEPENDENCE MEASURES

**Hilbert Schmidt Independence Criterion*

Amandine MARREL

CEA DES/IRESNE/DER, IMT Toulouse

In collaboration with M. Albert, V. Chabridon, M. De Lozzo, R. El Amri, , B. Laurent-Bonneau, A. Meynaoui, J. Pelamatti, H. Raguet.

June 11th, 2021 – **OpenTURNS UserDay #14**

IRESNE | DER | SESI | LEMS

Institut de recherche sur les systèmes nucléaires pour la production d'énergie bas carbone

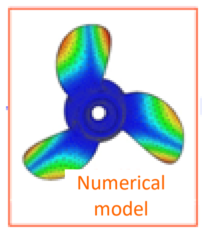
Sensitivity Analysis in Uncertainty framework

- **Numerical simulators:** fundamental tools to model & predict physical phenomena.
- **Large number of input parameters**, characterizing the studied phenomenon or related to its physical and numerical modelling.
- **Uncertainty on some input parameters** → impacts the **uncertainty on the output**
- **Black-box and time-expensive simulators** → limited number of simulations

?

Uncertain input parameters

$\boldsymbol{X} = (X_1, \dots, X_d)$



Numerical simulator

Uncertain output

$\boldsymbol{Y} = \mathcal{M}(X_1, \dots, X_d)$

$\Rightarrow$ Quantify how the variability of the input parameters influences the output

→ Aim of **Sensitivity Analysis (SA)**

Sensitivity Analysis in Uncertainty framework

- **Quantitative SA and Ranking purpose:**
  - Quantify the impact of each uncertain input and interaction → Ranking

  → Reduce the uncertainty of model output

  → Identify the variables to be fixed or further characterized in order to obtain the largest reduction of the output uncertainty

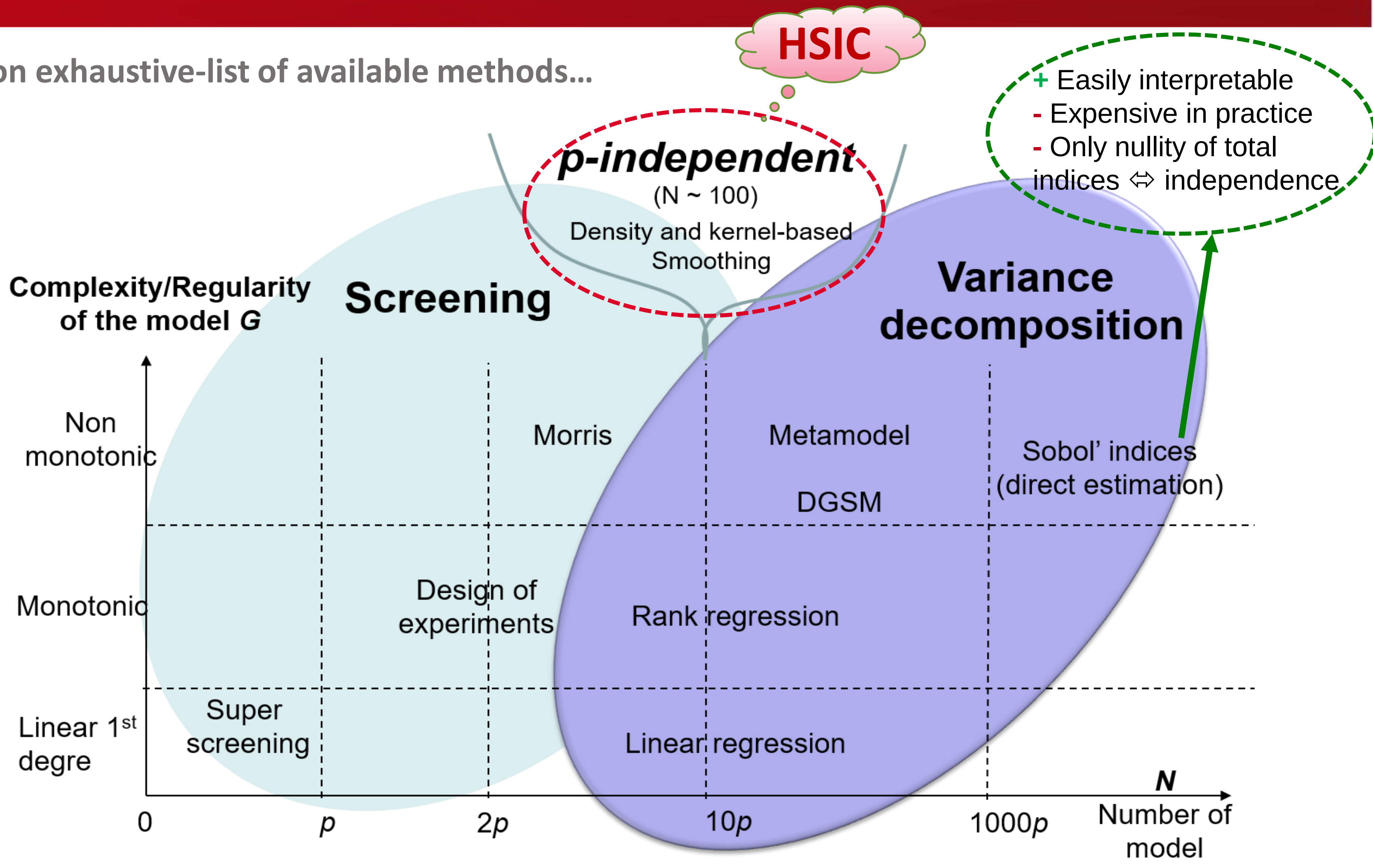
- **Screening purpose: Separate the inputs into two groups influential and non-influential**
  - Non-influential variables fixed without consequences on the output uncertainty
  - Reduction of the model
  - Build a simplified model, a metamodel

Global SA within a probabilistic framework

→ Valuable information to understand $\mathcal{M}$ and underlying phenomenon

Global Sensitivity Analysis

Non exhaustive-list of available methods…



Modified from Iooss et Lemaître [2015]

HSIC Review

IRESNE | DER | SESI | LEMS

Institut de recherche sur les systèmes nucléaires pour la production d'énergie bas carbone

Some notations

$$Y = \mathcal{M}(X_1, ..., X_d)$$

where $X_1, ..., X_d$ are the d input parameters and Y the output

- $X_1, ..., X_d$ are independent and evolve in domain $\mathcal{X}_1, ..., \mathcal{X}_d$
- Y evolves in domain $\mathcal{Y}$
- Function $\mathcal{M}$ is unknown analytically
- Only a sample of n draws of inputs and associated outputs $(X^{(i)}, Y^{(i)})_{1 \leq i \leq n}$, where $Y^{(i)} = \mathcal{M}(X^{(i)})$ for $i = 1, .., n$ is available
- P_X denotes the probability measure of a variable X and p_X its density if X is a continuous variable
- $P_{Y|X}$ conditional probability distribution of Y given X
- $P_{X,Y}$ joint probability measure and $P_X \otimes P_Y$ product of marginal distributions

