



HAL
open science

New statistical methodology for second level global sensitivity analysis

Anouar Meynaoui, Amandine Marrel, Béatrice Laurent

► **To cite this version:**

Anouar Meynaoui, Amandine Marrel, Béatrice Laurent. New statistical methodology for second level global sensitivity analysis. 2019. hal-02019412v1

HAL Id: hal-02019412

<https://hal.science/hal-02019412v1>

Preprint submitted on 18 Feb 2019 (v1), last revised 5 Jan 2023 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

New statistical methodology for second level global sensitivity analysis

Anouar Meynaoui^{1,2}, Amandine Marrel¹, and Béatrice Laurent-Bonneau²

¹CEA, DEN, DER, SESI, LEMS, 13108 Saint-Paul-lez-Durance, France

²INSA Toulouse, France

February 18, 2019

Abstract

Global sensitivity analysis (GSA) of numerical simulators aims at studying the global impact of the input uncertainties on the output. To perform the GSA, statistical tools based on inputs/output dependence measures are commonly used. We focus here on dependence measures based on reproducing kernel Hilbert spaces: the Hilbert-Schmidt Independence Criterion denoted HSIC. Sometimes, the probability distributions modeling the uncertainty of inputs may be themselves uncertain and it is important to quantify the global impact of this uncertainty on GSA results. We call it here the second-level global sensitivity analysis (GSA2). However, GSA2, when performed with a double Monte Carlo loop, requires a large number of model evaluations which is intractable with CPU time expensive simulators. To cope with this limitation, we propose a new statistical methodology based on a single Monte Carlo loop with a limited calculation budget. Firstly, we build a unique sample of inputs from a well chosen probability distribution and the associated code outputs are computed. From this inputs/output sample, we perform GSA for various assumed probability distributions of inputs by using weighted HSIC measures estimators. Statistical properties of these weighted estimators are demonstrated. Finally, we define 2nd-level HSIC-based measures between the probability distributions of inputs and GSA results, which constitute GSA2 indices. The efficiency of our GSA2 methodology is illustrated on an analytical example, thereby comparing several technical options. Finally, an application to a test case simulating a severe accidental scenario on nuclear reactor is provided.

1 Introduction

Numerical simulators (or computer codes) are fundamental tools for understanding, modeling and predicting phenomena. They are widely used nowadays in several fields such as physics, chemistry and biology, but also in economics and social science. These numerical simulators take a large number of input parameters more or less uncertain, characterizing the studied phenomenon. Consequently, the output which is provided by the numerical code is also uncertain. It is therefore important to consider not only the nominal values of inputs, but also the set of all possible values in the range of variation of each uncertain parameter [13, 24]. In the framework of a probabilistic approach, the inputs and the output are considered as random variables and their uncertainties are modeled by probability distributions. The objective is then to evaluate the impact of the input uncertainties on the variability of the output. For this, sensitivity analysis studies can be performed, using statistical methods based on code simulations (also called realizations or observations). To choose these numerical simulations, experimental design techniques can be used (see e.g. [11]).

Generalities on sensitivity analysis. Sensitivity analysis [35] aims at determining how the variability of inputs contributes, qualitatively or quantitatively, to the output variability. Sensitivity analysis can yield a screening of the inputs, which consists in separating the inputs into two subgroups: those that significantly influence the output value (significant inputs) and those whose influence on the output can be neglected. More generally, sensitivity analysis can be divided into two main areas:

- local sensitivity analysis (LSA) which studies the output variability for a small input variation around nominal values (reference values);
- global sensitivity analysis (GSA) which studies the impact of the input uncertainties on the output, considering the whole range of input variation.

We focus here on GSA and we call it in the following, first-level GSA, denoted GSA1.

Use of dependence measures for GSA1. Among GSA1 tools [25], one of the most popular methods used in industrial applications is based on a variance decomposition of the output [38]. The sensitivity indices thus obtained by this decomposition are called Sobol' indices. These indices have the advantage of being easily interpretable but are in practice very expensive in computing time (several tens of thousands of code simulations required). More recently, tools based on dependence measures have been proposed for GSA1 purpose [10]. These measures aim at quantifying, from a probabilistic point of view, the dependence between the output random variable and the input random variables. Among these measures, we can mention the f -divergence of Csiszár which, for a given input, compares the distribution of the output and its distribution when this input is fixed, thanks to a function with specific properties (see [9] for more details). Always on the same principle, the distance correlation is an other dependence measure which compares the characteristic function of a couple of random input/output variables, with the product of the joint characteristic functions of the two variables [42]. Last but not least, the Hilbert-Schmidt independence criterion denoted HSIC [22], generalizes the notion of covariance between two random variables and takes into account a very large spectrum of forms of dependence between variables. Initially developed by statisticians [22] to perform independence tests, these dependence measures offer the advantage of having a low cost of estimation (in practice a few hundred simulations against several tens of thousands for Sobol' indices) and their estimation for all inputs does not depend on the number of inputs. In addition, recent work proposed by [12] showed the efficiency of these measures to perform a screening of the input variables, from various HSIC-based statistical tests of significance. Finally, HSIC measures can easily be extended to non-vector inputs (functional, categorical, etc.). For all these reasons, we will focus here on HSIC measures for GSA1 of numerical simulators.

Second-level input uncertainties and GSA2. In some cases, the probability distributions characterizing the uncertain inputs may themselves be uncertain. This uncertainty may be related to a divergence of expert opinion on the probability distribution assigned to each input or a lack of information to characterize this distribution. The modeling of this lack of knowledge on input laws can take many forms:

- the type of the input distribution is uncertain (uniform, triangular, normal law, ...);
- the distribution is known but its parameters are uncertain (*e.g.*, known normal distribution with unknown mean and variance, eventually estimated on data).

In both cases, the resulting uncertainties on the input laws are referred to here as *second-level uncertainties*. As part of a probabilistic approach, these uncertainties can be modeled by a probability law on a set of possible probability laws of inputs or by a probability law on the parameters of a given input law (*e.g.* Gaussian distribution with probability law on mean and/or variance). In any case, these 2nd-level uncertainties can significantly change the GSA1 results performed by HSIC or any other dependence measure. In this framework, the main purpose of *second-level GSA denoted GSA2* is to answer the following questions: «What impact do 2nd-level uncertainties have on the GSA1 results?» and «What are the most influential ones and those whose influence is negligible?». The GSA2 results and conclusion can then be used to prioritize the characterization efforts on the inputs whose uncertainties on probability laws have the greatest impact on GSA1 results. Note that, we assume here that the inputs are independent and continuous random variables with a probability density function, denoted here pdf.

Practical problems raised by GSA2. In practice, the realization of GSA2 raises several issues and technical locks. First, it is necessary to characterize GSA1 results, *i.e.* to define a representative quantity of interest in order to compare the results obtained for different uncertain input pdf. Then, the impact of each uncertain input pdf on this quantity of interest has to be evaluated. For this, sensitivity indices measuring the dependence between GSA1 results and each input pdf have to be defined. We

propose to call them 2^{nd} -level GSA indices. In order to estimate these measures, an approach based on a "double Monte Carlo loop" could be considered. In the outer loop, a Monte Carlo sample of input pdfs is sorted, while the inner loop aims at evaluating the GSA1 results associated to each pdf. For each pdf selected in the outer loop, the inner loop consists in generating a Monte Carlo sample of code simulations (set of inputs/output) and to compute GSA1 results. The process is repeated for each input pdf. At the end of the outer loop, the impact of input pdf on the GSA1 results can be observed and quantify by computing 2^{nd} -level GSA. Unfortunately, this type of double loop approach requires in practice a very large number of simulations which is intractable for time expensive computer codes. Therefore, other less expensive approaches must be developed.

To answer these different issues (choice of the quantity of interest, definition of 2^{nd} -level sensitivity indices and reduction of the budget of simulations), we propose in this paper a "single loop" Monte Carlo methodology for GSA2 based on both 1^{st} -level and 2^{nd} -level HSIC dependence measures.

The paper is organized as follows. In Section 2, we introduce HSIC measures, before presenting the statistical estimators of these measures, as well as the associated characteristics (bias, variance, asymptotic law). Then, we show that these measures can be formulated and estimated with a sample generated from a different distribution than the *prior* distribution of the inputs. For this, new estimators are proposed and their characteristics are detailed, these new estimators being a key point for the proposed GSA2 methodology. In Section 3, the full methodology for GSA2 is presented: a single inputs/output sample is used, taking advantage of the new HSIC estimators. The GSA2 principle and the related practical issues are first introduced. The general algorithm is then detailed, followed by dedicated sections focusing on major technical elements. In Section 4, the methodology is illustrated on an analytical example, thereby comparing different options and technical choices of the methodology. Finally, an application on a test case simulating a severe accidental scenario on a nuclear reactor is proposed.

2 Statistical inference around Hilbert-Schmidt dependence measures (HSIC)

Throughout the rest of this document, the numerical model is represented by the relation:

$$Y = F(X_1, \dots, X_d),$$

where X_1, \dots, X_d and Y are respectively the d uncertain inputs and the uncertain output, evolving in one-dimensional real areas respectively denoted $\mathcal{X}_1, \dots, \mathcal{X}_d$ and \mathcal{Y} . F denotes the numerical simulator. We note $\mathbf{X} = (X_1, \dots, X_d)$ the vector of inputs. As part of the probabilistic approach, the d inputs are considered as continuous and independent random variables with known densities. These densities are respectively denoted f_1, \dots, f_d . Finally, $f(x_1, \dots, x_d) = f_1(x_1) \times \dots \times f_d(x_d)$ denotes the density of the random vector \mathbf{X} . As the model F is not known analytically, a direct computation of the output probability density as well as dependence measures between \mathbf{X} and Y is impossible. Only observations (or realisations) of F are available. It is therefore assumed in the following that we have a n -sample of inputs and associated outputs $(\mathbf{X}^{(i)}, Y^{(i)})_{1 \leq i \leq n}$, where $Y^{(i)} = F(\mathbf{X}^{(i)})$ for $i = 1, \dots, n$.

2.1 Review on HSIC measures

After introducing their theoretical definition, the estimation of HSIC dependence measures and their use for GSA1 are detailed.

2.1.1 Definition and description

To define the HSIC measure between X_k and Y (where $k \in \{1, \dots, d\}$), [22] associate to X_k a *reproducing kernel Hilbert space* (denoted RKHS, see [3] for more details) \mathcal{F}_k composed of functions mapping from \mathcal{X}_k to \mathbb{R} and characterized by a kernel l_k . The same transformation is carried out for Y , considering a RKHS denoted \mathcal{G} and a kernel l . The scalar products on \mathcal{F}_k and \mathcal{G} are respectively denoted $\langle \cdot, \cdot \rangle_{\mathcal{F}_k}$ and $\langle \cdot, \cdot \rangle_{\mathcal{G}}$. Under this RKHS framework, [4] defines the cross-covariance operator C_k between \mathcal{F}_k and \mathcal{G} as the linear operator from \mathcal{G} to \mathcal{F}_k defined for all $f \in \mathcal{F}_k$ and all $g \in \mathcal{G}$ by:

$$\langle f, C_k g \rangle_{\mathcal{F}_k} = \text{Cov}(f(X_k), g(Y)).$$

The operator C_k generalizes the notion of covariance, taking into account a large spectrum of relationships between X_k and Y (not only linear).

Finally, the Hilbert-Schmidt independence criterion (HSIC) is defined by [22] as the Hilbert-Schmidt norm of the operator C_k :

$$\text{HSIC}(X_k, Y)_{\mathcal{F}_k, \mathcal{G}} = \|C_k\|_{\text{HS}}^2 = \sum_{i,j} \langle u_i, C_k(v_j) \rangle_{\mathcal{F}_k}^2, \quad (1)$$

where $(u_i)_{i \geq 0}$ and $(v_j)_{j \geq 0}$ are respectively orthonormal bases of \mathcal{F}_k and \mathcal{G} .

Remark 2.1. In the following, the notation $\text{HSIC}(X_k, Y)_{\mathcal{F}_k, \mathcal{G}}$ is replaced by $\text{HSIC}(X_k, Y)$ in order to lighten the expressions.

Authors of [22] show that the HSIC measure between an input X_k and the output Y can be expressed using the kernels l_k and l in a more convenient form:

$$\begin{aligned} \text{HSIC}(X_k, Y) &= \mathbb{E}[l_k(X_k, X'_k)l(Y, Y')] + \mathbb{E}[l_k(X_k, X'_k)] \mathbb{E}[l(Y, Y')] \\ &\quad - 2\mathbb{E}[\mathbb{E}[l_k(X_k, X'_k) | X_k] \mathbb{E}[l(Y, Y') | Y]], \end{aligned} \quad (2)$$

where (X'_1, \dots, X'_d) is an independent and identically distributed copy of (X_1, \dots, X_d) and $Y' = F(X'_1, \dots, X'_d)$.

Independence characterization. The nullity of $\text{HSIC}(X_k, Y)$ is not always equivalent to the independence between X_k and Y : this characteristic depends on the RKHS associated to X_k and Y . To ensure equivalence between HSIC nullity and independence, the kernels l_k and l must belong to the specific class of *universal kernels* [32]. A kernel is said to be universal if the associated RKHS is dense in the space of continuous functions w.r.t the infinity norm. However, the universality is a very strong assumption, especially on non-compact spaces. Let us mention as example the Gaussian kernel (the most commonly used for real variables) which is universal only on compact subsets \mathcal{Z} of \mathbb{R}^q [40]. This kernel is defined for a pair of variables $(z, z') \in \mathbb{R}^q \times \mathbb{R}^q$ by:

$$k_\lambda(z, z') = \exp(-\lambda \|z - z'\|_2^2), \quad (3)$$

where λ is a positive real parameter (fixed) and $\|\cdot\|_2$ is the euclidean norm in \mathbb{R}^q .

First referred to as *probability-determining kernels* [18], the notion of *characteristic kernels* [19], which is a weaker assumption than universality, has been lately introduced. It has been proven that this last assumption is sufficient for independence characterization using HSIC. In fact, when the kernels l and l_k are characteristic then, $\text{HSIC}(X_k, Y) = 0$ iff X_k and Y are independent (see e.g. [41]). In particular, the Gaussian kernel defined in Formula (3) is characteristic on the entire \mathbb{R}^q [19].

Remark 2.2. There is no theoretical result for the optimal choice for the kernel width λ in (3). In practice, two main options are adopted for the adjustment of λ : whether the inverse of empirical variance of z , or the inverse of empirical median of $\|z - z'\|_2^2$. In the following, for the computation of $\text{HSIC}(X_k, Y)$, we choose the first option for both kernels associated to X_k and Y . We propose to call the HSIC measures built with these kernels: HSIC measures with standardized Gaussian kernel.

2.1.2 Statistical estimation

In this paragraph, we present HSIC estimators, as well as their characteristics. As a reminder, we assume that we have a n -sample of independent realizations $(\mathbf{X}^{(i)}, Y^{(i)})_{1 \leq i \leq n}$ of the inputs/output couple (\mathbf{X}, Y) where $\mathbf{X} = (X_1, \dots, X_d)$.

