New statistical methodology for second level global sensitivity analysis

Anouar Meynaoui, Amandine Marrel, Béatrice Laurent

New statistical methodology for second level global sensitivity analysis

Anouar Meynaoui *[1,2,3], Amandine Marrel[1,2], and Béatrice Laurent[2,3]

[1]CEA, DES, IRESNE, DER, Cadarache F-13108 Saint-Paul-Lez-Durance, France.
[2]Institut de Mathématiques de Toulouse, UMR5219, F-31062 Toulouse, France.
[3]Université de Toulouse; CNRS, INSA, F-31077 Toulouse, France.

January 5, 2023

Abstract

Numerical simulators are widely used to model physical phenomena and global sensitivity analysis (GSA) aims at studying the global impact of the input uncertainties on the simulator output. To perform GSA, statistical tools based on inputs/output dependence measures are commonly used. We focus here on the Hilbert-Schmidt independence criterion (HSIC). Sometimes, the probability distributions modeling the uncertainty of inputs may be themselves uncertain and it is important to quantify their impact on GSA results. We call it here the second-level global sensitivity analysis (GSA2). However, GSA2, when performed with a Monte Carlo double-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 Monte Carlo single-loop with a limited calculation budget. First, we build a unique sample of inputs and simulator outputs, from a well-chosen probability distribution of inputs. From this 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. Subsequently, we define 2^{nd}-level HSIC-based measures between the 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 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 simulator 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 input [12, 22]. 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 a sample of realizations from the simulator. To choose these numerical simulations, experimental design techniques can be used [10].

Generalities on sensitivity analysis. Sensitivity analysis [34] 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 them into two subgroups: those that mainly influence the output (most influential inputs) and those whose influence on the output can be neglected. More generally, sensitivity analysis can be divided into two main areas:

*anouar.meynaoui@gmail.com

- 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 [23], one of the most popular methods used in industrial applications is based on a variance decomposition of the output [37]. 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 simulations required). More recently, tools based on dependence measures have been proposed for GSA1 purpose [9]. 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 [8]. 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 [43]. Last but not least, the Hilbert-Schmidt independence criterion denoted HSIC [20], 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 [20] 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 [11] 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, ...);
- 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 2^{nd}-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 2^{nd}-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 obstacles. 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 "Monte Carlo double-loop" could be considered. In the outer loop, a Monte Carlo sample of input pdfs is generated, 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 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 simulators. 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. Note that this work was initiated in the framework of Meynaoui's PhD [31], the interested reader could find more technical elements and detailed demonstrations in this document.

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 and 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:

$$Y = \mathcal{M}\left(X_1, \ldots, X_d\right),$$

where $X_1, \ldots, X_d$ and Y are respectively the d uncertain inputs and the uncertain output, evolving in one-dimensional real sets respectively denoted $\mathcal{X}_1, \ldots, \mathcal{X}_d$ and $\mathcal{Y}$. $\mathcal{M}$ denotes the numerical simulator. We note $\mathbf{X} = (X_1, \ldots, X_d)$ 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 $f_1, \ldots, f_d$. Finally, $f : (x_1, \ldots, x_d) \mapsto f_1(x_1) \times \ldots \times f_d(x_d)$ denotes the density of the random vector $\mathbf{X}$. As the model $\mathcal{M}$ 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 realizations) of $\mathcal{M}$ are available. It is therefore assumed in the following that we have a n-sample of inputs and associated outputs $\left(\mathbf{X}^{(i)}, Y^{(i)}\right)_{1 \leq i \leq n}$, where $Y^{(i)} = \mathcal{M}(\mathbf{X}^{(i)})$.