GSA in a probabilistic framework

- **How to evaluate the sensitivity in a probabilistic way? ⇔ independence**

Solution 1: Quantify the impact of X on the probability distribution of the output Y

→ **By comparing P_Y with $P_{Y|X}$**

Solution 2: Measure and test the dependence between Y and X

→ **By comparing $P_{X,Y}$ with $P_X \otimes P_Y$**

- **In both cases, comparison can be based on:**
  - Cumulative distribution functions
  - Probability density functions
  - Characteristic functions

GSA in a probabilistic framework

- **How to evaluate the sensitivity in a probabilistic way? ⇔ independence**

Solution 1: **Quantify the impact of X on the probability distribution of the output Y**

→ By comparing P_Y with $P_{Y|X}$ with:

$$S_i = \mathbb{E}_{X_i}\left[d\left(P_Y, P_{Y|X_i}\right)\right]$$ *Baucells & Borgonovo [2013]*

where d a **dissimilarity measure** between two probablity distributions based on:

- d comparing the mean: $d\left(P_Y, P_{Y|X_i}\right) = (E[Y] - E[Y|X_i])^2$ → **S_i = 1st Sobol indices**

- **Sobol:**
  - \+ Interesting invariance properties & interpretation in terms of variance decomposition
  - \- Nullity non equivalent with independence ⇨ Only for total Sobol indices
  - \- Only focusing on conditional mean
  - \- Estimation cost for high-order indices ⇨ not directly computable from expensive model

Most of the dependence measures based on comparing P_Y with $P_{Y|X}$ suffer from a high estimation cost and curse of dimensionality

GSA in a probabilistic framework

- **Ok… But HSIC??**

Solution 2: **Measure and test the dependence between *Y* and *X***

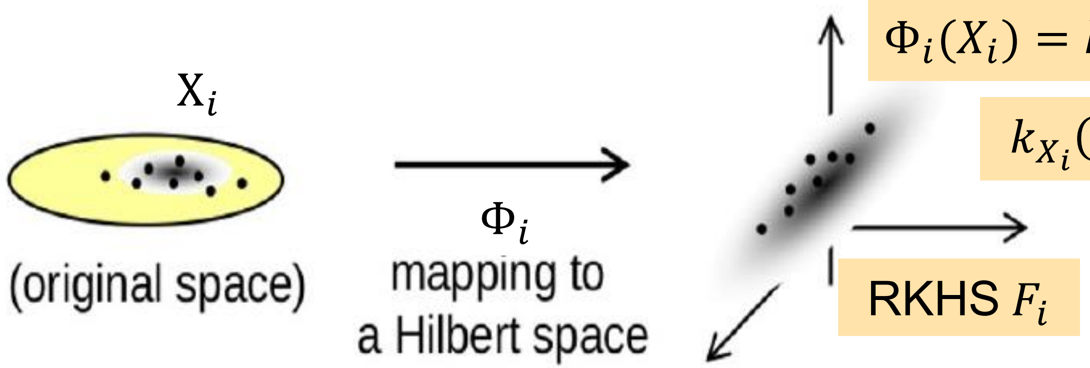
→ **By directly comparing $P_{X,Y}$ with $P_X \otimes P_Y$**

→ **Nonparametric dependence measure** based on a **dissimilarity measure** between **joint** probability distribution of *(X*$_i$*,Y), i.e.* $\mathbb{P}_{X_i,Y}$ and **product of marginals** $\mathbb{P}_{X_i} \otimes \mathbb{P}_Y$ (joint distribution under independence)

For this, association of Reproducing kernel Hilbert spaces F_i and G to X_i and *Y:* *with* Φ_i and Ψ mapping functions to F_i and G (with characteristic kernel k_{X_i} and k_Y).

$\Phi_i(X_i) = k_{X_i}(\bullet, X_i)$ *- canonical feature*

$k_{X_i}(x, x') = \langle \Phi_i(x), \Phi_i(x') \rangle$



Picture modified from Arlot's slides [2014]

HSIC Definition and Estimation

➢ Hilbert-Schmidt independence criterion (HSIC):

(Gretton et al. [2005])

- <u>**One definition:**</u> ***"generalized covariance between two transformations of X_i and Y"***,

Based on cross-covariance operator $C_{X_i,Y}$: covariance between the feature maps, applied respectively to X_i and Y *(tensorised product of covariance between features)*

$$\mathbb{COV}(\Phi_i(X_i),\psi(Y)) = \mathbb{E}_{X_iY}[\Phi_i(X_i)\otimes\psi(Y)] - \mathbb{E}_{X_i}[\Phi_i(X_i)]\otimes\mathbb{E}_Y[\psi(Y)]$$

With Φ_i and Ψ mapping functions to particular functional spaces (F_i and G) associated to X_i and Y (RKHS with characteristic kernel $k_{X_i}$and k_Y).

⇨ **HSIC** is defined as the squared **Hilbert-Schmidt norm of the cross-covariance operator**

$$\mathrm{HSIC}(X_i,Y)_{F_i,G} = \left\|C_{X_i,Y}\right\|_{HS}^2 = \sum_{l,m}\left|\left\langle u_l, C_{X_i,Y}[v_m]\right\rangle_{F_i}\right|^2$$

$$\text{with } \left\langle u_l, C_{X_i,Y}[v_m]\right\rangle_{\mathcal{F}_i} = \mathrm{COV}(u_l(X_i), v_m(Y))$$

and where $(u_l)_{l\geq 0}$ and $(v_m)_{m\geq 0}$ are orthonormal bases of F_i and G.

⇨ A larger panel of input-output dependency can be captured by this operator,

HSIC somehow "summarizes" the set of cross-cov between features applied to X_i and Y

HSIC Definition and Estimation

➢ **Hilbert-Schmidt independence criterion (HSIC):** *(Gretton et al. [2005])*

► ***Kernel trick*** ⇨ *Feature map linked to the positive definite kernel function*

$$k_i(x,x') = \langle \Phi_i(x), \Phi_i(x') \rangle_{\mathscr{F}_i} \text{ and } k(y,y') = \langle \psi(y), \psi(y') \rangle_{\mathscr{G}}$$

$$\Rightarrow \quad \mathrm{HSIC}(X_i,Y)_{\mathscr{F}_i,\mathscr{G}} = \mathbb{E}\left[\kappa_i(X_i,X_i')\kappa(Y,Y')\right] + \mathbb{E}\left[\kappa_i(X_i,X_i')\right]\mathbb{E}\left[\kappa(Y,Y')\right] - 2\mathbb{E}\left[\mathbb{E}[\kappa_i(X_i,X_i')|X_i]\ \mathbb{E}[\kappa(Y,Y')|Y]\right]$$

where $(\boldsymbol{X_i'}, \boldsymbol{Y'})$ is an independent and identically distributed copy of $(\boldsymbol{X_i}, \boldsymbol{Y})$.