Monte Carlo estimation. From Formula (2), authors of [22] propose to estimate each $\text{HSIC}(X_k, Y)$ by:

$$\widehat{\text{HSIC}}(X_k, Y) = \frac{1}{n^2} \sum_{1 \leq i, j \leq n} (L_k)_{i,j} L_{i,j} + \frac{1}{n^4} \sum_{1 \leq i, j, q, r \leq n} (L_k)_{i,j} L_{q,r} - \frac{2}{n^3} \sum_{1 \leq i, j, r \leq n} (L_k)_{i,j} L_{j,r}, \quad (4)$$

where L_k and L are the matrices defined for all $(i, j) \in \{1, \dots, n\}$ by $(L_k)_{i,j} = l_k(X_k^{(i)}, X_k^{(j)})$ and $(L)_{i,j} = l(Y^{(i)}, Y^{(j)})$.

These V-statistic estimators can also be written in the following more compact form (see [22]):

$$\widehat{\text{HSIC}}(X_k, Y) = \frac{1}{n^2} \text{Tr}(L_k H L H), \quad (5)$$

where H is the matrix defined by $H = \left(\delta_{i,j} - \frac{1}{n} \right)_{1 \leq i, j \leq n}$, with $\delta_{i,j}$ the Kronecker symbol between i and j which is equal to 1 if $i = j$ and 0 otherwise.

Remark 2.3. *The estimator of Equation (4) is part of a class of estimators called V-statistics (in the name of Richard Von Mises [33]), which, although being biased (but asymptotically unbiased), are more easily computable than an unbiased version based on U-statistics.*

Characteristics of HSIC estimators. Under the assumption of independence between X_k and Y and the assumption $l_k(x_k, x_k) = l(y, y) = 1$ (as in the case of Gaussian kernels), the estimator $\widehat{\text{HSIC}}(X_k, Y)$ is asymptotically unbiased, its bias converges in $O(\frac{1}{n})$, while its variance converges to 0 in $O(\frac{1}{n^2})$. Moreover, the asymptotic distribution of $n \times \widehat{\text{HSIC}}(X_k, Y)$ is an infinite sum of independent χ^2 random variables, which can be approximated by a Gamma law [36] with shape and scale parameters, respectively denoted γ_k and β_k :

$$\gamma_k \simeq \frac{e_k^2}{v_k}, \quad \beta_k \simeq \frac{n \cdot v_k}{e_k}$$

where e_k and v_k respectively are the expectation and the variance of $\widehat{\text{HSIC}}(X_k, Y)$, i.e. $e_k = \mathbb{E} \left[\widehat{\text{HSIC}}(X_k, Y) \right]$ and $v_k = \text{Var} \left(\widehat{\text{HSIC}}(X_k, Y) \right)$. The reader can refer to [23] and [12] for more details on e_k and v_k and their estimation.

2.1.3 Use for first-level GSA

Several methods based on the use of HSIC measures have been developed for GSA1. In this paragraph, we mention three possible approaches: sensitivity indices [10], asymptotic tests [23] and permutation (also referred to as bootstrap) tests [12].

HSIC-based sensitivity indices. These indices directly derived from HSIC measures, classify the input variables X_1, \dots, X_d by order of influence on the output Y . They are defined for all $k \in \{1, \dots, d\}$ by:

$$R_{\text{HSIC},k}^2 = \frac{\text{HSIC}(X_k, Y)}{\sqrt{\text{HSIC}(X_k, X_k) \text{HSIC}(Y, Y)}}. \quad (6)$$

The normalization in (6) implies that $R_{\text{HSIC},k}^2$ is bounded and included in the range $[0, 1]$ which makes its interpretation easier. In practice, $R_{\text{HSIC},k}^2$ can be estimated using a plug-in approach:

$$\widehat{R}_{\text{HSIC},k}^2 = \frac{\widehat{\text{HSIC}}(X_k, Y)}{\sqrt{\widehat{\text{HSIC}}(X_k, X_k) \widehat{\text{HSIC}}(Y, Y)}}. \quad (7)$$

Asymptotic tests. The independence test between the input X_k and the output Y based on HSIC rejects the independence assumption between these two random variables (hypothesis denoted $\mathcal{H}_{0,k}$) when the p-value¹ of the test based on the statistic $n \times \widehat{\text{HSIC}}(X_k, Y)$ is less than a threshold α (in practice α

¹The p-value of the test is the probability that, under $\mathcal{H}_{0,k}$, the test statistic (in this case, $n \times \widehat{\text{HSIC}}(X_k, Y)$) is greater than or equal to the value observed on the data.

is set at 5% or 10%). Within the asymptotic framework, this p-value denoted P_k is approximated under $\mathcal{H}_{0,k}$ using the Gamma approximation (denoted G_k) of $n \times \widehat{\text{HSIC}}(X_k, Y)$ law:

$$P_k \simeq 1 - F_{G_k} \left(n \times \widehat{\text{HSIC}}(X_k, Y)_{obs} \right), \quad (8)$$

where F_{G_k} is the cumulative distribution function of G_k and $\widehat{\text{HSIC}}(X_k, Y)_{obs}$ is the observed value of the random variable $\widehat{\text{HSIC}}(X_k, Y)$.

Permutation tests. Outside the asymptotic framework, independence tests based on permutation technique can be used. For this, the observed n -sample is resampled B independent times considering random permutations denoted τ on the set $\{1, \dots, n\}$. For each bootstrap sample, the permutation is applied. This permutation is applied only to the vector \mathbf{X} of inputs. We thus obtain B bootstrap-samples $\left(\mathbf{X}^{(\tau^{[b]}(i)}), Y^{(i)} \right)_{1 \leq i \leq n}$ with $1 \leq b \leq B$. The HSIC measures computed on these samples are denoted $\left(\widehat{\text{HSIC}}^{[b]} \right)_{1 \leq b \leq B}$. The p-value (denoted p_k) of the test is then computed by:

$$p_k = \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{\widehat{\text{HSIC}}^{[b]}(X_k, Y) > \widehat{\text{HSIC}}(X_k, Y)}. \quad (9)$$

These different approaches can be used to screen the input variables by order of influence and consequently, to characterize, in the following, the GSA1 results.

2.2 Estimation of HSIC with a sample generated from an alternative distribution

In this part, we first demonstrate that HSIC measures presented in Section 2.1.1, can be expressed and then estimated using a sample generated from a probability distribution of inputs which is not their prior distribution. This sampling distribution will be called "alternative law" or "modified law". The characteristics of these new HSIC estimators (bias, variance, asymptotic law) will be presented. **These estimators will then be used in the proposed methodology for 2nd-level global sensitivity analysis in Section 3.**

2.2.1 Expression and estimation of HSIC measures under an alternative law

The purpose of this paragraph is to express HSIC measures between the inputs X_1, \dots, X_d and the output Y , using d random variables $\tilde{X}_1, \dots, \tilde{X}_d$ whose laws are different from those of X_1, \dots, X_d . We assume that their densities denoted $\tilde{f}_1, \tilde{f}_2, \dots, \tilde{f}_d$ have respectively the same supports as f_1, \dots, f_d . We denote in the following by $\tilde{\mathbf{X}}$ and \tilde{Y} respectively the random vector $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_d)$ and the associated output $\tilde{Y} = F(\tilde{\mathbf{X}})$. Finally, we designate by $\tilde{f}(x_1, \dots, x_d) = \tilde{f}_1(x_1) \times \tilde{f}_2(x_2) \times \dots \times \tilde{f}_d(x_d)$ the density of $\tilde{\mathbf{X}}$.

Changing the probability laws in HSIC expression is based on a technique commonly used in the context of importance sampling (see *e.g.* [5]). This technique consists in expressing an expectation $\mathbb{E}[g(Z)]$, where Z is a random variable with density f_Z , by using a random variable \tilde{Z} with density $f_{\tilde{Z}}$ whose support is the same as that of f_Z . This gives the following expression for $\mathbb{E}[g(Z)]$:

$$\mathbb{E}[g(Z)] = \int_{\text{Supp}(Z)} g(z) f_Z(z) dz = \int_{\text{Supp}(Z)} g(z) \frac{f_Z(z)}{f_{\tilde{Z}}(z)} f_{\tilde{Z}}(z) dz = \mathbb{E}_{\tilde{f}} \left[g(\tilde{Z}) \frac{f_Z(\tilde{Z})}{f_{\tilde{Z}}(\tilde{Z})} \right], \quad (10)$$

where the notation $\mathbb{E}_{\tilde{f}}[h(Z)]$ designates the expectation of $h(Z)$ under the hypothesis $Z \sim \tilde{f}$ and $\text{Supp}(Z)$ denote the support of Z .

The HSIC measures, formulated as a sum of expectations in Equation (2), can then be expressed under the density $f_{\tilde{Z}}$ by adapting Equation (10) to more general forms of expectations. Hence, we obtain:

$$\text{HSIC}(X_k, Y) = H_k^1 + H_k^2 H_k^3 - 2H_k^4, \quad (11)$$

where $(H_k^l)_{1 \leq l \leq 4}$ are the real numbers defined by:

- $H_k^1 = \mathbb{E} \left[l_k(\tilde{X}_k, \tilde{X}'_k) l(\tilde{Y}, \tilde{Y}') w(\tilde{X}) w(\tilde{X}') \right];$
- $H_k^2 = \mathbb{E} \left[l_k \left(\tilde{X}_k, \tilde{X}'_k \right) w(\tilde{X}) w(\tilde{X}') \right];$
- $H_k^3 = \mathbb{E} \left[l \left(\tilde{Y}, \tilde{Y}' \right) w(\tilde{X}) w(\tilde{X}') \right];$
- $H_k^4 = \mathbb{E} \left[\mathbb{E} \left[l_k \left(\tilde{X}_k, \tilde{X}'_k \right) w(\tilde{X}') \mid \tilde{X}_k \right] \mathbb{E} \left[l \left(\tilde{Y}, \tilde{Y}' \right) w(\tilde{X}') \mid \tilde{Y} \right] w(\tilde{X}) \right]$

where $\tilde{\mathbf{X}}'$ is an independent and identically distributed copy of $\tilde{\mathbf{X}}$, $\tilde{Y}' = F(\tilde{\mathbf{X}}')$ and $w = \frac{f}{\tilde{f}}$.

Formula (11) shows that $\text{HSIC}(X_k, Y)$ with $k = 1, \dots, d$ can then be estimated using a sample generated from \tilde{f} , provided that \tilde{f} has the same support than the original density f . Thus, if we consider a n -sample of independent realizations $\left(\tilde{\mathbf{X}}^{(i)}, \tilde{Y}^{(i)} \right)_{1 \leq i \leq n}$, where $\tilde{\mathbf{X}}$ is generated from \tilde{f} and $\tilde{Y}^{(i)} = F(\tilde{\mathbf{X}}^{(i)})$ for $i = 1, \dots, n$, we propose the following V-statistic estimator of $\text{HSIC}(X_k, Y)$:

$$\widetilde{\text{HSIC}}(X_k, Y) = \tilde{H}_k^1 + \tilde{H}_k^2 \tilde{H}_k^3 - 2\tilde{H}_k^4, \quad (12)$$

where $(\tilde{H}_k^l)_{1 \leq l \leq 4}$ are the V-statistics estimators of $(H_k^l)_{1 \leq l \leq 4}$.

Proposition 2.1. *Similarly to Equation (5), this estimator can be rewritten as:*

$$\widetilde{\text{HSIC}}(X_k, Y) = \frac{1}{n^2} \text{Tr} \left(W \tilde{L}_k W H_1 \tilde{L} H_2 \right), \quad (13)$$

where W , \tilde{L}_k , \tilde{L} , H_1 and H_2 are the matrices defined by:

- $\tilde{L}_k = \left(l_k \left(\tilde{X}_k^{(i)}, \tilde{X}_k^{(j)} \right) \right)_{1 \leq i, j \leq n};$
- $\tilde{L} = \left(l \left(\tilde{Y}^{(i)}, \tilde{Y}^{(j)} \right) \right)_{1 \leq i, j \leq n};$
- $W = \text{Diag} \left(w(\tilde{X}^{(i)}) \right)_{1 \leq i \leq n};$
- $H_1 = I_n - \frac{1}{n} U W;$
- $H_2 = I_n - \frac{1}{n} W U;$
- I_n the identity matrix of size n ;
- U the matrix filled with 1.

The proof of this proposition is detailed in Appendix A.

Remark 2.4. *Similarly to Equation (7), the sensitivity index $R_{\text{HSIC},k}^2$ can also be estimated using the sample $\left(\tilde{\mathbf{X}}^{(i)}, \tilde{Y}^{(i)} \right)_{1 \leq i \leq n}$ by:*

$$\tilde{R}_{\text{HSIC},k}^2 = \frac{\widetilde{\text{HSIC}}(X_k, Y)}{\sqrt{\widetilde{\text{HSIC}}(X_k, X_k) \widetilde{\text{HSIC}}(Y, Y)}}. \quad (14)$$

2.2.2 Statistical properties of HSIC modified estimators

In this section we show that the estimator $\widetilde{\text{HSIC}}(X_k, Y)$ has asymptotic properties similar to those of the estimator $\widehat{\text{HSIC}}(X_k, Y)$: same asymptotic behaviors of expectation and variance and same type of asymptotic distribution. The properties presented in the following are proved in Appendix B, C and D.

Proposition 2.2 (Bias). *The estimator $\widetilde{\text{HSIC}}(X_k, Y)$ is asymptotically unbiased and its bias converges in $O(\frac{1}{n})$. Moreover, under the hypothesis of independence between X_k and Y and the assumption $l_k(x_k, x_k) = l(y, y) = 1$, its bias is:*

$$\begin{aligned} \mathbb{E} \left[\widetilde{\text{HSIC}}(X_k, Y) \right] - \text{HSIC}(X_k, Y) &= \frac{2}{n} (E_\omega^k - E_{x_k, \omega}) (E_\omega^{-k} - E_{y, \omega}) - \frac{1}{n} (E_\omega - E_{x_k}) (E_\omega - E_y) \\ &\quad + \frac{1}{n} E_\omega (E_\omega - 1) + O\left(\frac{1}{n^2}\right), \end{aligned} \quad (15)$$

where

$$\begin{aligned} E_\omega &= \mathbb{E} \left[\omega^2(\tilde{X}) \right], & E_{x_k} &= \mathbb{E} \left[l_k(\tilde{X}_k, \tilde{X}'_k) \omega_k(\tilde{X}_k) \omega_k(\tilde{X}'_k) \right], \\ E_y &= \mathbb{E} \left[l(\tilde{Y}, \tilde{Y}') \omega_{-k}(\tilde{X}_{-k}) \omega_{-k}(\tilde{X}'_{-k}) \right], & E_{x_k, \omega} &= \mathbb{E} \left[l_k(\tilde{X}_k, \tilde{X}'_k) \omega_k^2(\tilde{X}_k) \omega_k(\tilde{X}'_k) \right], \\ E_{y, \omega} &= \mathbb{E} \left[l(\tilde{Y}, \tilde{Y}') \omega_{-k}^2(\tilde{X}_{-k}) \omega_{-k}(\tilde{X}'_{-k}) \right], & E_\omega^k &= \mathbb{E} \left[\omega_k^2(\tilde{X}_k) \right], \\ E_\omega^{-k} &= \mathbb{E} \left[\omega_{-k}^2(\tilde{X}_{-k}) \right], \end{aligned}$$

ω and ω_k respectively denote the functions $\frac{f}{f}$, $\frac{f_k}{f_k}$, \tilde{X}_{-k} is the random vector extracted from $\tilde{\mathbf{X}}$ by removing the k -th coordinate, \tilde{X}'_{-k} an independent and identically distributed copy of \tilde{X}_{-k} and $\omega_{-k}(x_{-k}) = \prod_{\substack{r=1 \\ r \neq k}}^d \omega_r(x_r)$ with x_{-k} the vector extracted from the vector (x_1, \dots, x_d) by removing the k -th coordinate.