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 $k \in \{1, \ldots, d\}$, [20] associate to X_k a *reproducing kernel Hilbert space* (denoted RKHS, see [4] for more details) $\mathcal{H}_k$ 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 $\mathcal{H}_k$ and $\mathcal{G}$ are respectively denoted $\langle .,. \rangle_{\mathcal{H}_k}$ and $\langle .,. \rangle_{\mathcal{G}}$. Under this RKHS framework, [5] defines the cross-covariance operator C_k between $\mathcal{H}_k$ and $\mathcal{G}$ as the linear operator from $\mathcal{G}$ to $\mathcal{H}_k$ defined for all $h \in \mathcal{H}_k$ and all $g \in \mathcal{G}$ by

$$\langle h, C_k g \rangle_{\mathcal{H}_k} = \text{Cov}\left(h(X_k), g(Y)\right).$$

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 ones). Finally, the Hilbert-Schmidt independence criterion (HSIC) is defined by [20] as the Hilbert-Schmidt norm of the operator C_k:

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

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

Remark 1 *In the following, the notation* $\text{HSIC}(X_k, Y)_{\mathcal{H}_k,\mathcal{G}}$ *is replaced by* $\text{HSIC}(X_k, Y)$ *in order to lighten the expressions.*

Authors of [20] 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:

$$\begin{aligned}\text{HSIC}(X_k, Y) = \mathbb{E}\left[l_k(X_k, X_k')l(Y, Y')\right] + \mathbb{E}\left[l_k\left(X_k, X_k'\right)\right]\mathbb{E}\left[l\left(Y, Y'\right)\right] \\ - 2\mathbb{E}\left[\mathbb{E}\left[l_k\left(X_k, X_k'\right) \mid X_k\right]\mathbb{E}\left[l\left(Y, Y'\right) \mid Y\right]\right],\end{aligned} \tag{2}$$

where $(X_1', \ldots, X_d')$ is an independent and identically distributed copy of $(X_1, \ldots, X_d)$ and $Y' = \mathcal{M}(X_1', \ldots, X_d')$.

Independence characterization. To ensure equivalence between HSIC nullity and independence, the kernels l_k and l must belong to the specific class of *characteristic kernels* [42]. A most commonly used characteristic kernel for real variables is the Gaussian kernel, which is defined for a pair of variables $(z, z') \in \mathbb{R}^q \times \mathbb{R}^q$ by

$$k_\lambda(z, z') = exp\left(-\lambda\|z - z'\|_2^2\right), \tag{3}$$

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

Remark 2 *Despite that theoretically* $\text{HSIC}(X_k, Y) = 0$ *is equivalent to the independence between* X_k *and* Y*, a good choice of the kernel widths is required in practice. Indeed, a wise choice of these parameters guarantees a better behavior of HSIC estimators and better properties of the associated independence tests. Unfortunately, the best choice is unknown in practice, it depends on the joint density of* (X_k, Y)*. For this, intrinsic characteristics of these random variables are usually used. In particular, two main options are usually adopted in practice for the adjustment of* λ *in Equation* (3)*: whether the inverse of empirical variance of* z*, or the inverse of empirical median of* $\|z - z'\|_2^2$ *[11, 40, 49]. In the sequel, we refer to Standardized Gaussian kernel as the one with* λ *being the empirical variance. Note that, some existing works such as [41] propose methods based on cross-validation to suitably select widths. Very recently, [2] proposed aggregated HSIC-based tests: a well-chosen collection of HSIC tests is aggregated through a unique independence test to improve the power.*

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 $\left(\mathbf{X}^{(i)}, Y^{(i)}\right)_{1\leq i\leq n}$ of the inputs/output couple $(\mathbf{X}, Y)$, where $\mathbf{X} = (X_1, \ldots, X_d)$.

Monte Carlo estimation. From Equation (2), authors of [20] propose to estimate each $\text{HSIC}(X_k, Y)$ by

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

where L_k and L are the matrices defined for all $i, j \in \{1, \ldots, n\}$ 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 [45] (named after Richard Von Mises) can also be written in the following more compact form [20]:

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

where H is the matrix defined for all $i, j \in \{1, \ldots, n\}$ by $H_{i,j} = \delta_{i,j} - 1/n$, with $\delta_{i,j}$ the Kronecker symbol between i and j which is equal to 1 if $i = j$ and 0 otherwise.