Expression with only expectations of kernels ⇒ Monte-Carlo estimator

► ***Characteristic kernels*** ⇨ *Injective feature map* ⇨ *Equivalence to independence:*

$$\boldsymbol{HSIC(X_i,Y) = 0 \iff X_i \perp Y}$$

Gaussian Kernel

$$k(x_i, x_i') = exp\left(-\frac{(x_i - x_i')^2}{2\lambda^2}\right)$$

HSIC Definition and Estimation

➢ Hilbert-Schmidt independence criterion (HSIC):

(Gretton et al. [2005])

- **Case of continuous shift-invariant kernels *k(x,x') = k(x-x')***

⇨ *k* is the Fourier transform of a probability measure

⇨ HSIC somehow consists of comparing characteristic functions (Fourier transform of the probability density function), weighted by this probability measure on frequency space

- **Interpretation of features in particular cases:**
  - $k(x,x') = (< x,x' > \ + \mathbf{1})^{\boldsymbol{p}}$ → involves up to the p^th moments of P_X *(k not characteristic)*
  - $k(x,x') = \boldsymbol{e}^{<x,x'>}$ → moment generating function of P_X *(k characteristic)*
  - $k(x,x') = \boldsymbol{e}^{\boldsymbol{i}x^{\boldsymbol{T}}x'}$ → characteristic function of P_X *(k characteristic)*

HSIC Definition and Estimation

- **Normalization for sensitivity analysis:** *(Da Veiga [2015])*

$$\boxed{R^2_{HSIC,i} = \frac{HSIC\,(X_i,Y)}{\sqrt{HSIC(X_i,X_i)HSIC\,(Y,Y)}}}$$

⇨ $R^2_{HSIC} \in [0,1]$ for easier interpretation

- **Estimation in practice**:

⇒ **Monte Carlo** estimator from a ***n*-sample of simulations** $\left(X_i^{(j)}, Y^{(j)}\right)_{1\le j\le n}$

$$\widehat{\mathrm{HSIC}}(X_i,Y) = \frac{1}{n^2} Tr(K_i HLH)$$

where $H = I_n - \frac{1}{n}$, $K_i = \left(k_i\left(X_i^{(j)}, X_i^{(j')}\right)\right)_{1\le j,j'\le n}$ and $L = \left(k\left(Y^{(j)}, Y^{(j')}\right)\right)_{1\le j,j'\le n}$

- **Statistical properties of $\widehat{\mathrm{HSIC}}$**:
  - Asymptotically unbiased (unbiased estimator also exist, less practical)
  - Variance of order $O(1/n)$
  - Under independence, $n\widehat{\mathrm{HSIC}}(\mathbf{X_i},\mathbf{Y})$ converges asymptotically to a Gamma distribution, whose parameters can be estimated by simple M-C estimators.

HSIC-Based Independence test

- **Use HSIC for screening → with Independence test**

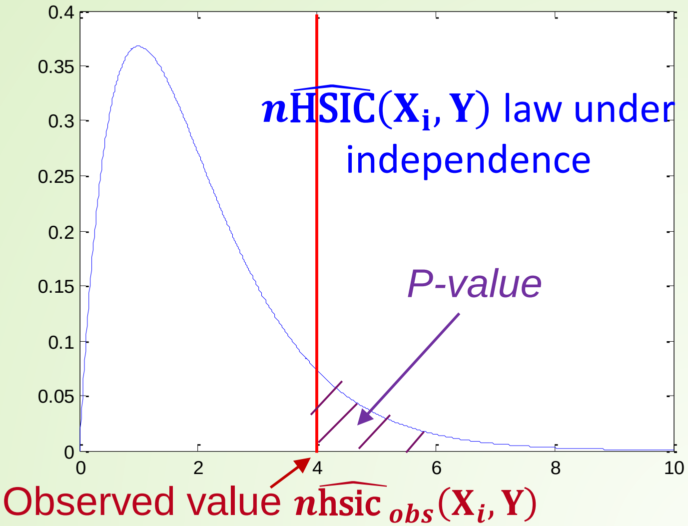
$$HSIC(X_i, Y) = 0 \Leftrightarrow X_i \perp Y$$

- Null hypothesis: $\mathcal{H}_{0,i} : X_i \perp Y_i$ against $\mathcal{H}_{1,i} : X_i \not\perp Y_i$
- Test statistics: $n\widehat{\mathrm{HSIC}}(X_i, Y)$
- Decision rule to obtain a test of level $\alpha = \mathbb{P}_{\mathcal{H}_0}[\text{reject } \mathcal{H}_0]$ (α fixed at 5% or 10%)

$\mathcal{H}_{0,i}$ rejected iff $n\widehat{\mathrm{HSIC}}(X_i, Y) > q_{1-\alpha}$ where $q_{1-\alpha}$ is the $(1-\alpha)$quantile under $\mathcal{H}_{0,i}$

- In practice, computation of p-value:

p-value $= \mathbb{P}[\widehat{\mathrm{HSIC}}(X_i, Y) > \widehat{\mathrm{hsic}}_{obs}(\mathrm{X_i}, \mathrm{Y})]$



Interpretation of *p-value* for a level α ($\alpha = 5\%$ or 10%) for screening:

➢ **pval < α** ⇒ H_0 (Independence) rejected ⇒ X_i **is significantly influential**

HSIC-Based Independence test

- **How to compute $q_{1-\alpha}$? or *p-value*?**

$$p\text{-value} = \mathbb{P}[\,\widehat{\mathrm{HSIC}}(X_i, Y) > \widehat{\mathrm{hsic}}_{obs}(\mathbf{X}_i, \mathbf{Y})]$$

  - **Asymptotic** computation with Gamma approximation of $n\widehat{\mathrm{HSIC}}(\mathbf{X}_i, \mathbf{Y})$ under $X_i \perp Y_i$ for large size sample *(Gretton et al. (2008])*

  - **Permutation-based** approximation for smaller size sample *(De Lozzo & Marrel (2016a], Meynaoui et al. [2019])*

Algorithm 1 – Permutation-based independence test (for each X_i)