Under the independence assumption, an asymptotically unbiased estimator of the bias of $\widetilde{\text{HSIC}}(X_k, Y)$ can be obtained by replacing each expectation in (15) by its empirical estimator.

Proposition 2.3 (Variance). *Under the independence hypothesis between X_k and Y , the variance of $\widetilde{\text{HSIC}}(X_k, Y)$ (denoted here ϑ_k) converges to 0 in $O(\frac{1}{n^2})$. More precisely, the variance ϑ_k can be expressed as:*

$$\vartheta_k = \frac{72(n-4)(n-5)}{n(n-1)(n-2)(n-3)} \mathbb{E}_{1,2} \left[\mathbb{E}_{3,4} [\tilde{h}_{1,2,3,4}]^2 \right] + O\left(\frac{1}{n^3}\right), \quad (16)$$

where $\tilde{h}_{1,2,3,4} = \frac{1}{4!} \sum_{(t,u,v,s)}^{(1,2,3,4)} \left[(\tilde{l}_k)_{t,u} \tilde{l}_{t,u} + (\tilde{l}_k)_{t,u} \tilde{l}_{v,s} - 2 (\tilde{l}_k)_{t,u} \tilde{l}_{t,v} \right]$, where the notation corresponds to the sum over all permutations (t, u, v, s) of $(1, 2, 3, 4)$. The notations $(\tilde{l}_k)_{p,q}$, $\tilde{l}_{p,q}$ and w_p respectively denote $l_k(\tilde{X}_k^{(p)}, \tilde{X}_k^{(q)})$, $l(\tilde{Y}^{(p)}, \tilde{Y}^{(q)})$ and $\frac{f}{f}(\tilde{\mathbf{X}}^{(p)})$. Finally, the notation $\mathbb{E}_{p,q}$ means that the expectation is done by integrating only with respect to the variables indexed by p and q .

An estimator $\tilde{\vartheta}_k$ of ϑ_k can be deduced from Equation (16):

$$\tilde{\vartheta}_k = \frac{2(n-4)(n-5)}{n^3(n-1)(n-2)(n-3)} \mathbf{1}^T (\tilde{B} \odot \tilde{B}) \mathbf{1}, \quad (17)$$

with $\tilde{B} = \sum_{r=1}^4 \tilde{B}_r$, where the matrices $(\tilde{B}_r)_{1 \leq r \leq 4}$ are defined for all $(i, j) \in \{1, \dots, n\}^2$ by:

$$\begin{aligned} (\tilde{B}_1)_{i,j} &= (\tilde{l}_k)_{i,j} w_i w_j (\tilde{l}_{i,j} - \tilde{l}_{i,\cdot} - \tilde{l}_{\cdot,j} + \tilde{l}), & (\tilde{B}_2)_{i,j} &= -(\tilde{l}_k)_{i,\cdot} w_i (\tilde{l}_{i,j} w_j - \tilde{l}_{i,\cdot} - \tilde{l}_{\cdot,j} w_j + \tilde{l}), \\ (\tilde{B}_3)_{i,j} &= -(\tilde{l}_k)_{\cdot,j} w_j (\tilde{l}_{i,j} w_i - \tilde{l}_{i,\cdot} w_i - \tilde{l}_{\cdot,j} + \tilde{l}), & (\tilde{B}_4)_{i,j} &= (\tilde{l}_k) (\tilde{l}_{i,j} w_i w_j - \tilde{l}_{i,\cdot} w_i - \tilde{l}_{\cdot,j} w_j + \tilde{l}), \end{aligned}$$

where $w_i = W_{i,i}$ and the terms $(\tilde{l}_k)_{i,\cdot}$, $(\tilde{l}_k)_{\cdot,j}$, (\tilde{l}_k) , $\tilde{l}_{i,\cdot}$, $\tilde{l}_{\cdot,j}$ and \tilde{l} are the empirical means (denoted with a bar above):

$$\begin{aligned} (\tilde{l}_k)_{i,\cdot} &= L_k \left(\overline{\tilde{X}_k^{(i)}, \tilde{X}_k} \right) \omega_k(\tilde{X}_k), & \tilde{l}_{i,\cdot} &= L \left(\overline{\tilde{Y}^{(i)}, \tilde{Y}} \right) \omega_{-k}(\tilde{X}_{-k}), \\ (\tilde{l}_k) &= L_k \left(\overline{\tilde{X}'_k, \tilde{X}_k} \right) \omega_k(\tilde{X}_k), & \tilde{l} &= L \left(\overline{\tilde{Y}', \tilde{Y}} \right) \omega_{-k}(\tilde{X}_{-k}). \end{aligned}$$

Theorem 2.1 (Asymptotic law). *In a similar way as $n \times \widehat{\text{HSIC}}(X_k, Y)$, one can prove that the asymptotic distribution of $n \times \widetilde{\text{HSIC}}(X_k, Y)$ can be approximated by a Gamma distribution, whose parameters $\tilde{\gamma}_k$ and $\tilde{\beta}_k$ are given by $\tilde{\gamma}_k = \frac{\varepsilon_k^2}{\vartheta_k}$ and $\tilde{\beta}_k = \frac{n\vartheta_k}{\varepsilon_k}$, where ε_k and ϑ_k are the expectation and variance of $\widetilde{\text{HSIC}}(X_k, Y)$, i.e. $\varepsilon_k = \mathbb{E} \left[\widetilde{\text{HSIC}}(X_k, Y) \right]$ and $\vartheta_k = \text{Var} \left(\widetilde{\text{HSIC}}(X_k, Y) \right)$.*

In practice, these parameters are respectively estimated by the empirical estimator for ε_k and the estimator given by Equation (16) for ϑ_k .

Remark 2.5. *From a practical point of view, the greater $\text{Var} \left(\omega_k(\tilde{X}_k) \right)$, the greater the number of simulations required to accurately estimate $\text{HSIC}(X_k, Y)$. It is therefore highly recommended to check that $\text{Var} \left(\omega_k(\tilde{X}_k) \right)$ is finite. For instance, in the case of densities with compact supports, it is enough to check that ω_k is finite on its support.*

2.2.3 Illustration on an analytical example

In this paragraph, we illustrate via a numerical application the behavior and the convergence of the modified estimators $\widetilde{\text{HSIC}}$, according to the size of the inputs/outputs sample. For that, we consider the analytic model h inspired from Ishigami's model [26] and defined by:

$$h(X_1, X_2, X_3) = \sin(X_1) + 1.8 \sin^2(X_2) + 0.5 X_3^4 \sin(X_1), \quad (18)$$

where the inputs X_1 , X_2 et X_3 are assumed to be independent and follow a triangular distribution on $[0, 1]$ with a mode equal to $\frac{1}{2}$. We denote by Y the output variable $Y = h(X)$.

We consider standardized Gaussian kernel HSIC measures (see remark 2.2) between each input X_k , $k = 1 \dots 3$ and the output Y . The objective is to estimate these measures from samples where the inputs are independent and identically distributed but generated from a uniform distribution on $[0, 1]$ (modified law). For this, we consider Monte Carlo samples of size $n = 100$ to $n = 1500$ and for each sample size, the estimation process is repeated 200 times, with independent random samples.

Figure (1) presents as a boxplot the convergence graphs of the estimators $\widetilde{\text{HSIC}}(X_k, Y)$. Results for the estimator $\widehat{\text{HSIC}}(X_k, Y)$ computed with samples generated from the original law (namely triangular) are also given. Theoretical values of HSIC are represented in red dotted lines. We observe that for small sample sizes ($n < 500$), modified estimators $\widetilde{\text{HSIC}}(X_k, Y)$ have more bias and variance than estimators $\widehat{\text{HSIC}}(X_k, Y)$. But, from size $n = 700$, both estimators have similar behaviors.

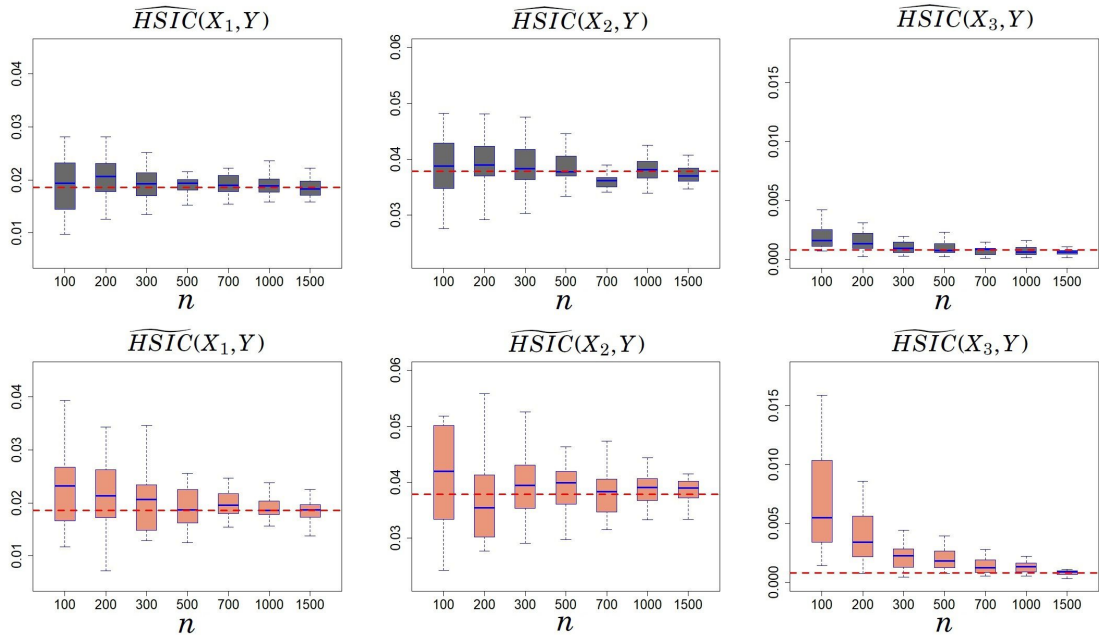


Figure 1: Convergence plots of the estimators $\widehat{\text{HSIC}}(X_k, Y)$ and $\widetilde{\text{HSIC}}(X_k, Y)$ for Ishigami function, according to the sample size n . Theoretical values are represented in red dashed lines.

On this same example, we are now interested in the classification of input variables by order of influence. For this, the sensitivity index $R_{\text{HSIC},k}^2$ ($1 \leq k \leq d$) is computed from $\widehat{\text{HSIC}}(X_k, Y)$ with Equation (14) and the inputs are ordered by decreasing $\widetilde{R}_{\text{HSIC},k}^2$. For each sample size, the rates of good ranking of inputs given by $\widetilde{R}_{\text{HSIC}}^2$ estimators are also computed. **Results are given in Table (1): they illustrate that, even for small sample sizes (e.g. $n = 200$), modified estimators $\widetilde{R}_{\text{HSIC}}^2$ have good ranking ability.**

| $n = 100$ | $n = 200$ | $n = 300$ | $n \geq 500$ |
|-----------|-----------|-----------|--------------|
| 88% | 93.5% | 97% | 100% |

Table 1: Good ranking rates of input variables using modified estimators $\widetilde{R}_{\text{HSIC}}^2$ for the Ishigami function, for different sample sizes n .

3 New methodology for second-level GSA

In this section, we first define the 2nd-level global sensitivity analysis (GSA2) before listing the issues raised by its implementation. We then present and detail the methodology that we propose, in order to tackle these issues.

3.1 Principle and objective

In this part, we assume that the inputs X_1, \dots, X_d vary according to unknown probability distributions $\mathbb{P}_{X_1}, \dots, \mathbb{P}_{X_d}$. This lack of knowledge can take many forms; we mention as examples, the case of parameterized distributions where the parameters are unknown, or the case where the distribution characteristics (e.g. mean, variance, etc.) are known but not the nature of the law (uniform, triangular, normal, etc.). As part of a probabilistic approach, this lack of knowledge is modeled by probability laws (on a set of distributions) denoted $\mathbb{P}_{\mathbb{P}_{X_1}}, \dots, \mathbb{P}_{\mathbb{P}_{X_d}}$.

Each assumed joint distribution $\mathbb{P}_X = \mathbb{P}_{X_1} \times \dots \times \mathbb{P}_{X_d}$ of inputs yields potentially different results of 1st-level global sensitivity analysis (GSA1). It is therefore important to take into account the impact of these uncertainties on GSA1 results. Thus, we will designate by GSA2 the statistical methods whose purpose is to quantify for each input parameter X_k , the impact of the uncertainty of \mathbb{P}_{X_k} on GSA1 results, this uncertainty being modeled by $\mathbb{P}_{\mathbb{P}_{X_k}}$. From GSA2 results, the probability distributions of inputs can be separated into two groups: those which significantly modify GSA1 results and those whose influence is negligible. Subsequently, probability distributions with a small impact can be set to a reference law and the efforts of characterization will be focused on the most influential distributions to improve their knowledge (strategy of uncertainty reduction).

3.2 Issues raised by GSA2

We present in what follows the different issues and technical locks raised by the realization of a GSA2.

3.2.1 Characterization of GSA1 results

The realization of GSA2 requires a prior characterization of GSA1 results. This characterization consists in associating to a given input distribution $\mathbb{P}_X = \mathbb{P}_{X_1} \times \dots \times \mathbb{P}_{X_d}$, a measurable quantity \mathcal{R} which represents GSA1 results. To choose this quantity of interest, we propose the following options, all based on HSIC (see 2.1.3):

- **Vector $R_{\text{HSIC}}^2 = (R_{\text{HSIC},1}^2, \dots, R_{\text{HSIC},d}^2)$ of sensitivity indices.** The quantity of interest $\mathcal{R} = R_{\text{HSIC}}^2$ is thereby a vector of d real components.
- **Ranking of inputs X_1, \dots, X_d using the indices $R_{\text{HSIC},1}^2, \dots, R_{\text{HSIC},d}^2$.** In this case, the quantity of interest \mathcal{R} is a permutation on the set $\{1, \dots, d\}$, which verifies that $\mathcal{R}(k) = j$ if and only if the variable X_j is the k -th in the ranking.
- **Vector $P = (P_1, \dots, P_d)$ of p-values associated with asymptotic independence tests.** In this case, the quantity of interest $\mathcal{R} = P$ is a vector of d components in $[0, 1]^d$.
- **Vector $p = (p_1, \dots, p_d)$ of p-values associated with permutation independence tests.** The quantity of interest $\mathcal{R} = p$ is a vector of d components in $[0, 1]^d$.