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 $\mathcal{O}(1/n)$, while its variance converges to 0 in $\mathcal{O}(1/n^2)$. 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 [35] with shape and scale parameters, respectively denoted γ_k and β_k:

$$\gamma_k \simeq \frac{e_k^2}{v_k} \quad \text{and} \quad \beta_k \simeq \frac{n.v_k}{e_k},$$

where e_k and v_k respectively are the expectation and the variance of $\widehat{\text{HSIC}}(X_k, Y)$, i.e.

$$e_k = \mathbb{E}\left[\widehat{\text{HSIC}}(X_k, Y)\right] \quad \text{and} \quad v_k = \text{Var}\left(\widehat{\text{HSIC}}(X_k, Y)\right).$$

The reader can refer to [18] and [11] for more details on e_k and v_k and their estimation.

2.1.3 Use for first-level GSA

Several methods based on HSIC measures have been developed for GSA1. In this section, we mention three possible HSIC-based approaches for screening and ranking the inputs: sensitivity indices [9], asymptotic tests [18] and permutation tests [11].

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

$$\text{R}^2_{\text{HSIC},k} = \frac{\text{HSIC}(X_k, Y)}{\sqrt{\text{HSIC}(X_k, X_k)\,\text{HSIC}(Y, Y)}}. \tag{6}$$

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

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

Asymptotic tests. The independence test between the input X_k and the output Y based on HSIC rejects the independence assumption (hypothesis denoted $\mathcal{H}_{0,k}$), when the p-value[1] of the test based on the statistic $n \times \widehat{\text{HSIC}}(X_k, Y)$ is less than a threshold α (in practice α 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:

$$\text{P}_\text{k} \simeq 1 - \text{F}_{\text{G}_\text{k}}\left(n \times \widehat{\text{HSIC}}(X_k, Y)_{obs}\right), \tag{8}$$

where $\text{F}_{\text{G}_\text{k}}$ 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 B random permutations on the set $\{1, \ldots, n\}$, denoted $(\tau^{[b]})_{1 \leq b \leq B}$. These permutations are applied only to the vector $\mathbf{X}$ of inputs. We thus obtain B bootstrap-samples $\left(\mathbf{X}^{(\tau^{[b]}(i))}, Y^{(i)}\right)_{1 \leq i \leq n}$. 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

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

More details and demonstration of test properties are available in [31] (see Proposition 3.5). In addition, sequential algorithms have been recently proposed by [15], to optimize the number of permutations B, while having reliable p-value estimation.

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 statistical properties of these new HSIC estimators are also presented.

[1] The 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.

2.2.1 Expression and estimation of HSIC measures under an alternative law

We consider here d continuous and independent random variables $\widetilde{\mathbf{X}} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)$ whose densities (different from those of $X_1, \ldots, X_d$) are denoted $\widetilde{f}_1, \ldots, \widetilde{f}_d$. We assume that they have the same supports as $f_1, \ldots, f_d$. The associated output is denoted $\widetilde{Y} = \mathcal{M}(\widetilde{\mathbf{X}})$. Finally, the density of $\widetilde{\mathbf{X}}$ is designated by $\widetilde{f}$.

Changing the probability laws in HSIC expression is based on a technique commonly used in the context of importance sampling (see *e.g.* [21]). 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 $\widetilde{Z}$ with density $f_{\widetilde{Z}}$ whose support is the same as that of f_Z. This gives the following expression for $\mathbb{E}[g(Z)]$:

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

where the notation $\mathbb{E}_{\widetilde{f}}[h(Z)]$ designates the expectation of $h(Z)$ for $Z \sim \widetilde{f}$ and Supp(Z) denotes the support of Z.

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

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

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

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

$$H_k^3 = \mathbb{E}\left[l(\widetilde{Y}, \widetilde{Y}') w(\widetilde{X}) w(\widetilde{X}') \right] \text{ and } H_k^4 = \mathbb{E}\left[\mathbb{E}\left[l_k(\widetilde{X}_k, \widetilde{X}_k') w(\widetilde{X}') \mid \widetilde{X}_k \right] \mathbb{E}\left[l(\widetilde{Y}, \widetilde{Y}') w(\widetilde{X}') \mid \widetilde{Y} \right] w(\widetilde{X}) \right],$$

where $\widetilde{\mathbf{X}}'$ is an independent and identically distributed copy of $\widetilde{\mathbf{X}}$, $\widetilde{Y}' = \mathcal{M}(\widetilde{\mathbf{X}}')$ and $w = f/\widetilde{f}$.