Require: The learning sample $(\mathbf{X}_i, \mathbf{Y})$ of n inputs/outputs $\{(X_i^{(1)}, Y^{(1)}), \ldots, (X_i^{(n)}, Y^{(n)})\}$, B and α

Compute $\widehat{\mathrm{HSIC}}_{obs}(X_i, Y)$ from Eq. (2)

Generate B permutation-based samples $(\mathbf{X}_i, \mathbf{Y}_{[b]})_{1 \leq b \leq B}$

Compute the B permutation-based estimators $\left(\widehat{\mathrm{HSIC}}_b(X_i, Y)\right)_{1 \leq b \leq B}$ by replacing $\mathbf{Y}$ by $\mathbf{Y}_{[b]}$ in Eq. (2)

Estimate the p-value by Monte-Carlo estimator $\hat{p}^B_{val,i} = \dfrac{1}{B}\displaystyle\sum_{b=1}^{B} \mathbb{1}_{\widehat{\mathrm{HSIC}}_b(X_i,Y) > \widehat{\mathrm{HSIC}}_{obs}(X_i,Y)}$

if $\hat{p}^B_{val,i} < \alpha$ **then**

    return reject $(\mathscr{H}_0^i)$

else

    return accept $(\mathscr{H}_0^i)$

end if

HSIC-Based Independence test

- **How to compute $q_{1-\alpha}$? or *p-value*?**

$$p\text{-value} = \mathbb{P}[\, \widehat{\mathrm{HSIC}}(X_i, Y) > \widehat{\mathrm{hsic}}_{obs}(\mathrm{X_i}, \mathrm{Y})]$$

  - **Asymptotic** computation with Gamma approximation of $n\widehat{\mathrm{HSIC}}(\mathbf{X_i}, \mathbf{Y})$ under $X_i \perp Y_i$ for large size sample *(Gretton et al. (2008])*

  - **Permutation-based** approximation for smaller size sample *(De Lozzo and Marrel (2016a], Meynaoui et al. [2019])*

→ Theoretical demonstration: permuted-test is of level α

→ Empirically observed: power of asymptotic and permutation test equivalent for a sufficient number of permutations

→ Guidance and comparison of tests according to *n* in *De Lozzo and Marrel (2016a]*

- Estimation of p-value with **permutation**-based tests *(El Amri & Marrel [2021])*

In practice, which number *B* of permutations required?

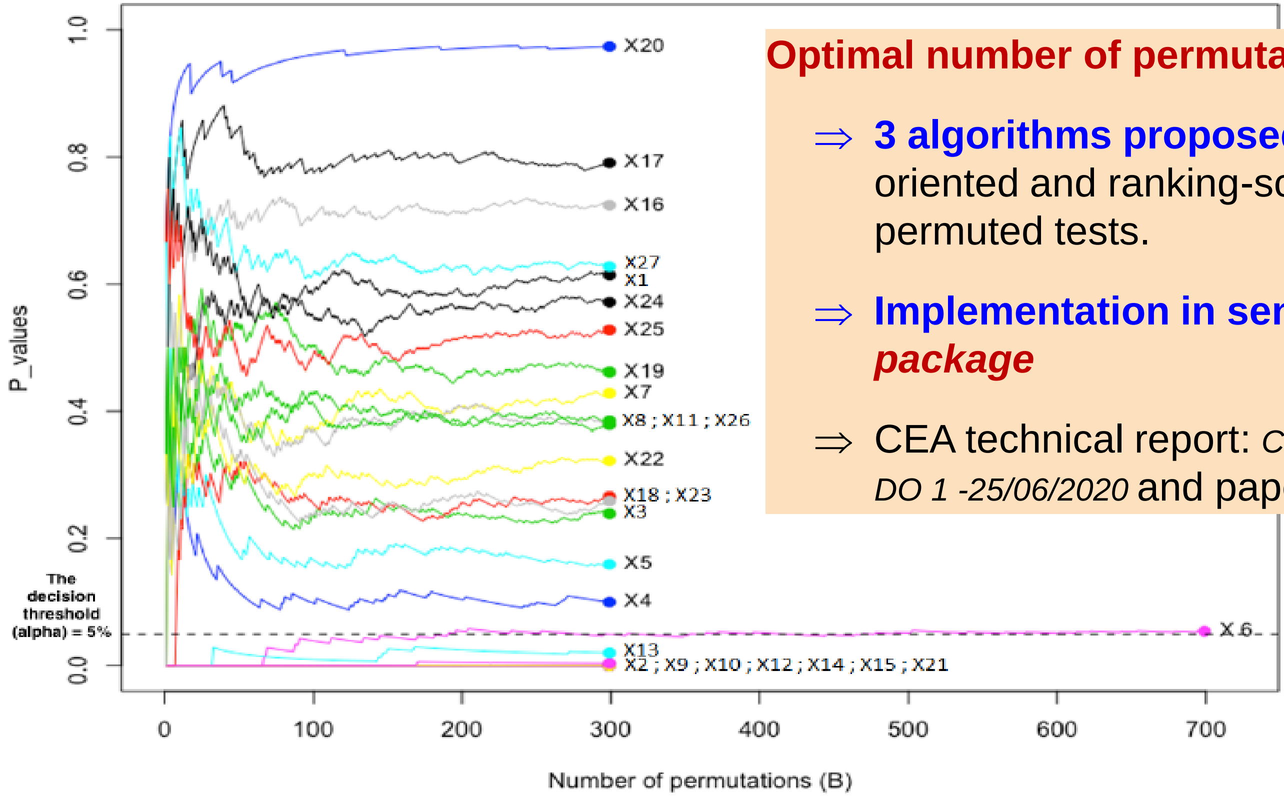


Figure 11: IBLOCA test case – Sequential estimation of p-values by Algorithm 2 (screening), according to the number of permutations.

Optimal number of permutations according to **the final goal**

⇒ **3 algorithms proposed**: screening-oriented, ranking-oriented and ranking-screening-oriented sequential permuted tests.