3.2.2 Definition of 2nd-level sensitivity indices

By analogy with formulas (2), it is possible to build 2nd-level HSIC measures between the probability distributions $\mathbb{P}_{X_1}, \dots, \mathbb{P}_{X_d}$ and the quantity of interest \mathcal{R} . This involves to define RKHS kernels $l_{\mathcal{D}_1}, \dots, l_{\mathcal{D}_d}$ on input distributions and a RKHS kernel $l_{\mathcal{R}}$ on the quantity of interest \mathcal{R} . This point will be further detailed in Section 3.4. Thus, assuming all the kernels are defined, we propose the 2nd-level HSIC measures defined for $k = 1..d$ by:

$$\begin{aligned} \text{HSIC}(\mathbb{P}_{X_k}, \mathcal{R}) &= \mathbb{E} [l_{\mathcal{D}_k}(\mathbb{P}_{X_k}, \mathbb{P}'_{X_k}) l_{\mathcal{R}}(\mathcal{R}, \mathcal{R}')] + \mathbb{E} [l_{\mathcal{D}_k}(\mathbb{P}_{X_k}, \mathbb{P}'_{X_k})] \mathbb{E} [l_{\mathcal{R}}(\mathcal{R}, \mathcal{R}')] \\ &\quad - 2\mathbb{E} [\mathbb{E} [l_{\mathcal{D}_k}(\mathbb{P}_{X_k}, \mathbb{P}'_{X_k}) \mid \mathbb{P}_{X_k}] \mathbb{E} [l_{\mathcal{R}}(\mathcal{R}, \mathcal{R}') \mid \mathcal{R}]], \end{aligned} \quad (19)$$

where \mathbb{P}'_{X_k} is an independent and identically distributed copy of \mathbb{P}_{X_k} and \mathcal{R}' the GSA1 results associated to \mathbb{P}'_{X_k} .

From 2nd-level HSIC measures, we can define GSA2 indices by:

$$R_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R}) = \frac{\text{HSIC}(\mathbb{P}_{X_k}, \mathcal{R})}{\sqrt{\text{HSIC}(\mathbb{P}_{X_k}, \mathbb{P}_{X_k}) \text{HSIC}(\mathcal{R}, \mathcal{R})}}, \text{ for } k = 1..d. \quad (20)$$

3.2.3 Monte Carlo estimation

To estimate $R_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R})$, for $k = 1..d$, one has to dispose of a n_1 -sized sample $(\mathbb{P}_{\mathbf{X}}^{(i)}, \mathcal{R}^{(i)})_{1 \leq i \leq n_1}$ of $(\mathbb{P}_{\mathbf{X}}, \mathcal{R})$. For this, we could consider a double loop Monte Carlo approach. In the outer loop, at each iteration i , a distribution $\mathbb{P}_{\mathbf{X}}^{(i)}$ is randomly generated from $\mathbb{P}_{\mathbb{P}_{\mathbf{X}}} = \mathbb{P}_{\mathbb{P}_{X_1}} \times \dots \times \mathbb{P}_{\mathbb{P}_{X_d}}$. The quantity of

interest $\mathcal{R}^{(i)}$ associated to this distribution $\mathbb{P}_{\mathbf{X}}^{(i)}$ is provided by a 2nd loop. This inner loop consists in generating a n_2 -sized sample $(X_1^{(i,j)}, \dots, X_d^{(i,j)})_{1 \leq j \leq n_2}$ where \mathbf{X} follows $\mathbb{P}_{\mathbf{X}}^{(i)}$. The n_2 corresponding outputs $(Y^{(i,j)})_{1 \leq j \leq n_2}$ are computed in this inner loop. Once this loop performed, the quantity of interest $\mathcal{R}^{(i)}$ is computed from $\mathcal{E}^{(i)} = (X_1^{(i,j)}, \dots, X_d^{(i,j)}, Y^{(i,j)})_{1 \leq j \leq n_2}$. This process is repeated for each $\mathbb{P}_{\mathbf{X}}^{(i)}$ of the outer loop. At the end, 2nd-level HSIC can be estimated from the sample $(\mathbb{P}_{\mathbf{X}}^{(i)}, \mathcal{R}^{(i)})_{1 \leq i \leq n_1}$ by:

$$\widehat{\text{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R}) = \frac{1}{n_1^2} \text{Tr}(L_{\mathcal{D}_k} H L_{\mathcal{R}} H), \quad (21)$$

where $L_{\mathcal{D}_k}$ and $L_{\mathcal{R}}$ are the matrices defined for all $(i, j) \in \{1, \dots, n_1\}$ by: $(L_{\mathcal{D}_k})_{i,j} = l_{\mathcal{D}_k}(\mathbb{P}_{X_k}^{(i)}, \mathbb{P}_{X_k}^{(j)})$, $(L_{\mathcal{R}})_{i,j} = l_{\mathcal{R}}(\mathcal{R}^{(i)}, \mathcal{R}^{(j)})$ and H the matrix defined in Formula (5).

From 2nd-level HSIC estimators, 2nd-level R_{HSIC}^2 indices can be estimated using plug-in and Formula (21) by:

$$\widehat{R}_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R}) = \frac{\widehat{\text{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R})}{\sqrt{\widehat{\text{HSIC}}(\mathbb{P}_{X_k}, \mathbb{P}_{X_k}) \widehat{\text{HSIC}}(\mathcal{R}, \mathcal{R})}}. \quad (22)$$

Consequently, this Monte Carlo double-loop approach requires a total of $n_1 n_2$ code simulations. For example, if $n_1 = 100$ and $n_2 = 1000$, the computation of 2nd-level sensitivity indices HSIC requires 10^5 code calls. This approach is therefore not tractable for CPU-time expensive simulators.

To overcome this problem and reduce the number of code-calls, we propose a single-loop Monte Carlo approach to obtain the sample $(\mathbb{P}_{\mathbf{X}}^{(i)}, \mathcal{R}^{(i)})_{1 \leq i \leq n_1}$, which requires only n_2 simulations, and allows to consider a large sample \mathcal{P} of distributions $\mathbb{P}_{\mathbf{X}}$. This new algorithm is detailed in the next section.

3.3 Algorithm for computing 2nd-level sensitivity indices with a single Monte Carlo loop

In this part, we detail the algorithm to estimate the 2nd-level HSIC (and R_{HSIC}^2) from a unique inputs/output sample \mathcal{E} . We assume that inputs are generated from a unique and known probability distribution denoted $\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{X_1} \times \dots \times \mathbb{P}_{X_d}$ with density denoted $\bar{f}(x_1, \dots, x_d) = \bar{f}_1(x_1) \times \dots \times \bar{f}_d(x_d)$. The options for choosing \bar{f} will be discussed in Section 3.5. The algorithm consists of 3 steps:

► **Step 1. Build a unique n_2 -sized sample \mathcal{E} from \bar{f}**

In this step, we first draw a sample $\bar{\mathbf{X}} = (\mathbf{X}^{(i)})_{1 \leq i \leq n_2}$ according to \bar{f} , then we compute the associated outputs $\bar{Y} = (Y^{(i)})_{1 \leq i \leq n_2}$, to obtain a sample $\mathcal{E} = (\bar{\mathbf{X}}, \bar{Y})$ of inputs/output. Thus, in what follows, all the formulas for modified HSIC will be used with the alternative sample \mathcal{E} , \bar{f} being the alternative distribution. Hence, in all the modified HSIC formulas, the alternative sample will be $(\bar{\mathbf{X}}, \bar{Y}) = (\bar{\mathbf{X}}, \bar{Y})$.

► **Step 2. Perform n_1 GSA1 from \mathcal{E}**

First, we generate a n_1 -sized sample of input distributions according to $\mathbb{P}_{\mathbf{X}}$. This sample of distributions is denoted $\mathcal{P} = (\mathbb{P}_{\mathbf{X}}^{(i)})_{1 \leq i \leq n_1}$ and the density associated to each distribution $\mathbb{P}_{\mathbf{X}}^{(i)}$ is denoted $f^{(i)} = (f_1^{(i)}, \dots, f_d^{(i)})$. The objective is then to compute the GSA1 results $\mathcal{R}^{(i)}$ associated to each distribution $\mathbb{P}_{\mathbf{X}}^{(i)}$, using only \mathcal{E} . The options proposed for $\mathcal{R}^{(i)}$ in Section 3.2.1 are distinguished:

- **Vector $\mathcal{R}^{(i)} = (R_{\text{HSIC},1}^{2,(i)}, \dots, R_{\text{HSIC},d}^{2,(i)})$ of sensitivity indices.** In this case, each $R_{\text{HSIC}}^{2,(i)}$ is estimated by $\widetilde{R}_{\text{HSIC},k}^2$ given by Equation (14) with $\mathcal{E} = (\bar{\mathbf{X}}, \bar{Y})$.
- **Ranking of inputs X_1, \dots, X_d using the indices $R_{\text{HSIC},1}^2, \dots, R_{\text{HSIC},d}^2$.** These rankings are obtained by ordering the coordinates of $\widetilde{R}_{\text{HSIC}}^2$ vectors; still estimated from \mathcal{E} and Equation (14).

- **Vector $\mathcal{R}^{(i)} = (P_1^{(i)}, \dots, P_d^{(i)})$ of p-values associated with asymptotic independence tests.** By analogy with Equation (8), each $P_k^{(i)}$ is estimated thanks to the properties of the modified estimators:

$$\tilde{P}_k^{(i)} \simeq 1 - \tilde{F}_{G_k} \left(n_2 \times \widetilde{\text{HSIC}}(X_k^{(i)}, Y)_{obs} \right), \quad k = 1, \dots, d \quad (23)$$

where \tilde{F}_{G_k} is the cumulative distribution function of Gamma law approximating the asymptotic law of $n_2 \times \widetilde{\text{HSIC}}(X_k^{(i)}, Y)$.

- **Vector $\mathcal{R}^{(i)} = (p_1^{(i)}, \dots, p_d^{(i)})$ of p-values associated with permutation independence tests.** Using the same notations as in Formula (9), each $p_k^{(i)}$ is estimated by:

$$\tilde{p}_k^{(i)} = \frac{1}{B} \sum_{b=1}^B \mathbb{1}_{\widetilde{\text{HSIC}}^{[b]}(X_k^{(i)}, Y) > \widetilde{\text{HSIC}}(X_k^{(i)}, Y)}, \quad k = 1, \dots, d. \quad (24)$$

► Step 3. Estimate 2nd-level sensitivity indices

Finally, the 2nd-level sensitivity indices $R_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ are estimated with the sample $(\mathbb{P}_{\mathbf{X}}^{(i)}, \tilde{\mathcal{R}}^{(i)})_{1 \leq i \leq n_1}$ using Formulas (21) and (22). The computation of matrices $L_{\mathcal{D}_k}$, $k = 1, \dots, d$ and $L_{\mathcal{R}}$ requires the definition of specific RKHS kernels $l_{\mathcal{D}_k}$, $k = 1, \dots, d$ and $l_{\mathcal{R}}$. This item is detailed in the next section.

3.4 Choice of characteristic kernels for probability distributions and for quantities of interest

In this part, we present examples of characteristic RKHS kernels for probability distributions and for the different quantities of interest \mathcal{R} , these kernels being involved in Formula (21) (and as a result in Equation (22)).

Characteristic RKHS kernel for probability distributions. Before defining a characteristic kernel for distributions, we first introduce the Maximum Mean Discrepancy (MMD) defined in [21]. If we consider two distributions \mathbb{P}_1 and \mathbb{P}_2 having the same support and if K denotes a RKHS kernel defined on the common support of \mathbb{P}_1 and \mathbb{P}_2 , then the MMD between \mathbb{P}_1 and \mathbb{P}_2 induced by K is defined as:

$$\text{MMD}_K(\mathbb{P}_1, \mathbb{P}_2) = \mathbb{E}[K(Z_1, Z'_1)] - 2\mathbb{E}[K(Z_1, Z_2)] + \mathbb{E}[K(Z_2, Z'_2)], \quad (25)$$

where Z_1, Z_2 are random variables respectively with laws $\mathbb{P}_1, \mathbb{P}_2$ and Z'_1, Z'_2 are independent and identically distributed copies respectively of Z_1, Z_2 .

Authors of [21] establish that when K is characteristic, the MMD associated to K defines a distance. From MMD distance, [39] defines Gaussian RKHS kernels between probability distributions in a similar way to Formula (3):

$$l_{\mathcal{D}}(\mathbb{P}_1, \mathbb{P}_2) = \exp(-\lambda \text{MMD}_K^2(\mathbb{P}_1, \mathbb{P}_2)), \quad (26)$$

where λ is a positive real parameter.

It has been shown in [7] that when the common support of distributions is compact, the Gaussian MMD-based kernel is universal (and consequently characteristic). We can then define kernels $l_{\mathcal{D}_k}$, $k = 1, \dots, d$ introduced in Formula (19) by:

$$l_{\mathcal{D}_k}(\mathbb{P}_{X_k}, \mathbb{P}_{X'_k}) = \exp\left(-\lambda_k \text{MMD}_{l_k}^2(\mathbb{P}_{X_k}, \mathbb{P}_{X'_k})\right), \quad (27)$$

where λ_k , $k = 1, \dots, d$ are positive real parameters.

From a practical point of view, one can choose λ_k as the inverse of s_k^2 , the empirical variance w.r.t MMD distance (i.e. $\lambda_k = 1/s_k^2$):

$$s_k^2 = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \text{MMD}_{l_k}^2\left(\mathbb{P}_{X_k}^{(i)}, \overline{\mathbb{P}}_{X_k}\right),$$

where the distribution $\bar{\mathbb{P}}_{X_k}$ is defined as, $\bar{\mathbb{P}}_{X_k} = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{P}_{X_k}^{(i)}$.

Characteristic RKHS kernel for permutations as quantity of interest. When GSA1 results \mathcal{R} is a permutation (see Section 3.2.1), we propose to use Mallows kernel K_M [27], the Mallows kernel is universal (and characteristic) [31]. This kernel is given, for two permutations σ, σ' by:

$$K_M(\sigma, \sigma') = \exp(-\lambda n_d(\sigma, \sigma')), \quad (28)$$

where λ is a positive real parameter and n_d is the number of discordant pairs between σ and σ' :

$$n_d(\sigma, \sigma') = \sum_{1 \leq r < s \leq d} [\mathbb{1}_{\{\sigma(r) < \sigma(s)\}} \mathbb{1}_{\{\sigma'(r) > \sigma'(s)\}} + \mathbb{1}_{\{\sigma(r) > \sigma(s)\}} \mathbb{1}_{\{\sigma'(r) < \sigma'(s)\}}]. \quad (29)$$

In practice, if a n_1 -sample of σ is available, we propose to choose λ as the inverse of the empirical mean of $n_d(\sigma, \sigma')$.

Characteristic RKHS kernel for real vectors as quantities of interest. In cases where \mathcal{R} is a vector of \mathbb{R}^d , the usual Gaussian kernel defined in Formula (3) can be considered.

3.5 Possibilities for the unique sampling distribution

We propose three different possibilities for the single draw density $\bar{f}(x_1, \dots, x_d) = \bar{f}_1(x_1) \times \dots \times \bar{f}_d(x_d)$ which is used to generate the unique sample \mathcal{E} in Step 1 of the algorithm (Section 3.3). Note here that the support of each \bar{f}_k , $k = 1, \dots, d$ must be \mathcal{X}_k (the variation domain of X_k , see Section 2). To have a density \bar{f}_k close to the set of all possible densities and more particularly to the most likely ones, we propose to use either mixture distribution or two barycenter distributions, namely the Wasserstein barycenter and the symmetric Kullback-Leibler barycenter.

Option 1: mixture distribution. The mixture density \bar{f}_M (see e.g. [17, 43]) of a random density probability f is defined by:

$$\bar{f}_M = \mathbb{E}_{\mathbb{F}}[f] = \int_{\mathcal{F}} f \, d\mathbb{F}(f), \quad (30)$$

with f lying in \mathcal{F} with probability distribution measure $d\mathbb{F}$.