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

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

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

Proposition 1 *Similarly to Equation (5), this estimator can be rewritten as*

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

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

$$\widetilde{L}_k = \left(l_k(\widetilde{X}_k^{(i)}, \widetilde{X}_k^{(j)}) \right)_{1 \leq i,j \leq n}; \quad \widetilde{L} = \left(l(\widetilde{Y}^{(i)}, \widetilde{Y}^{(j)}) \right)_{1 \leq i,j \leq n}; \quad W = Diag\left(w(\widetilde{X}^{(i)}) \right)_{1 \leq i \leq n};$$

$$H_1 = I_n - \frac{1}{n} UW; \quad H_2 = I_n - \frac{1}{n} WU;$$

with I_n is the identity matrix of size n and U the matrix filled with 1.

The proof of this proposition is detailed in Appendix A. Similarly, the sensitivity index $\mathrm{R}^2_{\mathrm{HSIC},k}$ can also be estimated using the sample $\left(\widetilde{\mathbf{X}}^{(i)}, \widetilde{Y}^{(i)}\right)_{1 \leq i \leq n}$ by

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

We can demonstrate that these new estimators have statistical properties similar to those of classical estimators. More precisely, $\widetilde{\mathrm{HSIC}}(X_k, Y)$ is asymptotically unbiased and its bias converges in $\mathcal{O}(1/n)$.

Proposition 2 *Under the hypothesis of independence of X_k and Y, and the assumption $l_k(x_k, x_k) = l(y, y) = 1$, the bias and variance of $\widetilde{\mathrm{HSIC}}(X_k, Y)$ are respectively:*

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

$$\mathrm{Var}\left[\widetilde{\mathrm{HSIC}}(X_k, Y)\right] = \frac{72(n-4)(n-5)}{n(n-1)(n-2)(n-3)}\,\mathbb{E}_{1,2}\left[\mathbb{E}_{3,4}[\widetilde{h}_{1,2,3,4}]^2\right] + \mathcal{O}(1/n^3), \quad (16)$$

where

$$E_\omega = \mathbb{E}\left[\omega^2(\widetilde{X})\right], \qquad E_{x_k} = \mathbb{E}\left[l_k(\widetilde{X}_k, \widetilde{X}'_k)\omega_k(\widetilde{X}_k)\omega_k(\widetilde{X}'_k)\right],$$
$$E_y = \mathbb{E}\left[l(\widetilde{Y}, \widetilde{Y}')\omega_{-k}(\widetilde{X}_{-k})\omega_{-k}(\widetilde{X}'_{-k})\right], \qquad E_{x_k,\omega} = \mathbb{E}\left[l_k(\widetilde{X}_k, \widetilde{X}'_k)\omega_k^2(\widetilde{X}_k)\omega_k(\widetilde{X}'_k)\right],$$
$$E_{y,\omega} = \mathbb{E}\left[l(\widetilde{Y}, \widetilde{Y}')\omega_{-k}^2(\widetilde{X}_{-k})\omega_{-k}(\widetilde{X}'_{-k})\right], \qquad E_\omega^k = \mathbb{E}\left[\omega_k^2(\widetilde{X}_k)\right],$$
$$E_\omega^{-k} = \mathbb{E}\left[\omega_{-k}^2(\widetilde{X}_{-k})\right], \qquad \widetilde{h}_{1,2,3,4} = \frac{1}{4!}\sum_{(t,u,v,s)}^{(1,2,3,4)}\left[(\widetilde{l}_k)_{t,u}\widetilde{l}_{t,u} + (\widetilde{l}_k)_{t,u}\widetilde{l}_{v,s} - 2(\widetilde{l}_k)_{t,u}\widetilde{l}_{t,v}\right],$$