⇒ **Implementation in sensiHSIC of *R Sensitivity package***

⇒ CEA technical report: *CEA/DES/IRESNE/DER/SESI/LEMS/NT DO 1 -25/06/2020* and paper to appear

In practice reduction of *B* from *B* =5000 to *B* = 300, e.g.
⇒ Convergence studies and sensitivity studies tractable

Practical use of HSIC for industrial applications

IRESNE | *DER* | *SESI* | *LEMS*

Institut de recherche sur les systèmes nucléaires pour la production d’énergie bas carbone

HSIC for Ranking

- **HSIC-based sensitivity analysis:** *(Da Veiga [2015])*

$$\boxed{R^2_{H,X_{[i]}} = \frac{HSIC\,(X_i,Y)}{\sqrt{HSIC(X_i,X_i)HSIC\,(Y,Y)}}}$$

$\Rightarrow R^2_{H,X_{[i]}} \in [0,1]$ for easier interpretation

- **Use for ranking:**

$$\textbf{Influence}(X_{[1]}) > \textbf{Influence}(X_{[2]}) > \cdots > \textbf{Influence}(X_{[d]})$$

where $[\cdot]: i \in \{1,\dots,d\} \mapsto [i] \in \{1,\dots,d\}$ is such that $\widehat{R^2_{H,X_{[1]}}} > \widehat{R^2_{H,X_{[2]}}} > \cdots > \widehat{R^2_{H,X_{[d]}}}$

- **Several illustrations on analytical examples (Linear, Ishighami, G-Sobol, Morris…)**
  - HSIC indices detect non-influential factors easily and robustly, even with small sample size
  - HSIC indices can capture a large spectrum of dependence

    → Good ranking on usual GSA functions from sample size $n\sim 100$

    → Efficiency for screening → Even better: **HSIC independence test**

Use the HSIC independence tests

1/ Screening: **Asymptotic** or **Non-asymptotic** tests, depending on n

H_0 : « X_i and Y are independent. »

$\Rightarrow$ In practice computation of ***p-value:***

P-value: Pr[$\widehat{\mathrm{HSIC}}(X_i, Y)$ > $hsic_{obs}$]

Interpretation of *p-value* for a level α ($\alpha = 5\%$ or 10%) for screening:

- **pval < α** $\Rightarrow$ H_0 (Independence) rejected $\Rightarrow$ X_i **is significantly influential**

2/ Ranking of inputs:

Interpretation of *p-value* for ranking:

Lower pval, stronger H_0 rejected and higher the influence of X_i

- **Inputs are ordered by decreasing influence using p-values:**

$$\text{Influence}(X_{[1]}) > \text{Influence}(X_{[2]}) > \cdots > \text{Influence}(X_{[d]})$$

where $[\cdot]: i \in \{1, \dots, d\} \mapsto [i] \in \{1, \dots, d\}$ is such that $\widehat{\text{pval}}_{[1]} < \widehat{\text{pval}}_{[2]} < \cdots < \widehat{\text{pval}}_{[d]}$.

Playing with kernels…

- **Applications on several industrial test cases with different kind of data**
  - Strategy for oil reservoir characterization test case *(Da Veiga [2015])*
  - Atmospheric dispersion model with spatio-temporal output *(De Lozzo & Marrel [2016b])*
  - Assess the impact of uncertain distribution of input X_i on GSA results, uncertain inputs = probability measures (*Meynaoui et al. [2021]*)

- **Technical point: choose the characteristic kernel according to the type of data**
  - For real and scalar/vector data: Gaussian, Laplacian, Matérn kernel

    → 1 or 2 parameters to be estimated

    Gaussian Kernel

    $$k_G(x_i, x_i') = exp\left(-\frac{(x_i - x_i')^2}{2\sigma^2}\right)$$

  - For binomial or discrete data: Dirac kernel
  - For categorical data: Discrete kernel
  - For functional data: semi-metric based kernels (⚠ not characteristic)

$k(\boldsymbol{x}_i, \boldsymbol{x}_i') = k_r(\Delta(\boldsymbol{x}_i, \boldsymbol{x}_i'))$ with semi-metric Δ (PCA e.g.) and k_r kernel defined on $\mathcal{R}$ (Gaussian..)

(Current CEA work with El Amri)

HSIC for Goal-Oriented SA

- Applications on several **industrial test cases with different kind of data**
  - **Goal-oriented SA for safety studies** *(Marrel & Chabridon [2021], Iooss & Marrel[2019])*
    - ⇒ To measure the input influence in a **restricted output domain**: $Y \in \mathcal{C}$
    - ⇒ Numerous applications for safety and risk assessment ($\mathcal{C}$: critical safety domain, e.g. $\mathcal{C} = \{Y | Y > q_{0.9}\}$)

- **Technical point: choose the characteristic kernel according to the type of data:**
  - Uncertain inputs : real data → Usual Gaussian kernel
  - Output = is Y in a restricted output domain $\mathcal{C}$? ($\mathcal{C}$: critical safety domain)
    - Target SA: measures the influence of X over **the occurrence** of $Y \in \mathcal{C}$

      → Bernouilli output: $\mathbf{1}_{Y \in \mathcal{C}}(Y) \sim \mathcal{B}(p_{\mathcal{C}})$ with $p_{\mathcal{C}} = \mathbb{P}[Y \in \mathcal{C}]$ ⇨ Dirac Kernel
    - Conditional SA: GSA performed within $\mathcal{C}$ only, ignoring what happens outside

      → Real output: $\boldsymbol{Y | Y \in \mathcal{C}}$ with $\mathbb{P}_{|Y\in\mathcal{C}}[\mathcal{A}] = \frac{\mathbb{P}[\mathcal{A} \cap Y \in \mathcal{C}]}{p_{\mathcal{C}}}$ ⇨ Gaussian kernel (if Y real)

HSIC for Goal-Oriented SA

- **Use of HSIC for Target and conditional SA** *(Marrel & Chabridon [2021])*

$\Rightarrow$ **Brute:**

- Target SA: $\boldsymbol{HSIC}(\boldsymbol{X}, \mathbf{1}_{\boldsymbol{Y} \in \boldsymbol{\mathcal{C}}}(\boldsymbol{Y}))$
- Conditional SA: $\boldsymbol{HSIC}(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{Y} \in \boldsymbol{\mathcal{C}})$

$\Rightarrow$ **Smoother versions** to cope with the loss of information and take into account some information outside $\mathcal{C}$ $\rightarrow$ **Use of weight function $\boldsymbol{W_{\mathcal{C}}}$ for relaxation**

$$W_{\mathcal{C}} : \mathcal{Y} \longrightarrow [0,1] \quad ; \quad y \longrightarrow e^{-d_{\mathcal{C}}(y)/s}$$

$\rightarrow \boldsymbol{HSIC}(\boldsymbol{X}, \boldsymbol{W_{\mathcal{C}}}(\boldsymbol{Y}))$ and $\boldsymbol{HSIC}(\boldsymbol{X}, \boldsymbol{W_{\mathcal{C}}}(\boldsymbol{Y})\boldsymbol{Y} | \boldsymbol{Y} \in \boldsymbol{\mathcal{C}})$

Similar use for optimization purpose in Spagnol et al. [2019]

Illustration within the ICSCREAM Methodology

IRESNE / *DER* / *SESI* / *LEMS*

Institut de recherche sur les systèmes nucléaires pour la production d’énergie bas carbone