If \mathbb{F} is discrete over a finite parametric set $\{f_{\theta_0}, \dots, f_{\theta_m}\}$, the mixture density is written as

$$\bar{f}_M = \sum_{r=0}^m f_{\theta_r} \mathbb{F}(f_{\theta_r}). \quad (31)$$

If the density f depends on a parameter which is generated according to a continuous density π over Θ , the mixture density is defined by

$$\bar{f}_M = \int_{\Theta} f_{\theta} \pi(\theta) \, d\theta. \quad (32)$$

Option 2: symmetrized Kullback-Leibler barycenter. The symmetric Kullback-Leibler distance [28] is a distance obtained by symmetrizing the Kullback-Leibler divergence. It is defined for two real distributions μ and ν by:

$$D_K^S(\mu, \nu) = \frac{1}{2} (KL(\mu||\nu) + KL(\nu||\mu)), \quad (33)$$

where $KL(Q_1||Q_2) = \int \log\left(\frac{dQ_1}{dQ_2}\right) dQ_1$ is the Kullback-Leibler divergence.

For a finite set of unidimensional and equiprobable densities $\{f_1, \dots, f_m\}$, the symmetrized Kullback-Leibler barycenter \bar{f}_K can not directly be expressed using densities. However, the distribution Q_K of density \bar{f}_K defined by:

$$\bar{f}_K = \frac{1}{2m} \sum_{r=1}^m f_r + \frac{\left(\prod_{r=1}^m f_r\right)^{\frac{1}{m}}}{2 \int \left(\prod_{r=1}^m f_r\right)^{\frac{1}{m}}}, \quad (34)$$

is a very good approximation of symmetrized Kullback-Leibler barycenter (see [44] for detailed proofs).

To generalize the Formula (34) to a probabilistic set of one-dimensional densities, we propose:

$$\bar{f}_K = \frac{1}{2} \mathbb{E}_{\mathbb{F}}[f] + \frac{e^{\mathbb{E}_{\mathbb{F}}[lnf]}}{2 \int e^{\mathbb{E}_{\mathbb{F}}[lnf]}}. \quad (35)$$

where $\mathbb{E}_{\mathbb{F}}[f]$ and $\mathbb{E}_{\mathbb{F}}[lnf]$ are mixture functions of random functions f and lnf (given by Equation (30)).

Option 3: Wasserstein barycenter distribution. The Wasserstein distance (see e.g. [20, 45]) of order p between two real distributions μ and ν with the same support A is defined by:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{A \times A} |x - y|^p d\gamma(x, y) \right)^{1/p}, \quad (36)$$

where $\Gamma(\mu, \nu)$ is the set of probability measures on $A \times A$ with marginals μ and ν .

Note that in the general case, when referring to the Wasserstein distance (without specifying the order) we refer to Wasserstein distance of order 2.

For a finite set \mathcal{F} of unidimensional and equiprobable densities, the Wasserstein barycenter density [1] is the one whose quantile function is the mean of the quantile functions of the elements of the set \mathcal{F} :

$$\bar{q}_W = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} q_f, \quad (37)$$

where \bar{q}_W denotes the quantile function of Wasserstein barycenter, $|\mathcal{F}|$ the cardinal of the set \mathcal{F} and q_f the quantile function associated to f .

To generalize Formula (37) to a probabilistic set of one-dimensional densities, we propose to use:

$$\bar{q}_W = \mathbb{E}_{\mathbb{F}}[q_f], \quad (38)$$

where $\mathbb{E}_{\mathbb{F}}[q_f]$ is the mixture quantile function of the quantile functions $(q_f)_{f \in \mathcal{F}}$.

4 Application of GSA2 methodology

In this part, our proposed methodology is first applied to an analytical model. The three different options proposed in Section 3.5 to generate the unique sample are studied and compared. Moreover, the benefit of this new methodology comparing to a "double loop" approach is highlighted. Thereafter, the whole methodology is applied to a nuclear study case simulating a severe nuclear reactor accident.

4.1 Analytical example

Our proposed "single loop" methodology for GSA2 is first tested on the analytical model presented in Section 2.2.3. We recall that this model is defined on the set $[0, 1]^3$ by $h(X_1, X_2, X_3) = \sin(X_1) + 1.5 \sin^2(X_2) + 0.5 X_3^4 \sin(X_1)$. The inputs X_1 , X_2 and X_3 are assumed here to be independent and their probability distributions \mathbb{P}_{X_1} , \mathbb{P}_{X_2} et \mathbb{P}_{X_3} can equiprobably be the laws \mathbb{P}_U , \mathbb{P}_T et \mathbb{P}_N , where

- \mathbb{P}_U is the uniform distribution on $[0, 1]$,
- \mathbb{P}_T is the triangular distribution on $[0, 1]$ with mode 0.4,
- \mathbb{P}_N is the truncated normal distribution on $[0, 1]$ with mean 0.6 and standard deviation 0.2.

We want here to estimate from a single inputs/output sample, the 2nd-level GSA indices $R_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R})_{k=1\dots d}$ indices of the model $Y = h(\mathbf{X})$ for different sample sizes. For this, we use HSIC measures for GSA1 (and $R_{\text{HSIC}}^2(X_k, Y)_{k=1\dots d}$ indices) with standardized Gaussian kernel. We characterize GSA1 results by the vector of 1st-level R_{HSIC}^2 (option 1 in Section 3.3). To compute the 2nd-level R_{HSIC}^2 indices, MMD-based kernels $l_{\mathcal{D}_k}$ (Equation 26) are used for input distributions and the standardized Gaussian kernel (Equation 3) is used for GSA1 results.

Remark 4.1. *The other quantities of interest characterizing GSA1 results could be studied in a similar way.*

4.1.1 Computation of theoretical values

In order to compute theoretical values of 2nd-level HSIC and R_{HSIC}^2 indices, we consider the finite set of the $n_1 = 27$ possible triplet of input probability distributions. The 1st-level R_{HSIC}^2 vector associated to each distribution is then computed with a sample of size $n_2 = 1000$ (which ensures the convergence of HSIC estimators). Theoretical values of 2nd-level HSIC are estimated with Formula (21):

- $\text{HSIC}(\mathbb{P}_{X_1}, \mathcal{R}) = 0.0414$,
- $\text{HSIC}(\mathbb{P}_{X_2}, \mathcal{R}) = 0.0261$,
- $\text{HSIC}(\mathbb{P}_{X_3}, \mathcal{R}) = 0.0009$.

The theoretical values of 2nd-level $R_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ indices can also be computed:

- $R_{\text{HSIC}}^2(\mathbb{P}_{X_1}, \mathcal{R}) = 0.4152$,
- $R_{\text{HSIC}}^2(\mathbb{P}_{X_2}, \mathcal{R}) = 0.2516$,
- $R_{\text{HSIC}}^2(\mathbb{P}_{X_3}, \mathcal{R}) = 0.0086$.

We observe that $R_{\text{HSIC}}^2(\mathbb{P}_{X_1}, \mathcal{R})$ is considerably larger than $R_{\text{HSIC}}^2(\mathbb{P}_{X_2}, \mathcal{R})$, while $R_{\text{HSIC}}^2(\mathbb{P}_{X_3}, \mathcal{R})$ is negligible compared to the other two. In this example, the lack of knowledge on \mathbb{P}_{X_3} has therefore no influence on 1st-level R_{HSIC}^2 . Furthermore, the uncertainty on \mathbb{P}_{X_1} has a much higher impact than the one of \mathbb{P}_{X_2} , which remains non-negligible. Consequently, characterization efforts must be targeted in priority on \mathbb{P}_{X_1} , followed-up by \mathbb{P}_{X_2} .

4.1.2 GSA2 with our single loop approach

In the following, $\widetilde{\text{HSIC}}_M(\mathbb{P}_{X_k}, \mathcal{R})$, $\widetilde{\text{HSIC}}_W(\mathbb{P}_{X_k}, \mathcal{R})$ and $\widetilde{\text{HSIC}}_K(\mathbb{P}_{X_k}, \mathcal{R})$, $k = 1, \dots, 3$ denote the 2nd-level HSIC measures respectively associated to mixture law, Wasserstein barycenter law and symmetrized Kullback-Leibler barycenter law. Similarly, $\widetilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_k}, \mathcal{R})$, $\widetilde{R}_{\text{HSIC},W}^2(\mathbb{P}_{X_k}, \mathcal{R})$ and $\widetilde{R}_{\text{HSIC},K}^2(\mathbb{P}_{X_k}, \mathcal{R})$, $k = 1, \dots, 3$ are the derived 2nd-level R_{HSIC}^2 indices.

In this section, we apply the methodology proposed in Section 3.3 to estimate GSA2 HSIC-based indices. For this, we consider Monte Carlo samples of sizes $n_2 = 100$ to $n_2 = 1500$. The estimations are repeated independently 200 times from independent samples. The results obtained with the three modified laws are given by Figure 2. The theoretical values of $R_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ are represented in dotted lines. In this case, the estimators $\widetilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_k}, \mathcal{R})$ and $\widetilde{R}_{\text{HSIC},K}^2(\mathbb{P}_{X_k}, \mathcal{R})$ seem to have similar behaviors for both small and higher sample sizes. The dispersion of these two estimators remains high for small sizes (especially for $n_2 \leq 200$) and becomes satisfying from $n_2 = 700$. The estimators $\widetilde{R}_{\text{HSIC},W}^2(\mathbb{P}_{X_k}, \mathcal{R})$ have a higher variance than the two previous estimators, particularly for small and medium sample sizes ($300 \leq n_2 \leq 700$).

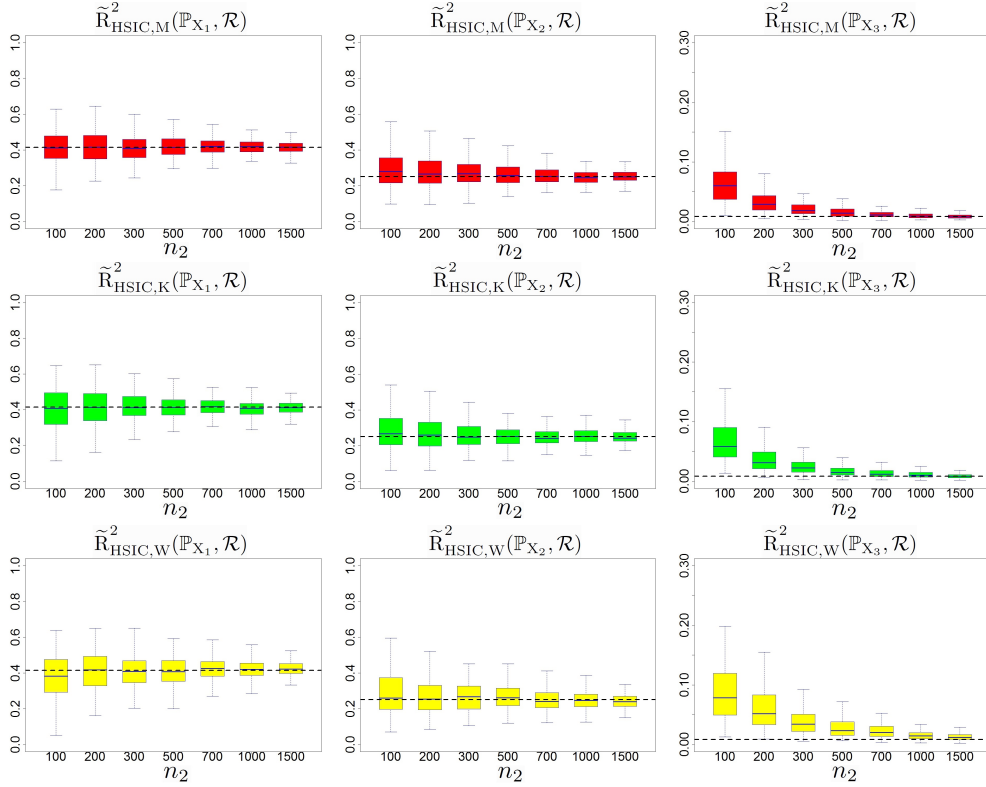


Figure 2: Convergence plots of the estimators $\tilde{R}_{\text{HSIC,M}}^2(\mathbb{P}_{X_k}, \mathcal{R})$, $\tilde{R}_{\text{HSIC,W}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ and $\tilde{R}_{\text{HSIC,K}}^2(\mathbb{P}_{X_k}, \mathcal{R})$, $k = 1, \dots, 3$, for the model h and w.r.t the size n_2 of samples. Theoretical values are represented in dotted lines.

In addition, we compare the ability of the three estimators to correctly order \mathbb{P}_{X_k} , $k = 1 \dots 3$ by order of influence. For this, we compute for each sample size, the ratio of times when they give the good theoretical ranking. Table (2) gives the good ranking rates of 2nd-level $\tilde{R}_{\text{HSIC}}^2$ estimators w.r.t the sample size. These results confirm that the estimators based on mixture and Kullback-Leibler barycenter laws outperform those based on Wasserstein barycenter law. Both $\tilde{R}_{\text{HSIC,M}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ and $\tilde{R}_{\text{HSIC,K}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ yield highly accurate ranking from $n_2 = 500$ against $n_2 = 700$ for $\tilde{R}_{\text{HSIC,W}}^2(\mathbb{P}_{X_k}, \mathcal{R})$. Furthermore, the Kullback-Leibler barycenter seems to give slightly better results for small samples $n_2 \leq 300$, this being reversed from $n_2 = 500$. The lower performance of Wasserstein barycenter law could be explained by the fact that the ratio $\frac{f}{f_W}$ becomes very high in the neighborhood of 0.

| n_2 | 100 | 200 | 300 | 500 | 700 | 1000 | 1500 |
|--|-------|-----|-----|-------|-----|-------|-------|
| $\tilde{R}_{\text{HSIC,M}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ | 74% | 79% | 84% | 94.5% | 97% | 100% | 100% |
| $\tilde{R}_{\text{HSIC,K}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ | 75.5% | 79% | 87% | 92% | 97% | 99.5% | 99.5% |
| $\tilde{R}_{\text{HSIC,W}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ | 57.5% | 71% | 77% | 82% | 91% | 93.5% | 98% |

Table 2: Good ranking rates of $(\mathbb{P}_{X_i})_{i=1..3}$ given by the estimators $\tilde{R}_{\text{HSIC,M}}^2(\mathbb{P}_{X_k}, \mathcal{R})$, $\tilde{R}_{\text{HSIC,K}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ and $\tilde{R}_{\text{HSIC,W}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ for the model h , w.r.t the size n_2 of samples.

4.1.3 Comparison with Monte Carlo "double loop" approach

In this part, we compare the "single loop" estimation of 2nd-level HSIC measures with the "double loop" estimation. For this, we consider a total budget $n = 1026$ simulations for both methods and propose the following test:

- **For the "double loop" approach**, a sample of size $n_2 = 38$ is generated for each triplet of input distributions ($n = n_1 \times n_2 = 1026$ simulations). The computed "double loop" estimators are denoted $\hat{R}_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R}), k = 1 \dots 3$.
- **For the "single loop" approach**, we apply the proposed methodology with $n_2 = 1026$ to compute the "single loop" estimators $\tilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_k}, \mathcal{R})$ and $\tilde{R}_{\text{HSIC},K}^2(\mathbb{P}_{X_k}, \mathcal{R}), k = 1 \dots 3$.

This numerical test is repeated 200 times with independent Monte Carlo samples. Figure 3 shows the dispersion of the obtained estimators. Theoretical values are shown in dotted lines. We observe that the "double loop" estimators have much more variability than "single loop" ones (especially for the distribution \mathbb{P}_{X_3}). We even observe a much larger bias (especially for \mathbb{P}_{X_3}) for the "double loop" approach. Good ranking rates are given by Table 3 and confirm that our proposed "single loop" approach significantly outperforms the "double loop" approach.