ω, ω_k and ω_{-k} respectively denote the functions $f/\widetilde{f}$, $f_k/\widetilde{f}_k$ and $\omega_{-k} : x_{-k} \mapsto \omega(x_1, \ldots, x_d)/\omega_k(x_k)$, with x_{-k} being the vector extracted from $(x_1, \ldots, x_d)$ by removing the k-th coordinate. Moreover, $\widetilde{X}'_{-k}$ is an independent and identically distributed copy of $\widetilde{X}_{-k}$ and $(\widetilde{l}_k)_{p,q}$, $\widetilde{l}_{p,q}$ respectively denote $l_k(\widetilde{X}_k^{(p)}, \widetilde{X}_k^{(q)})$, $l(\widetilde{Y}^{(p)}, \widetilde{Y}^{(q)})$. Finally, $\sum_{(t,u,v,s)}^{(1,2,3,4)}$ is the sum over all permutations (t, u, v, s) of $(1, 2, 3, 4)$ and $\mathbb{E}_{p,q}$ is the expectation only with respect to $\mathbf{X}_p$ and $\mathbf{X}_q$.

One can also prove that the distribution of $n \times \widetilde{\mathrm{HSIC}}(X_k, Y)$ can be approximated by a Gamma law, whose parameters $\widetilde{\gamma}_k$ and $\widetilde{\beta}_k$ are given by $\widetilde{\gamma}_k = \varepsilon_k^2/\vartheta_k$ and $\widetilde{\beta}_k = n\vartheta_k/\varepsilon_k$, where ε_k and ϑ_k are respectively the expectation and variance of $\widetilde{\mathrm{HSIC}}(X_k, Y)$. Proofs of all the propositions are provided in [31], as well as unbiaised estimators of bias and variance.

Remark 3 *From a practical point of view, the greater $(\mathrm{Var}(\omega_k(\widetilde{X}_k))_{1\leq k\leq d}$, the greater the number of simulations required to accurately estimate $(\mathrm{HSIC}(X_k, Y))_{1\leq k\leq d}$. It is therefore highly recommended to check that $(\mathrm{Var}(\omega_k(\widetilde{X}_k))_{1\leq k\leq d}$ are finite. For instance, in the case of densities with compact supports, it is enough to check that $(\omega_k)_{1\leq k\leq d}$ are finite on their supports.*

2.2.2 Illustration on an analytical example

To illustrate the statistical properties of $\widetilde{\mathrm{HSIC}}$, we consider a numerical application inspired from Ishigami's model [24] and defined on $[0, 1]^3$ by

$$\mathcal{M}(X_1, X_2, X_3) = \sin(X_1) + 1.5\sin^2(X_2) + 0.5X_3^4\sin(X_1), \quad (17)$$

where the inputs X_1, X_2 and X_3 are assumed to be independent and follow a triangular distribution with a mode equal to 0.5.

We consider HSIC measures based on Standardized Gaussian kernel (see Remark 2). To estimate them, we suppose that we have Monte Carlo samples of independent inputs generated from a uniform distribution on $[0, 1]^3$ (modified law). For each sample of size $n = 100$ to $n = 1500$, the estimation process is repeated 200 times, with independent random samples. The convergence of $\widetilde{\mathrm{HSIC}}(X_k, Y)$ estimators is illustrated by Figure 1. Results for $\widehat{\mathrm{HSIC}}(X_k, Y)$ computed with samples generated from the original law (namely triangular) are also given and theoretical values are represented in red dotted lines. We observe that for small sample sizes ($n < 500$), modified estimators $\widetilde{\mathrm{HSIC}}(X_k, Y)$ have more bias and variance than $\widehat{\mathrm{HSIC}}(X_k, Y)$ estimators. But, from size $n = 700$, both estimators have similar behaviors.

In addition, to assess the convergence of ranking, the sensitivity indices $R^2_{\mathrm{HSIC},k}$ are estimated from $\widetilde{\mathrm{HSIC}}(X_k, Y)$ with Equation (14). The inputs are ranked by decreasing indices and the resulting correct ranking rates are given by Table 1. Even for small sample sizes (*e.g* $n = 200$), the modified estimators $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC}}$ have good ranking ability.