Key element in ICSCREAM* methodology

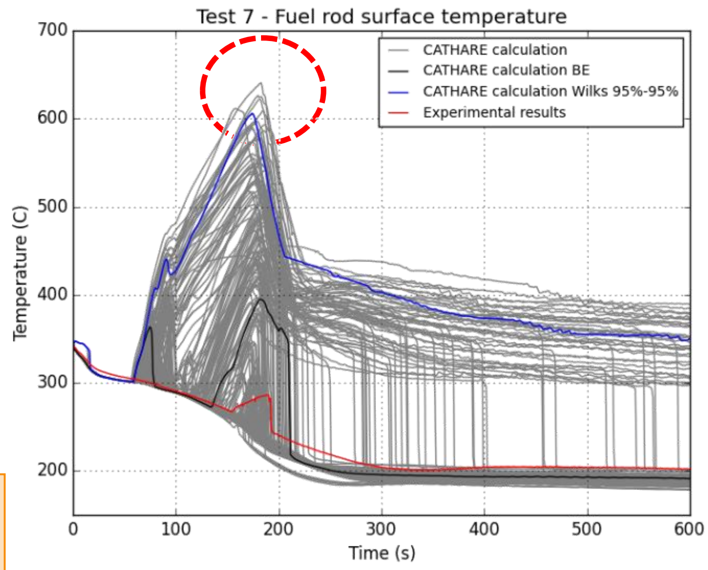
*Identification of penalizing Configurations using SCREening And Metamodel

Accidental scenario on pressurized water reactor: IB-LOCA

LOss of primary **C**oolant **A**ccident due to a **I**ntermediate **B**reak in cold leg

***d* (~ 100) input random variables :**

Critical flowrates, initial/boundary conditions, phys. eq. coef., …



Modelled with **CATHARE2 code**:

- Models complex thermal-hydraulic phenomena
- **Large CPU cost for one code run (> 1 hour)**

Variable of Interest:

2nd peak of cladding temperature (PCT) = scalar output

⇒ ICSCREAM objective

Identify the most **penalizing configurations** of **scenario** inputs for PCT, regardless to the uncertainties of the other inputs.

Key element in ICSCREAM* methodology

*Identification of penalizing Configurations using SCREening And Metamodel

X: Uncertain inputs + scenario inputs to be penalized X_{pen}

↓

Step 1: Monte Carlo sample of *n* simulations (X_S, Y_S)

↓

Step 2: Screening and ranking with HSIC and Target-HSIC independence tests from (X_S, Y_S)

↓

Step 3: Sequential Metamodeling with Gaussian process (GP), from (X_S, Y_S)

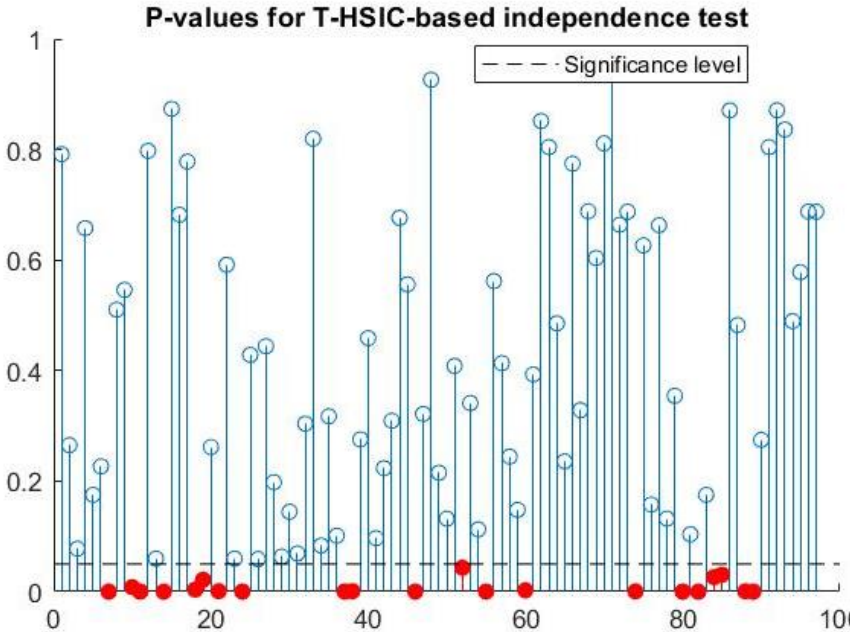
↓

Step 4: Identify with GP metamodel the penalizing values of X_{pen} under the uncertainty of the other inputs $\{X \backslash X_{pen}\}$

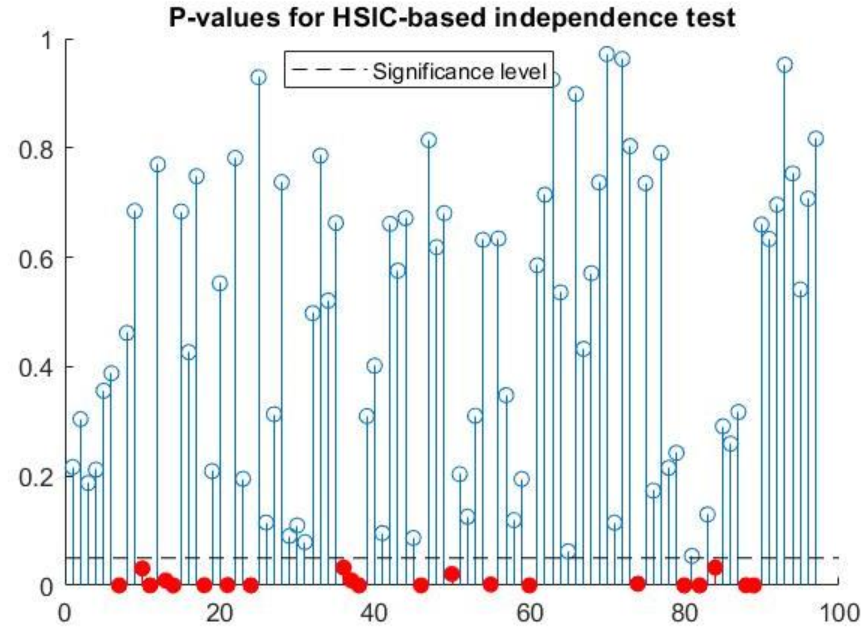
- Identify Primary Influential Inputs (PII) & reduce the dimension before Step 3
- Aggregation of HSIC and Target-HSIC independence tests, to capture both global influence and influence on penalizing configurations
- *Use for Screening (P-value < 5%):* expected reduction of explanatory input variables dim(PII) = $d_{PII} << d$
- *Use for Ranking:* inputs ordered by influence d°, using *P-values*
  ⇨Sequential and more robust metamodel building process