This example illustrates the interest of the "simple loop" approach which allows a much more accurate estimation of 2nd-level HSIC measures. Indeed, for a given total budget of n simulations, 1st-level HSIC are computed via modified HSIC from $n_2 = n$ simulations in our "single loop" approach against $n_2 = n/n_1$ in the "double loop" one. Even if classical estimators converge faster than modified ones, the number of simulations available for their estimation is drastically reduced with the double loop approach.

On this same analytical function, other numerical tests with different hypothesis on the input distributions (more different from each other) have been performed and yield similar results and conclusions.

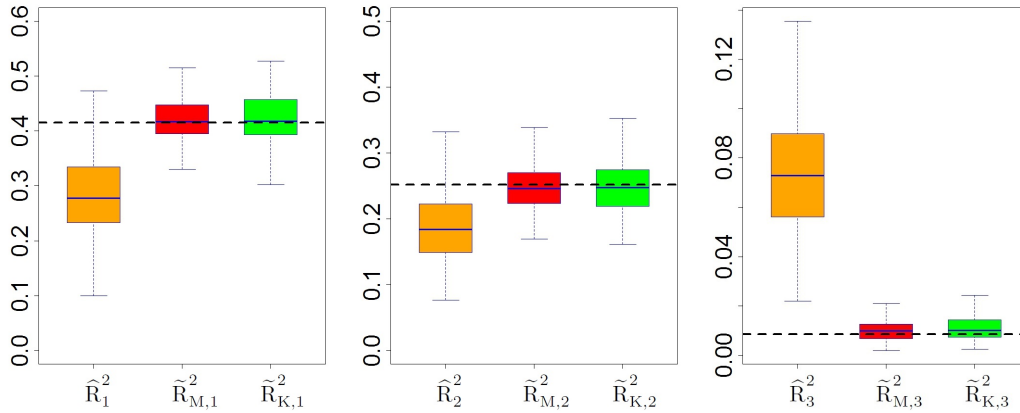


Figure 3: Comparison of convergence plots of 2nd-level GSA indices by "double loop" approach ($\hat{R}_k^2 = \hat{R}_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R})$) and by "single loop" approach ($\tilde{R}_{M,k}^2 = \tilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_k}, \mathcal{R})$ and $\tilde{R}_{K,k}^2 = \tilde{R}_{\text{HSIC},K}^2(\mathbb{P}_{X_k}, \mathcal{R})$) for the model h and $n = 1026$. Theoretical values are represented in dotted lines.

| Double loop | Single loop | |
|--|--|--|
| $\hat{R}_{\text{HSIC}}^2(\mathbb{P}_{X_k}, \mathcal{R})$ | $\tilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_k}, \mathcal{R})$ | $\tilde{R}_{\text{HSIC},K}^2(\mathbb{P}_{X_k}, \mathcal{R})$ |
| 67.5% | 100% | 99% |

Table 3: Comparison of good ranking rates of "double loop" and "simple loop" estimators, for model h and $n = 1026$.

4.2 Nuclear safety application

Within the framework of 4th-generation sodium-cooled fast reactor ASTRID: Advanced Sodium Technological Reactor for Industrial Demonstration (see Figure 4), the CEA (French *Commissariat à l'Énergie atomique et aux Énergies alternatives*) provides numerical tools in order to assess the safety in case of several accidents. For this, various physical modelling tools have been developed to study different severe accidental scenarios. Among these physical modelling tools, a numerical tool called MACARENa (French: *Modélisation de l'Accident d'Arrêt des pompes d'un REacteur refroidi au sodium*) developed by [15] simulates a primary phase of an Unprotected Loss Of Flow (ULOF) accident. During this type of accident, the power loss of primary pumps and the dysfunction of shutdown systems cause a gradual decrease of the sodium flow in the primary circuit, which subsequently may increase the temperature of sodium until it boils. This temperature increase can lead to a degradation of several components and structures of the reactor core.

Previous GSA studies were performed on MACARENa simulator with several tens of uncertain parameters whose pdf were assumed to be known and set at a reference pdf (see [14] for more details). These studies showed the predominant influence of only 3 parameters on the accident transient predicted by MACARENa:

- X_1 : error of measurement on external pressure loss,
- X_2 : primary half-flow time,
- X_3 : Lockart-Martinelli correction value.

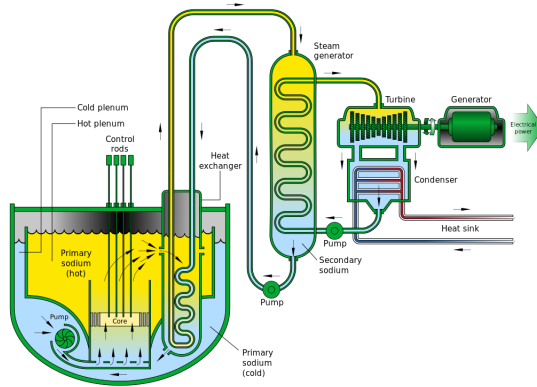


Figure 4: Basic architecture of a Sodium-cooled Fast Reactor

However, due to lack of data and knowledge, uncertainty remains on the distributions \mathbb{P}_{X_1} , \mathbb{P}_{X_2} and \mathbb{P}_{X_3} respectively of X_1 , X_2 and X_3 . To take into account this uncertainty, for each input, the type of law is assumed to be known but one of its parameters is uncertain, as described in Table (4). The notations $\mathcal{N}_t(a, b, m, \sigma)$, $T(a, b, c)$ and $\mathcal{U}(a, b)$ are respectively, the truncated normal law of mean m and standard deviation σ on $[a, b]$, the triangular law on $[a, b]$ with mode c and the uniform law on $[a, b]$.

| Law of input | Nature | Uncertain parameter |
|--------------------|---------------------------------------|---------------------------------------|
| \mathbb{P}_{X_1} | $\mathcal{N}_t(-0.1, 0.1, 0, \sigma)$ | $\sigma \sim \mathcal{U}(0.03, 0.05)$ |
| \mathbb{P}_{X_2} | $T(0, 20, c)$ | $c \sim \mathcal{U}(8, 15)$ |
| \mathbb{P}_{X_3} | $T(0.8, 2, m)$ | $m \sim \mathcal{U}(1, 1.5)$ |

Table 4: Uncertainties on the laws \mathbb{P}_{X_1} , \mathbb{P}_{X_2} and \mathbb{P}_{X_3} .

Among the outputs computed by MACARENa simulator to describe the ULOF accident, we focus on the first instant of sodium boiling denoted Y . The objective is then to assess how each uncertainty on input pdf can impact the results of sensitivity analysis of Y .

Methodological choices. In order to perform GSA2, we apply our proposed algorithm with the following methodological choices (see Section 3.3):

- the unique sample for each input is generated according to the mixture law,
- the quantity of interest characterizing GSA1 results is the vector R_{HSIC}^2 ,
- the RKHS kernel based on the MMD distance is used for input distributions and the standardized Gaussian kernel is used for GSA1 results.

Choices of sample sizes n_1 and n_2 . We consider a Monte Carlo sample of size $n_2 = 1000$ for the unique sample. This choice is motivated by two main reasons, firstly the calculation time of one simulation of MACARENA (between 2 and 3 hours on average) which limits the total number of simulations and secondly the analytical three-dimensional example of Section (4.1) for which a budget of 1000 simulations gave good results. Furthermore, for the sample of distributions, we consider a Monte Carlo sample of $n_1 = 200$ triplets of pdf. These two choices for n_1 and n_2 will then be justified later in this section, by checking the convergence of estimators.

By applying our 2nd GSA methodology, with all these choices, we obtain the following 2nd-level sensitivity indices values:

- $\tilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_1}, \mathcal{R}) = 0.5341$,
- $\tilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_2}, \mathcal{R}) = 0.3317$,
- $\tilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_3}, \mathcal{R}) = 0.0753$.

Consequently, uncertainty on \mathbb{P}_{X_1} mainly impacts GSA1 results, followed by \mathbb{P}_{X_2} , while \mathbb{P}_{X_3} has a negligible impact. Therefore, the efforts of characterization must be targeted on \mathbb{P}_{X_1} to improve the confidence in GSA1 results. Note that, similar results and conclusions are obtained considering for \mathcal{R} the ranking of X_1 , X_2 and X_3 using 1st-level R_{HSIC}^2 .

Remark 4.2. *A deeper analysis of the 200 GSA1 results shows that X_2 is almost all the time the predominant input (99% of cases). On the other hand, the rank of X_3 or X_1 varies: X_3 is the least influential input in 63% of cases, against 37% for X_1 .*

In the light of GSA2 results, this alternation between the rank of X_3 and X_1 is therefore mainly driven by the uncertainty on \mathbb{P}_{X_1} , to a lesser extent by \mathbb{P}_{X_2} , while \mathbb{P}_{X_3} has no impact. Moreover, X_2 whose distribution is not the most influential on GSA1 result, is surprisingly, the most influential inputs on Y . This example illustrates, if necessary, that GSA2 aims to capture an information that is different but complementary to that of GSA1.

In order to assess the accuracy of 2nd-level R_{HSIC}^2 estimation, we use a non-asymptotic bootstrapping approach (see e.g. [16]). For this, we first generate Monte Carlo subsamples with replacement from the initial sample (of 1000 simulations), then we re-estimate 2nd-level R_{HSIC}^2 using these samples. We consider in particular subsamples of sizes $n_2 = 100$ to $n_2 = 800$. For each size, the estimation is repeated independently $B = 20$ times. Furthermore, to reduce computational efforts, we consider a sample of distributions of reduced size $n_1 = 30$ and generated with a space-filling approach. More precisely, the vector (σ, c, m) is sampled with a Maximum Projection Latin Hypercube Design [29] of size $n_1 = 30$ and defined on the cubic domain $[0.03, 0.05] \times [8, 15] \times [1, 1.5]$.

Figure (5) presents as a boxplot the mismatch between the value estimated from the initial sample and the values estimated from subsamples. We first observe a robustness of estimation: the means of estimators seem to match the value given by the initial sample. We notice also high dispersions for small and medium sizes ($n_2 \leq 400$) and small dispersions for medium and big sizes ($n_2 \geq 500$). Therefore, it can be deduced that the estimations of GSA2 indices with the sample of $n_2 = 1000$ simulations have converged, the stabilization of the estimations being satisfactory from $n_2 = 700$.

We also test the robustness of the estimation in terms of ranking of input distributions. Table (5) gives for each subsample size, the rate of times that the ranking matches with the ranking obtained on the initial sample. The results given by Table (5) validate the conclusions drawn from convergence plots (5).

| $n_2 = 100$ | $n_2 = 200$ | $n_2 = 300$ | $n_2 = 400$ | $n_2 = 500$ | $n_2 = 600$ | $n_2 \geq 700$ |
|-------------|-------------|-------------|-------------|-------------|-------------|----------------|
| 45% | 55% | 70% | 75% | 95% | 95% | 100% |

Table 5: Good ranking rates given by the estimators $\tilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_k}, \mathcal{R})$ for MACARENa w.r.t the size n_2 of the unique sample.

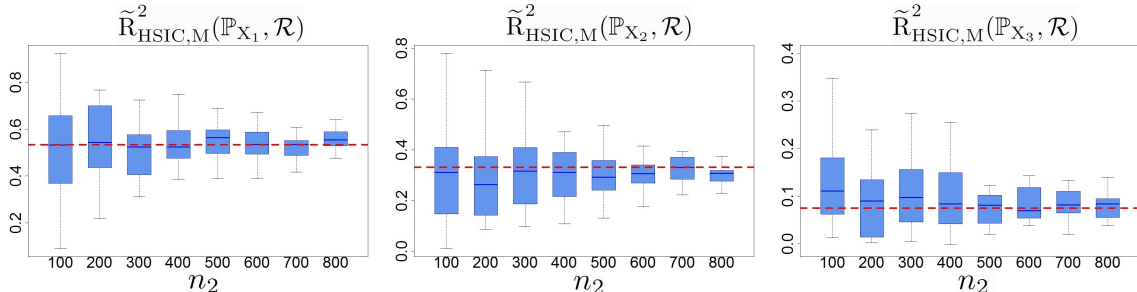


Figure 5: Convergence plots of the estimators $\tilde{R}_{\text{HSIC},M}^2(\mathbb{P}_{X_k}, \mathcal{R})$ for MACARENa, according to the sample size n_2 . Theoretical values are represented in red dashed lines.

5 Conclusion and Prospect

In this article, we proposed a new methodology for second-level Global Sensitivity Analysis (GSA2) based on Hilbert-Schmidt Independence Criterion (HSIC). For this, we first proposed new weighted estimators for HSIC, using an alternative sample generated according to a probability distribution which is not the prior distribution of the inputs. We also demonstrated the properties of these new estimators (bias, variance and asymptotic law), which are similar to those of classical estimators. Moreover, their convergence has been illustrated on an analytical example which has also highlighted their ability to correctly rank variables (even for small and medium sample sizes). Subsequently, 2nd-level GSA based on HSIC measures is discussed. When input distributions are uncertain, GSA2 purpose is to assess the impact of these uncertainties on GSA results. In order to perform GSA2, we presented a new "single loop" Monte Carlo methodology to address problems raised by GSA2: characterization of GSA results, definition of 2nd-level HSIC measures and limitation of the calculation budget. This methodology is based on a single sample generated according to a "reference distribution" (related to the set of all possible distributions). Three options have been proposed for this distribution: mixture law and barycentric laws w.r.t symmetrized Kullback-Leibler distance or Wasserstein distance. The estimation of 2nd-level HSIC seems to be more accurate using the two first options rather than the Wasserstein barycenter. We also illustrated the great interest of the "single loop" approach compared to the "double loop" approach. Finally, the whole methodology has been applied to a nuclear test case simulating a severe reactor accident and has shown how GSA2 can provide additional information to classical GSA.

Several points of the methodology could be more investigated in future research. First, we could focus on comparing Space Filling Design (see e.g. [34], [8], [46]) techniques and Monte Carlo methods for the sampling of input distribution in the case of probabilistic densities (pdf) with uncertain parameters. Indeed, sampling the uncertain parameters of pdf following a space-filling design could improve the accuracy of the estimators of GSA2 indices. Another interesting perspective would be to build independence tests based on 2nd-level HSIC measures estimators. This could be achieved by identifying the asymptotic distributions of these estimators under the assumption of independence between distributions and GSA1 results.

Furthermore, this new approach for GSA2 could also be compared to the classical approach of epistemic GSA in the framework of Dempster-Shafer theory (see [37], [2]). Indeed, Dempster-Shafer theory gives a description of random variables with epistemic uncertainty, which is to associate with an epistemic variable Z on a set A , a mass function representing a probability measure on the set $\mathcal{P}(A)$ of all A -subsets. This lack of knowledge is reflected in Dempster-Shafer theory by an upper and lower bound of the cumulative distribution function and can be viewed as 2nd-level of uncertainty.

An other potential prospect could be to make the connection between our approach and Perturbed-Law based Indices (PLI) [30]. These indices are used to quantify the impact of a perturbation of an input density on the failure probability (probability that a model output exceeds a given threshold). To compare our GSA2 indices with PLI, the probability of failure could be considered as the quantity of interest characterizing GSA results in our methodology. Last but not least, GSA2 method can be compared to the approach proposed in [6] which models 2nd-level uncertainties as a uni-level uncertainty on the vector (Θ, X) , where Θ is the vector of uncertain parameters.