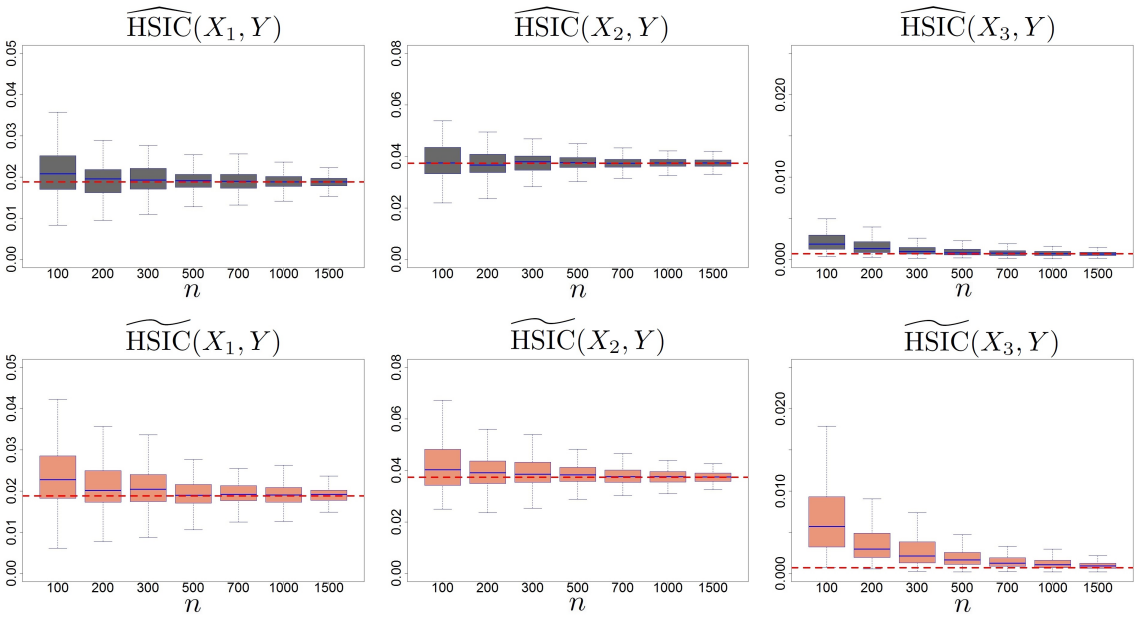

Figure 1: Model $\mathcal{M}$ – Convergence plots of the estimators $\widehat{\text{HSIC}}(X_k, Y)$ and $\widetilde{\text{HSIC}}(X_k, Y)$, according to the sample size n. Theoretical values are represented in red dashed lines.

$n = 100$	$n = 200$	$n = 300$	$n \geqslant 500$
88%	93.5%	97%	100%

Table 1: Model $\mathcal{M}$ – Good ranking rates of inputs based on $\widetilde{\text{R}}^2_{\text{HSIC}}$, for different sample sizes n.

3 New methodology for second-level GSA

We consider that the input probability distributions, $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_d}$, are uncertain. These uncertainties on $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_d}$ are modeled by probability distributions, respectively denoted $\mathbb{P}_{\mathbb{P}_{X_1}}, \ldots, \mathbb{P}_{\mathbb{P}_{X_d}}$. We also assume that the distributions $\mathbb{P}_{\mathbb{P}_{X_1}}, \ldots, \mathbb{P}_{\mathbb{P}_{X_d}}$ are independent and that all possible input distributions have a common support, which the set of all possible input values. Each assumed joint distribution $\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{X_1} \times \ldots \times \mathbb{P}_{X_d}$ of inputs yields potentially different results of 1^{st}-level global sensitivity analysis (GSA1). This impact must be quantified by GSA2. Based on GSA2 results, the probability distributions of inputs could be separated into two groups: those which significantly modify GSA1 results and those whose influence is negligible. Subsequently, the probability distributions with a small impact can be set to a reference distribution and the efforts of characterization will be focused on the most influential ones to improve their knowledge (strategy of uncertainty reduction).