Illustration of Step 2: Screening and ranking with HSIC and Target-HSIC

Global-HSIC tests



T-HSIC ⇨ on exceeding the 90%-quantile $\hat{q}_{0.9}(Y)$



From aggregation, selection of around 20 inputs

Building of a GP metamodel, assessment of its predictive abilities before estimating conditional probabilities

Key element in ICSCREAM* methodology

*Identification of penalizing Configurations using SCREening And Metamodel

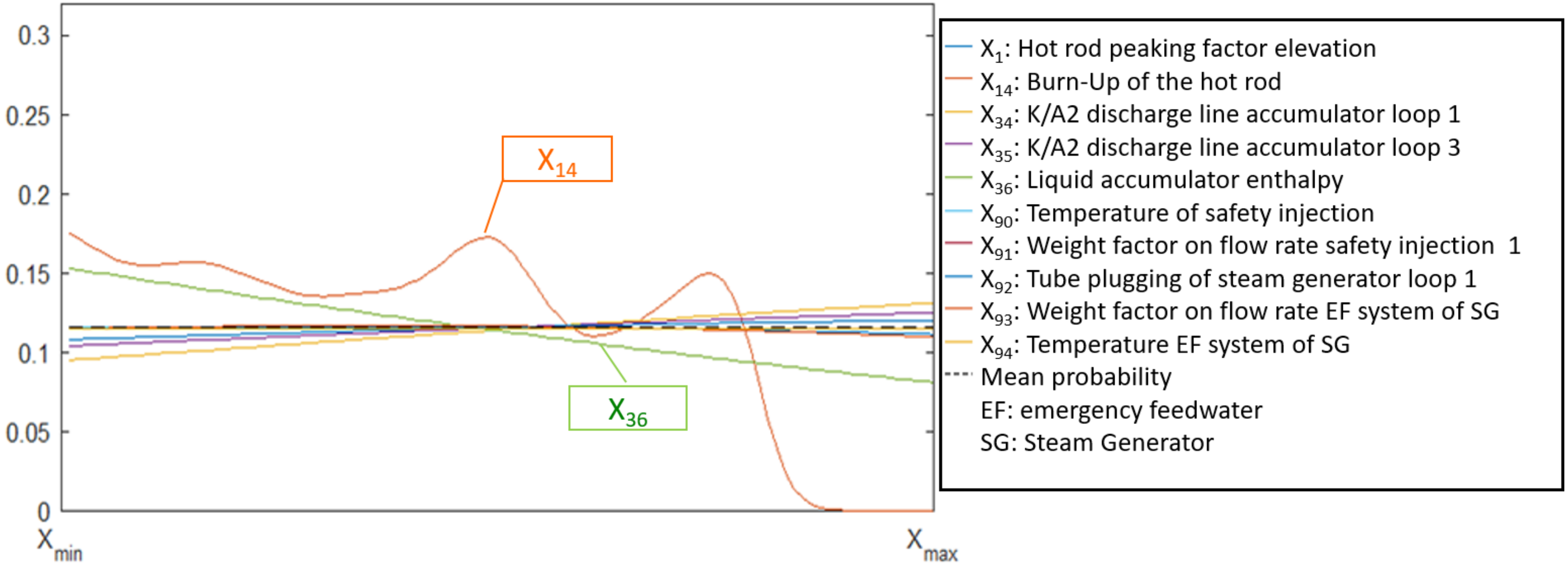
Illustration of Step 4: Capture critical configurations of inputs $\boldsymbol{X_{pen}}$

Leading to the highest probability of PCT exceeding $\hat{q}_{0.9}(Y)$ (under randomness of the other inputs)

$$\hat{P}(\mathbf{X_{pen}}) = P\left[Y_{Gp}(\mathbf{X_{exp}}) > \hat{q}_{0.9} \,|\mathbf{X_{pen}}\right]$$

With $\boldsymbol{X_{exp}} = \{\boldsymbol{X}_{PII} \cup \boldsymbol{X}_{pen}\}$, $\boldsymbol{X}_{PII}$ the inputs selected at Step2

1D- conditional probabilities $\hat{P}(X_{pen})$



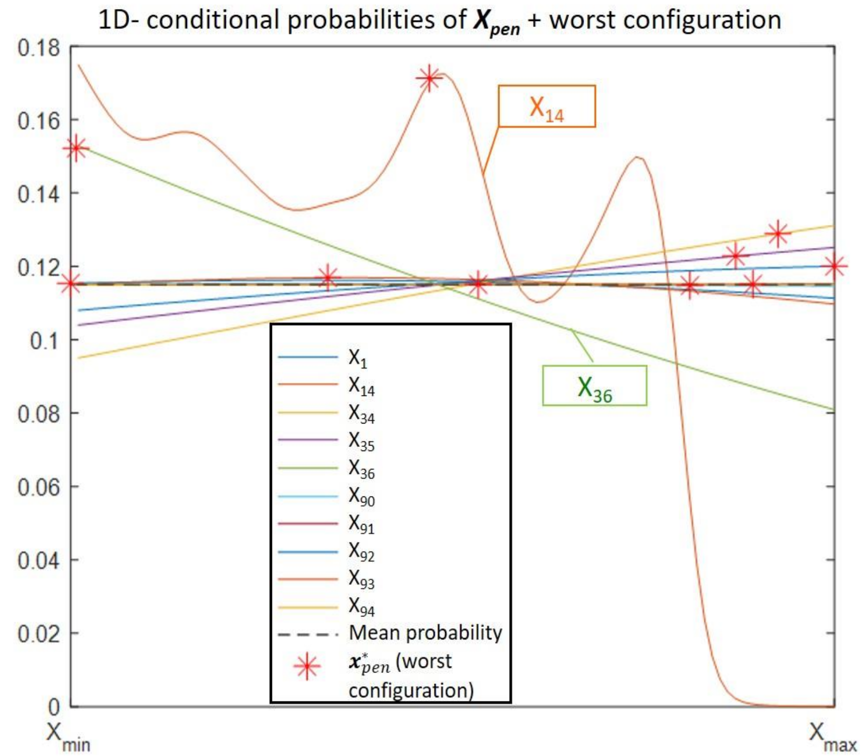
Key element in ICSCREAM* methodology

*Identification of penalizing Configurations using SCREening And Metamodel

Illustration of Step 4: Capture critical configurations of inputs X_{pen}

Identification of the worst configuration

Name	Input to be penalized	Lower bound	Upper bound	Value for the most penalizing configuration ✳
X_1	Hot rod peaking factor elevation [m]	2.4	3.2	**2.4**
X_{14}	Burn-Up of the hot rod [MWj/t]	515	59 000	**28 176.9**
X_{34}	K/A2 discharge line accumulator loop 1 [m^{-4}]	800	1 900	**1 818.8**
X_{35}	K/A2 discharge line accumulator loop 3 [m^{-4}]	800	1 900	**1 757.8**
X_{36}	Liquid accumulator enthalpy [J/kg]	33 544	213 105	**34 827.9**
X_{90}	Temperature of safety injection [°C]	7	50	**41.9**
X_{91}	Weight factor on flow rate safety injection 1	-1	+1	**0.068**
X_{92}	Tube plugging of steam generator loop 1	0	0.09	**0.09**
X_{93}	Weight factor on flow rate Emergency FeedWater System of steam generator	-1	+1	**-0.33**
X_{94}	Temperature Emergency FeedWater System of steam generator [°C]	7	55	**49.8**