Acknowledgments

We are grateful to Sébastien Da Veiga for his useful ideas and constructive conversations. We also thank Jean-Baptiste Droin for his assistance on the use of MACARENa, and Hugo Raguet for his helpful discussions all along this work.

References

- [1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
- [2] Diego A Alvarez. Reduction of uncertainty using sensitivity analysis methods for infinite random sets of indexable type. *International journal of approximate reasoning*, 50(5):750–762, 2009.
- [3] Nachman Aronszajn. Theory of reproducing kernels. *Matematika*, 7(2):67–130, 1963.
- [4] Charles R Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.
- [5] Claire Cannaméla. *Apport des méthodes probabilistes dans la simulation du comportement sous irradiation du combustible à particules*. PhD thesis, Paris 7, 2007.
- [6] Vincent Chabridon, Mathieu Balesdent, Jean-Marc Bourinet, Jérôme Morio, and Nicolas Gayton. Evaluation of failure probability under parameter epistemic uncertainty: application to aerospace system reliability assessment. *Aerospace Science and Technology*, 69:526–537, 2017.
- [7] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. In *Advances in neural information processing systems*, pages 406–414, 2010.
- [8] Thomas M Cioppa. *Efficient nearly orthogonal and space-filling experimental designs for high-dimensional complex models*. PhD thesis, 2002.
- [9] Imre Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4):191–213, 1972.
- [10] Sebastien Da Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, 2015.
- [11] Guillaume Damblin, Mathieu Couplet, and Bertrand Iooss. Numerical studies of space-filling designs: optimization of latin hypercube samples and subprojection properties. *Journal of Simulation*, 7(4):276–289, 2013.
- [12] Matthias De Lozzo and Amandine Marrel. New improvements in the use of dependence measures for sensitivity analysis and screening. *Journal of Statistical Computation and Simulation*, 86(15):3038–3058, 2016.
- [13] Etienne De Rocquigny, Nicolas Devictor, and Stefano Tarantola. *Uncertainty in industrial practice*. Wiley, 2008.
- [14] Jean-Baptiste Droin. *Modélisation d’un transitoire de perte de débit primaire non protégé dans un RNR-Na*. PhD thesis, Université Grenoble Alpes, 2016.

- [15] Jean-Baptiste Droin, Nathalie Marie, Andrea Bachrata, Frederic Bertrand, Elsa Merle, and Jean-Marie Seiler. Physical tool for unprotected loss of flow transient simulations in a sodium fast reactor. *Annals of nuclear energy*, 106:195–210, 2017.
- [16] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
- [17] B. Everitt and D.J. Hand. *Finite Mixture Distributions*. Monographs on applied probability and statistics. Chapman and Hall, 1981.
- [18] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004.
- [19] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in neural information processing systems*, pages 489–496, 2008.
- [20] Clark R Givens, Rae Michael Shortt, et al. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.
- [21] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.
- [22] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Scholkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, volume 16, pages 63–78. Springer, 2005.
- [23] Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pages 585–592, 2008.
- [24] J.C. Helton, J.D. Johnson, C.J. Sallaberry, and C.B. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, 91(10-11):1175 – 1209, 2006.
- [25] Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pages 101–122. Springer, 2015.
- [26] T Ishigami and Toshimitsu Homma. An importance quantification technique in uncertainty analysis for computer models. In *Uncertainty Modeling and Analysis, 1990. Proceedings., First International Symposium on*, pages 398–403. IEEE, 1990.
- [27] Yunlong Jiao and Jean-Philippe Vert. The kendall and mallows kernels for permutations. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1755–1769, 2018.
- [28] Don Johnson and Sinan Sinanovic. Symmetrizing the kullback-leibler distance. *IEEE Transactions on Information Theory*, 2001.
- [29] V Roshan Joseph, Evren Gul, and Shan Ba. Maximum projection designs for computer experiments. *Biometrika*, 102(2):371–380, 2015.
- [30] Paul Lemaître, Ekatarina Sergienko, Aurélie Arnaud, Nicolas Bousquet, Fabrice Gamboa, and Bertrand Iooss. Density modification-based reliability sensitivity analysis. *Journal of Statistical Computation and Simulation*, 85(6):1200–1223, 2015.
- [31] Horia Mania, Aaditya Ramdas, Martin J Wainwright, Michael I Jordan, Benjamin Recht, et al. On kernel methods for covariates that are rankings. *Electronic Journal of Statistics*, 12(2):2537–2577, 2018.
- [32] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.
- [33] R v Mises. On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18(3):309–348, 1947.

- [34] Luc Pronzato and Werner G Müller. Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22(3):681–701, 2012.
- [35] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.
- [36] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.
- [37] Philippe Smets. What is dempster-shafer’s model. *Advances in the Dempster-Shafer theory of evidence*, pages 5–34, 1994.
- [38] Ilya M Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993.
- [39] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.
- [40] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.
- [41] Zoltán Szabó and Bharath K Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.
- [42] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.
- [43] D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley,, 1985.
- [44] Raymond Veldhuis. The centroid of the symmetrical kullback-leibler distance. *IEEE signal processing letters*, 9(3):96–99, 2002.
- [45] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.
- [46] G Gary Wang and Songqing Shan. Review of metamodeling techniques in support of engineering design optimization. *Journal of Mechanical design*, 129(4):370–380, 2007.

Proofs

A Proof of Proposition 1

In this annex, we prove that:

$$\widetilde{\text{HSIC}}(X_k, Y) = \frac{1}{n^2} \text{Tr} \left(\mathbf{W} \tilde{\mathbf{L}}_k \mathbf{W} H_1 \tilde{\mathbf{L}} H_2 \right).$$

Firstly, we evaluate the matrix $\mathbf{W} \tilde{\mathbf{L}}_k \mathbf{W} H_1 \tilde{\mathbf{L}} H_2$ coefficients before computing its trace. The matrix \mathbf{W} being diagonal, we write for $i, j \in \{1, \dots, n\}$:

$$(\mathbf{W} \tilde{\mathbf{L}}_k \mathbf{W})_{i,j} = (\tilde{\mathbf{L}}_k)_{i,j} W_{i,i} W_{j,j}.$$

The coefficient of the matrix $\mathbf{W} \tilde{\mathbf{L}}_k \mathbf{W} H_1$ indexed by i and j can therefore be computed:

$$\begin{aligned} (\mathbf{W} \tilde{\mathbf{L}}_k \mathbf{W} H_1)_{i,j} &= \sum_{r=1}^n (\tilde{\mathbf{L}}_k)_{i,r} W_{i,i} W_{r,r} (H_1)_{r,j} \\ &= \sum_{r=1}^n (\tilde{\mathbf{L}}_k)_{i,r} W_{i,i} W_{r,r} (\delta_{r,j} - \frac{1}{n} W_{j,j}) \\ &= (\tilde{\mathbf{L}}_k)_{i,j} W_{i,i} W_{j,j} - \frac{1}{n} \sum_{r=1}^n (\tilde{\mathbf{L}}_k)_{i,r} W_{i,i} W_{r,r} W_{j,j}. \end{aligned}$$

Subsequently, the matrix $W \tilde{L}_k W H_1 \tilde{L}$ coefficients are obtained:

$$\begin{aligned}
(W \tilde{L}_k W H_1 \tilde{L})_{i,j} &= \sum_{r=1}^n (W \tilde{L}_k W H_1)_{i,r} \tilde{L}_{r,j} \\
&= \sum_{r=1}^n \left((\tilde{L}_k)_{i,r} W_{i,i} W_{r,r} - \frac{1}{n} \sum_{s=1}^n (\tilde{L}_k)_{i,s} W_{i,i} W_{s,s} W_{r,r} \right) \tilde{L}_{r,j} \\
&= \sum_{r=1}^n (\tilde{L}_k)_{i,r} \tilde{L}_{r,j} W_{i,i} W_{r,r} - \frac{1}{n} \sum_{s=1}^n (\tilde{L}_k)_{i,s} W_{i,i} W_{s,s} \sum_{r=1}^n \tilde{L}_{r,j} W_{r,r}.
\end{aligned}$$

Finally,

$$\begin{aligned}
(W \tilde{L}_k W H_1 \tilde{L} H_2)_{i,j} &= \sum_{r=1}^n (W \tilde{L}_k W H_1 \tilde{L})_{i,r} (H_2)_{r,j} \\
&= \sum_{r=1}^n (W \tilde{L}_k W H_1 \tilde{L})_{i,r} (\delta_{r,j} - \frac{1}{n} W_{r,r}) \\
&= (W \tilde{L}_k W H_1 \tilde{L})_{i,j} - \frac{1}{n} \sum_{r=1}^n (W \tilde{L}_k W H_1 \tilde{L})_{i,r} W_{r,r} \\
&= \sum_{r=1}^n (\tilde{L}_k)_{i,r} \tilde{L}_{r,j} W_{i,i} W_{r,r} - \frac{1}{n} \sum_{1 \leq r, s \leq n} (\tilde{L}_k)_{i,s} \tilde{L}_{r,j} W_{i,i} W_{s,s} W_{r,r} \\
&\quad - \frac{1}{n} \sum_{r=1}^n \left(\sum_{s=1}^n (\tilde{L}_k)_{i,s} \tilde{L}_{s,r} W_{i,i} W_{s,s} - \frac{1}{n} \sum_{1 \leq p, q \leq n} (\tilde{L}_k)_{i,q} \tilde{L}_{p,r} W_{i,i} W_{q,q} W_{p,p} \right) W_{r,r} \\
&= \sum_{r=1}^n (\tilde{L}_k)_{i,r} \tilde{L}_{r,j} W_{i,i} W_{r,r} - \frac{1}{n} \sum_{1 \leq r, s \leq n} (\tilde{L}_k)_{i,s} \tilde{L}_{r,j} W_{i,i} W_{s,s} W_{r,r} \\
&\quad - \frac{1}{n} \sum_{1 \leq r, s \leq n} (\tilde{L}_k)_{i,s} \tilde{L}_{s,r} W_{i,i} W_{s,s} W_{r,r} + \frac{1}{n^2} \sum_{1 \leq r, p, q \leq n} (\tilde{L}_k)_{i,q} \tilde{L}_{p,r} W_{i,i} W_{q,q} W_{p,p} W_{r,r}.
\end{aligned}$$

Summing up the matrix $W \tilde{L}_k W H_1 \tilde{L} H_2$ diagonal terms, then dividing by n^2 gives:

$$\begin{aligned}
\frac{1}{n^2} \text{Tr} \left(W \tilde{L}_k W H_1 \tilde{L} H_2 \right) &= \frac{1}{n^2} \sum_{1 \leq i, r \leq n} (\tilde{L}_k)_{i,r} \tilde{L}_{i,r} W_{i,i} W_{r,r} + \frac{1}{n^4} \sum_{1 \leq i, q \leq n} (\tilde{L}_k)_{i,q} W_{i,i} W_{q,q} \sum_{1 \leq p, r \leq n} \tilde{L}_{p,r} W_{p,p} W_{r,r} \\
&\quad - \frac{2}{n^3} \sum_{1 \leq i, r, s \leq n} (\tilde{L}_k)_{i,s} \tilde{L}_{i,r} W_{i,i} W_{s,s} W_{r,r}.
\end{aligned}$$

By definition of \tilde{L}_k , \tilde{L} and W , the three terms of the last equation are respectively the estimators defined in Formula (12).

B Proof of Proposition 2

Throughout the rest of the document, to lighten formulas, we denote $(l_k)_{i,j} = (\tilde{L}_k)_{i,j}$, $l_{i,j} = \tilde{L}_{i,j}$ and $w_i = W_{i,i}$. We also denote $\widehat{\text{HSIC}}_{\text{U}}(X_k, Y)$ the U-statistic associated to the estimator $\widehat{\text{HSIC}}(X_k, Y)$.

Under the null hypothesis $\mathcal{H}_{0,k}$, the estimator $\widehat{\text{HSIC}}_{\text{U}}(X_k, Y)$ is unbiased. The estimator $\widehat{\text{HSIC}}(X_k, Y)$ bias, is then equal to that of $\widehat{\text{HSIC}}(X_k, Y) - \widehat{\text{HSIC}}_{\text{U}}(X_k, Y)$ under this same assumption. We first compute the expression of $\widehat{\text{HSIC}}(X_k, Y) - \widehat{\text{HSIC}}_{\text{U}}(X_k, Y)$, before computing its expectation. We recall that,

$$\begin{aligned}
\widetilde{\text{HSIC}}(X_k, Y) &= \frac{1}{n^2} \sum_{1 \leq i, j \leq n} (l_k)_{i,j} l_{i,j} w_i w_j + \frac{1}{n^4} \sum_{1 \leq i, j, p, q \leq n} (l_k)_{i,j} l_{p,q} w_i w_j w_p w_q \\
&\quad - \frac{2}{n^3} \sum_{1 \leq i, j, r \leq n} (l_k)_{i,j} l_{i,r} w_i w_j w_r \\
&:= \widetilde{\text{H}}_k^1 + \widetilde{\text{H}}_k^2 - 2\widetilde{\text{H}}_k^3, \\
\widetilde{\text{HSIC}}_{\text{U}}(X_k, Y) &= \frac{1}{\binom{n}{2}} \sum_{(i,j) \in i_2^n} (l_k)_{i,j} l_{i,j} w_i w_j + \frac{1}{\binom{n}{4}} \sum_{(i,j,p,q) \in i_4^n} (l_k)_{i,j} l_{p,q} w_i w_j w_p w_q \\
&\quad - \frac{2}{\binom{n}{3}} \sum_{(i,j,r) \in i_3^n} (l_k)_{i,j} l_{i,r} w_i w_j w_r \\
&:= \widetilde{\text{H}}_k^{1,\text{U}} + \widetilde{\text{H}}_k^{2,\text{U}} - 2\widetilde{\text{H}}_k^{3,\text{U}},
\end{aligned}$$

where $\binom{n}{s} = \frac{n!}{(n-s)!}$ and i_s^n is the set of all s -tuples drawn without replacement from the set $\{1, \dots, n\}$.

Let us compute $\widetilde{\text{HSIC}}(X_k, Y) - \widetilde{\text{HSIC}}_{\text{U}}(X_k, Y)$ term by term:

$$\begin{aligned}
\widetilde{\text{H}}_k^1 - \widetilde{\text{H}}_k^{1,\text{U}} &= \frac{1}{n^2} \sum_{i=1}^n (l_k)_{i,i} l_{i,i} w_i^2 - \frac{1}{n^2(n-1)} \sum_{1 \leq i \neq j \leq n} (l_k)_{i,j} l_{i,j} w_i w_j, \\
\widetilde{\text{H}}_k^2 - \widetilde{\text{H}}_k^{2,\text{U}} &= \frac{1}{n^4} \sum_{(i,j,q) \in i_3^n} [(l_k)_{i,i} l_{j,q} w_i^2 w_j w_q + 4(l_k)_{i,j} l_{i,q} w_i^2 w_j w_q + (l_k)_{i,j} l_{q,q} w_i w_j w_q^2] \\
&\quad - \frac{6}{n \binom{n}{4}} \sum_{(i,j,p,q) \in i_4^n} (l_k)_{i,j} l_{p,q} w_i w_j w_p w_q + O\left(\frac{1}{n^2}\right), \\
\widetilde{\text{H}}_k^3 - \widetilde{\text{H}}_k^{3,\text{U}} &= \frac{1}{n^3} \sum_{1 \leq i \neq j \leq n} [(l_k)_{i,i} l_{i,j} w_i^2 w_j + (l_k)_{i,j} l_{i,i} w_i^2 w_j + (l_k)_{i,j} l_{i,j} w_i w_j^2] \\
&\quad - \frac{3}{n \binom{n}{3}} \sum_{(i,j,r) \in i_3^n} (l_k)_{i,j} l_{i,r} w_i w_j w_r + O\left(\frac{1}{n^2}\right).
\end{aligned}$$

These expressions can be simplified by replacing $(l_k)_{i,i} = l_{i,i} = 1$:

$$\begin{aligned}
\widetilde{\text{H}}_k^1 - \widetilde{\text{H}}_k^{1,\text{U}} &= \frac{1}{n^2} \sum_{i=1}^n w_i^2 - \frac{1}{n^2(n-1)} \sum_{1 \leq i \neq j \leq n} (l_k)_{i,j} l_{i,j} w_i w_j, \\
\widetilde{\text{H}}_k^2 - \widetilde{\text{H}}_k^{2,\text{U}} &= \frac{1}{n^4} \sum_{(i,j,q) \in i_3^n} [l_{j,q} w_i^2 w_j w_q + 4(l_k)_{i,j} l_{i,q} w_i^2 w_j w_q + (l_k)_{i,j} w_i w_j w_q^2] \\
&\quad - \frac{6}{n \binom{n}{4}} \sum_{(i,j,p,q) \in i_4^n} (l_k)_{i,j} l_{p,q} w_i w_j w_p w_q + O\left(\frac{1}{n^2}\right), \\
\widetilde{\text{H}}_k^3 - \widetilde{\text{H}}_k^{3,\text{U}} &= \frac{1}{n^3} \sum_{1 \leq i \neq j \leq n} [l_{i,j} w_i^2 w_j + (l_k)_{i,j} w_i^2 w_j + (l_k)_{i,j} l_{i,j} w_i w_j^2] \\
&\quad - \frac{3}{n \binom{n}{3}} \sum_{(i,j,r) \in i_3^n} (l_k)_{i,j} l_{i,r} w_i w_j w_r + O\left(\frac{1}{n^2}\right).
\end{aligned}$$

By computing the expectation of these three estimators under $\mathcal{H}_{0,k}$, we have:

$$\begin{aligned}
\mathbb{E} \left[\widetilde{\text{H}}_k^1 - \widetilde{\text{H}}_k^{1,\text{U}} \right] &= \frac{1}{n} (\mathbb{E}_\omega - \mathbb{E}_{x_k} \mathbb{E}_y), \\
\mathbb{E} \left[\widetilde{\text{H}}_k^2 - \widetilde{\text{H}}_k^{2,\text{U}} \right] &= \frac{1}{n} (\mathbb{E}_\omega \mathbb{E}_y + 4 \mathbb{E}_{x_k, \omega} \mathbb{E}_{y, \omega} + \mathbb{E}_\omega \mathbb{E}_{x_k}) - \frac{6}{n} \mathbb{E}_{x_k} \mathbb{E}_y + O\left(\frac{1}{n^2}\right), \\
\mathbb{E} \left[\widetilde{\text{H}}_k^3 - \widetilde{\text{H}}_k^{3,\text{U}} \right] &= \frac{1}{n} \left(\mathbb{E}_\omega^k \mathbb{E}_{y, \omega} + \mathbb{E}_\omega^{-k} \mathbb{E}_{x_k, \omega} + \mathbb{E}_{x_k, \omega} \mathbb{E}_{y, \omega} \right) - \frac{3}{n} \mathbb{E}_{x_k} \mathbb{E}_y + O\left(\frac{1}{n^2}\right).
\end{aligned}$$

From these last equations, we obtain:

$$\begin{aligned} \mathbb{E} \left[\widetilde{\text{HSIC}}(X_k, Y) - \widetilde{\text{HSIC}}_{\text{U}}(X_k, Y) \right] &= \frac{2}{n} (\mathbb{E}_{\omega}^k - \mathbb{E}_{x_k, \omega}) (\mathbb{E}_{\omega}^{-k} - \mathbb{E}_{y, \omega}) - \frac{1}{n} (\mathbb{E}_{\omega} - \mathbb{E}_{x_k}) (\mathbb{E}_{\omega} - \mathbb{E}_y) \\ &\quad + \frac{1}{n} \mathbb{E}_{\omega} (\mathbb{E}_{\omega} - 1) + O\left(\frac{1}{n^2}\right). \end{aligned}$$

Finally, The bias of $\widetilde{\text{HSIC}}(X_k, Y)$ under $\mathcal{H}_{0,k}$ is written:

$$\begin{aligned} \mathbb{E}[\widetilde{\text{HSIC}}(X_k, Y)] - \text{HSIC}(X_k, Y) &= \frac{2}{n} (\mathbb{E}_{\omega}^k - \mathbb{E}_{x_k, \omega}) (\mathbb{E}_{\omega}^{-k} - \mathbb{E}_{y, \omega}) - \frac{1}{n} (\mathbb{E}_{\omega} - \mathbb{E}_{x_k}) (\mathbb{E}_{\omega} - \mathbb{E}_y) \\ &\quad + \frac{1}{n} \mathbb{E}_{\omega} (\mathbb{E}_{\omega} - 1) + O\left(\frac{1}{n^2}\right). \end{aligned}$$

C Proof of Proposition 3

In order to compute the variance of $\widetilde{\text{HSIC}}(X_k, Y)$ and to determine its asymptotic law under $\mathcal{H}_{0,k}$, general theorems on V-statistics must be used. For this, we write this last estimator as a single V-statistic. By analogy with theorem 1 of [23], we have:

$$\widetilde{\text{HSIC}}(X_k, Y) = \frac{1}{n^4} \sum_{1 \leq i, j, q, r \leq n} \tilde{h}_{ijqr},$$

where

$$\tilde{h}_{i,j,q,r} = \frac{1}{4!} \sum_{(t,u,v,s)}^{(i,j,q,r)} (l_k)_{t,u} l_{t,u} w_t w_u + (l_k)_{t,u} l_{v,s} w_t w_u w_v w_s - 2(l_k)_{t,u} l_{t,v} w_t w_u w_v,$$

the sum represents all ordered quadruples (t, u, v, s) drawn without replacement from (i, j, q, r) .

This equality is easily obtained by decomposing the last sum into three sums, then by writing that:

$$\begin{aligned} \frac{1}{n^4} \sum_{1 \leq i, j, q, r \leq n} \frac{1}{4!} \sum_{(t,u,v,s)}^{(i,j,q,r)} (l_k)_{t,u} l_{t,u} w_t w_u &= \frac{1}{n^2} \sum_{1 \leq i, j \leq n} (l_k)_{i,j} l_{i,j} w_i w_j, \\ \frac{1}{n^4} \sum_{1 \leq i, j, q, r \leq n} \frac{1}{4!} \sum_{(t,u,v,s)}^{(i,j,q,r)} (l_k)_{t,u} l_{v,s} w_t w_u w_v w_s &= \frac{1}{n^4} \sum_{1 \leq i, j, q, r \leq n} (l_k)_{i,j} l_{q,r} w_i w_j w_q w_r, \\ \frac{1}{n^4} \sum_{1 \leq i, j, q, r \leq n} \frac{1}{4!} \sum_{(t,u,v,s)}^{(i,j,q,r)} (l_k)_{t,u} l_{t,v} w_t w_u w_v &= \frac{1}{n^3} \sum_{1 \leq i, j, r \leq n} (l_k)_{i,j} l_{i,r} w_i w_j w_r. \end{aligned}$$

The result is then obtained by combining the last three equalities.

Remark .1. The U-statistic associated to the estimator $\widetilde{\text{HSIC}}(X_k, Y)$ is written:

$$\widetilde{\text{HSIC}}_{\text{U}}(X_k, Y) = \frac{1}{(n)_4} \sum_{(i,j,q,r) \in i_4^n} \tilde{h}_{ijqr}.$$

Under $\mathcal{H}_{0,k}$, the estimators $n \times \widetilde{\text{HSIC}}(X_k, Y)$ et $n \times \widetilde{\text{HSIC}}_{\text{U}}(X_k, Y)$ have the same asymptotic behavior (see e.g. [36]). Moreover, Hoeffding variance decomposition of $\widetilde{\text{HSIC}}_{\text{U}}(X_k, Y)$ is written:

$$\text{Var} \left(\widetilde{\text{HSIC}}_{\text{U}}(X_k, Y) \right) = \binom{n}{4}^{-1} \sum_{r=1}^4 \binom{r}{4} \binom{n-4}{4-r} \zeta_r,$$

where $\zeta_r = \text{Var} \left(\mathbb{E}[\tilde{h}_{i,j,p,q} \mid X_1, \dots, X_r] \right)$, $r = 1, \dots, 4$.

Moreover, under $\mathcal{H}_{0,k}$, the variance of $\widetilde{\text{HSIC}}_{\text{U}}(X_k, Y)$ converges to 0 in $O(\frac{1}{n^2})$:

$$\text{Var}(\widetilde{\text{HSIC}}_{\text{U}}(X_k, Y)) = \frac{72(n-4)(n-5)}{n(n-1)(n-2)(n-3)} \zeta_2 + O(\frac{1}{n^3}).$$

Under $\mathcal{H}_{0,k}$, $\zeta_2 = \mathbb{E}_{i,j} \left[\mathbb{E}_{q,r}[\tilde{h}_{i,j,p,q}]^2 \right]$, where the notation $\mathbb{E}_{r,s}$ designates the expectation by integrating only w.r.t variables X_r and X_s .

Moreover, by detailing the different terms of $\tilde{h}_{i,j,p,q}$, we easily show that:

$$\begin{aligned} 6\tilde{h}_{ijqr} &= (l_k)_{i,j} w_i w_j (l_{i,j} + l_{q,r} w_q w_r) + (l_k)_{i,q} w_i w_q (l_{i,q} + l_{j,r} w_j w_r) + (l_k)_{i,r} w_i w_r (l_{i,r} + l_{j,q} w_j w_q) \\ &+ (l_k)_{j,q} w_j w_q (l_{j,q} + l_{i,r} w_i w_r) + (l_k)_{j,r} w_j w_r (l_{j,r} + l_{q,i} w_q w_i) + (l_k)_{q,r} w_q w_r (l_{q,r} + l_{i,j} w_i w_j) \\ &- \frac{1}{2} \sum_{\substack{(i,j,q,r) \\ (t,u,v)}} (l_k)_{t,u} w_t w_u (l_{t,v} w_v + l_{u,v} w_v) \end{aligned}$$

where

$$\begin{aligned} (l_k)_{i,\cdot} &= \mathbb{E} \left[\tilde{\text{L}}_k \left(X_k^{(i)}, X_k \right) \omega_k(X_k) \right], & l_{i,\cdot} &= \mathbb{E} \left[\tilde{\text{L}} \left(Y^{(i)}, Y \right) \omega_{-k}(X_{-k}) \right], \\ (l_k)_{\cdot,j} &= \mathbb{E} \left[\tilde{\text{L}}_k \left(X_k^{(j)}, X_k \right) \omega_k(X_k) \right], & l_{j,\cdot} &= \mathbb{E} \left[\tilde{\text{L}} \left(Y^{(j)}, Y \right) \omega_{-k}(X_{-k}) \right], \\ (l_k) &= \mathbb{E} \left[\tilde{\text{L}}_k \left(X'_k, X_k \right) \omega_k(X_k) \right], & l &= \mathbb{E} \left[\tilde{\text{L}} \left(Y', Y \right) \omega_{-k}(X_{-k}) \right]. \end{aligned}$$

We therefore write under $\mathcal{H}_{0,k}$:

$$\begin{aligned} 6E_{q,r} \left[\tilde{h}_{ijqr} \right] &= (\tilde{l}_k)_{i,j} w_i w_j \left(\tilde{l}_{i,j} - \tilde{l}_{i,\cdot} - \tilde{l}_{j,\cdot} + \tilde{l} \right) - (\tilde{l}_k)_{i,\cdot} w_i (l_{i,j} w_j - l_{i,\cdot} - l_{j,\cdot} w_j + l) \\ &- (l_k)_{\cdot,j} w_j (l_{i,j} w_i - l_{i,\cdot} w_i - l_{j,\cdot} + l) - (l_k) (l_{i,j} w_i w_j - l_{i,\cdot} w_i - l_{j,\cdot} w_j + l). \end{aligned}$$

$E_{q,r} \left[\tilde{h}_{ijqr} \right]$ can be estimated empirically by $\frac{1}{6}(\tilde{B})_{i,j}$, where \tilde{B} is the matrix defined in Formula (16).

The variance ζ_2 can be estimated by: $\hat{\zeta}_2 = \frac{1}{36n^2} \mathbf{1}^T (\tilde{B} \odot \tilde{B}) \mathbf{1}$. Formula (16) is then obtained by replacing the expression of $\hat{\zeta}_2$, in Hoeffding's decomposition.

D Proof of Theorem 1

The asymptotic law of the V-statistic $n \times \widetilde{\text{HSIC}}(X_k, Y)$ (as well as the U-statistic $n \times \widetilde{\text{HSIC}}_{\text{U}}(X_k, Y)$) is given by Theorem 5.5.2, page 194 of [36], which gives a formulation of the asymptotic laws of degenerate V-statistics (and U-statistics). Indeed, under $\mathcal{H}_{0,k}$ the statistic $\widetilde{\text{HSIC}}(X_k, Y)$ is degenerate, that is: $\forall (i, j, q, r) \in \{1, \dots, n\} : \mathbb{E} \left[\tilde{h}_{ijqr} \right] = 0$.

Theorem. *Under $\mathcal{H}_{0,k}$ we have the following two law convergence theorems:*

$$\begin{aligned} n \times \widetilde{\text{HSIC}}^V(\tilde{X}_k, \tilde{Y}) &\xrightarrow{\mathcal{L}} \sum_{l=1}^{+\infty} \lambda_l \tilde{z}_l^2, \\ n \times \widetilde{\text{HSIC}}^U(\tilde{X}_k, \tilde{Y}) &\xrightarrow{\mathcal{L}} \sum_{l=1}^{+\infty} \lambda_l (\tilde{z}_l^2 - 1) \end{aligned}$$

where $(\tilde{z}_l)_{l \geq 1}$ are independent and identically distributed random variables of law $\mathcal{N}(0, 1)$ and $(\lambda_l)_{l \geq 1}$ are the eigenvalues of the following operator:

$$A(g) : z \mapsto \int \tilde{h}_{ijqr}(\tilde{z}, \tilde{z}_j, \tilde{z}_q, \tilde{z}_r) g(\tilde{z}_j) dF_{jqr}$$

where dF_{jqr} denotes random variables \tilde{z}_j , \tilde{z}_q and \tilde{z}_r .

To conclude, the distribution $\sum_{l=1}^{+\infty} \lambda_l \tilde{z}_l^2$ can be approximated by a Gamma law according to [23]. In fact, it is an infinite sum of random variables independent of law χ^2 (Chi two). The asymptotic law of the V-statistic $n \times \widetilde{\text{HSIC}}(X_k, Y)$ under $\mathcal{H}_{0,k}$ is a Gamma law, whose parameters can be estimated based on the empirical expectation and variance of $n \times \widetilde{\text{HSIC}}(X_k, Y)$ (see section 2.2.2).