Implementation in OpenTURNS

IRESNE | *DER* | *SESI* | *LEMS*

Institut de recherche sur les systèmes nucléaires pour la production d'énergie bas carbone

Implementation in OpenTURNS

- **Proposed features:**
  - Various estimators for different types of sensitivity analysis
    - **Global** (U-stat, V-stat)
    - **Target** (U-stat, V-stat) with various weight functions (smooth and hard, multiple critical regions)
    - **Conditional** (V-stat) with a user-defined weight function (multiple critical regions)
  - Independence statistical tests
    - Permutation-based approach
    - Asymptotic approach
  - Visualization tools
    - Indices and standardized indices (R2-HSIC and others)
    - P-values
- **Will be available in the next OpenTURNS release (v. 1.18)0**

Implementation in OpenTURNS

- **Future developments:**
  - Implementation of dedicated simulation-based algorithms
  - Computation of the sup-HSIC metric (robustness w.r.t. input kernel parametrization)
  - Advanced statistical tests
    - Sequential permutation-based approach (work of [El Amri & Marrel, 2021])
    - Spectral approach (work of [Zhang, Filippi, Gretton & Sejdinovic, 2018])
    - P-values aggregation strategies
  - Visualization tools for high-dimensional problems
    - Automated clustering tools
- **Any other user-related suggestion (or need) can be of interest! Don't hesitate!**
- **Use-cases from the OT community are welcome in order to test and validate the use of HSIC indices for real-world industrial problems!**

Conclusion and prospects

IRESNE | *DER* | *SESI* | *LEMS*

Institut de recherche sur les systèmes nucléaires pour la production d'énergie bas carbone

Conclusions

- **HSIC as GSA indices**
  - Focus the SA analysis on the difference between $P_{X,Y}$ with $P_X \otimes P_Y$
  - Power of RKHS → HSIC=one of the most successful non-parametric dependence measure
  - Capture a large spectrum of relationships
  - Able to deal with many factors and purposes (goal-oriented SA, metamodel, optimization)
  - **Characterize independence** → efficient for screening !

- **HSIC-tests of independence for screening**
  - Rigorous statistical framework, control of 1st and 2nd kind error
  - **P-value** of test → Really efficient for screening and use for quantitative SA

→ **Efficiency demonstrated in numerous industrial applications, especially with small sample size and large dimension**

Prospects

- **Limitations remain in HSIC SA indices**

  - Decomposition into main effects & interactions must be investigated

    ⇒ Assess the use of **HSIC with** **ANOVA-like kernels** and **Shapley-HSIC** **for dependent inputs** (*Da Veiga [2021]*)

  - Multidimensional extension → impact of kernel?

  - Invariance properties → Preliminary isoprobabilistic transformation? (*Poczos et al. (2018)*)

  - Sensitivity to the choice of kernel and of its parameter (bandwidth parameter):

    ⇒ **Aggregated tests** with a collection of bandwidths (Albert et al. [2021])

    ⇒ HSIC-test with optimal bandwidth

  - Extend HSIC-tests to **non i.i.d.** samples → Quasi Monte-Carlo or space-filling design

    ⇒ A first corrected test proposed for scrambled Sobol' sequences (CEA technical report CEA/DES/IRESNE/DER/SESI/LEMS/NT DO 07 of 30/03/2021)

  - Increasing selection rate of HSIC tests with n ⇒ correction to control family-wise error rate and alternative with ANOVA-like kernels

Some References

- Albert, M, Laurent, B., Marrel, A. and Meynaoui, A. Aggregated test of independence based on hsic measures. https://arxiv.org/abs/1902.06441. 2019.
- S. Da Veiga, Global sensitivity analysis with dependence measures, *Journal of Statistical Computation and Simulation*, 85:1283-1305, 2015.
- S. Da Veiga, Kernel-based anova decomposition and shapley effects–application to global sensitivity
- analysis. arXiv preprint arXiv:2101.05487. 2021.
- M. De Lozzo and A. Marrel, New improvements in the use of dependence measures for sensitivity analysis and screening, *Journal of Statistical Computation and Simulation*, 86:3038-3058, 2016
- R. El Amri and A. Marrel. HSIC-based independence tests with optimal sequential permutations: application for sensitivity analysis of numerical simulators, *To appear in Quality and Reliability Engineering International*, 2021.
- B. Iooss and A. Marrel, Advanced methodology for uncertainty propagation in computer experiments with large number of inputs, *Accepted in Journal of Nuclear Technology*, 2019.
- Gretton, Bousquet, Smola and Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. *In: Proceedings Algorithmic Learning Theory*, 2015.
- Iooss B., and Lemaître P. A review on global sensitivity analysis methods. In Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications, Springer, 2015.
- Marrel, A. ICSCREAM methodology for the Identification of penalizing Configurations using SCREening And Metamodel Application to IB-LOCA on PWR (CATHSBI). CEA/DES/IRESNE/DER/SESI/LEMS/NT DR 17 of 15/10/2020.
- A. Marrel and V. Chabridon. Statistical developments for target and conditional sensitivity analysis: application on safety studies for nuclear reactor, Reliability Engineering and System Safety 214, 2021.
- Meynaoui, A. New developments around dependence measures for sensitivity analysis: application to severe accident studies for generation IV reactors. PhD thesis, University of Toulouse, 2019.
- Spagnol A., Le Riche R. and Da Veiga S. Global Sensitivity Analysis for Optimization with Variable Selection, SIAM/ASA Journal on Uncertainty Quantification, Vol. 7, No. 2 : pp. 417-443, 2019.
- Poczos, Z. Ghahramani and J. Schneider. Copula-based Kernel Dependency Measures. https://arxiv.org/abs/1206.4682, 2018.
- Zhang, Q., Filippi, S., Gretton, A. et al. Large-scale kernel methods for independence testing. Stat Comput 28, 113–130 (2018).