3.1 Issues raised by GSA2

We present in the following the main steps for GSA2 realization and some related technical issues. Our approach is based on the extension of HSIC measures for non-vectorial data. The idea is to define 2^{nd}-level sensitivity indices between input distributions $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_d}$ and GSA1 results. To do so, we first characterize GSA1 results. This means that we associate to each possible input distribution $\mathbb{P}_{\mathbf{X}} = \mathbb{P}_{X_1} \times \ldots \times \mathbb{P}_{X_d}$, a mathematical quantity denoted $\mathcal{R}$ representing the associated GSA1 results. To choose this quantity of interest, we propose the following options, all based on HSIC (see Section 2.1.3):

- **vector of sensitivity indices** $\text{R}^{\mathbf{2}}_{\text{HSIC}} = (\text{R}^2_{\text{HSIC},1}, \ldots, \text{R}^2_{\text{HSIC},d})$;
- **ranking of inputs** $X_1, \ldots, X_d$ using the indices $\text{R}^2_{\text{HSIC},1}, \ldots, \text{R}^2_{\text{HSIC},d}$ **.** This quantity of interest $\mathcal{R}$ is a permutation on the set $\{1, \ldots, d\}$, verifying that $\mathcal{R}(k) = j$ if and only if the variable X_j is the k-th in the ranking;
- **vector of p-values from asymptotic independence tests** $\text{P} = (\text{P}_1, \ldots, \text{P}_\text{d})$;

- **vector of p-values from permuted tests** $\mathrm{p} = (\mathrm{p}_1, \ldots, \mathrm{p}_\mathrm{d})$.

Thanks to the kernel trick, we build 2^{nd}-level HSIC measures between the probability distributions $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_d}$ and the quantity of interest $\mathcal{R}$. Assume that $l_{\mathcal{D}_1}, \ldots, l_{\mathcal{D}_\mathrm{d}}$ and $l_\mathcal{R}$ are RKHS kernels respectively associated to $\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_d}$ and $\mathcal{R}$. Some examples of these kernels are provided in Section 3.3. We define similarly to Equation (2), the 2^{nd}-level HSIC measure between $\mathbb{P}_{X_k}$ and $\mathcal{R}$ as

$$\begin{aligned}\mathrm{HSIC}(\mathbb{P}_{X_k}, \mathcal{R}) = \mathbb{E}\left[l_{\mathcal{D}_\mathrm{k}}(\mathbb{P}_{X_k}, \mathbb{P}'_{X_k}) l_\mathcal{R}(\mathcal{R}, \mathcal{R}')\right] + \mathbb{E}\left[l_{\mathcal{D}_\mathrm{k}}(\mathbb{P}_{X_k}, \mathbb{P}'_{X_k})\right] \mathbb{E}\left[l_\mathcal{R}\left(\mathcal{R}, \mathcal{R}'\right)\right] \\ - 2\mathbb{E}\left[\mathbb{E}\left[l_{\mathcal{D}_\mathrm{k}}(\mathbb{P}_{X_k}, \mathbb{P}'_{X_k}) \mid \mathbb{P}_{X_k}\right] \mathbb{E}\left[l_\mathcal{R}\left(\mathcal{R}, \mathcal{R}'\right) \mid \mathcal{R}\right]\right], \end{aligned} \tag{18}$$

where $(\mathbb{P}'_{X_1}, \ldots, \mathbb{P}'_{X_d})$ is an independent and identically distributed copy of $(\mathbb{P}_{X_1}, \ldots, \mathbb{P}_{X_d})$ and $\mathcal{R}'$ the GSA1 results associated to $(\mathbb{P}'_{X_1}, \ldots, \mathbb{P}'_{X_d})$. The GSA2 indice between $\mathbb{P}_{X_k}$ and $\mathcal{R}'$ is then defined as

$$\mathrm{R}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R}) = \frac{\mathrm{HSIC}(\mathbb{P}_{X_k}, \mathcal{R})}{\sqrt{\mathrm{HSIC}(\mathbb{P}_{X_k}, \mathbb{P}_{X_k})\, \mathrm{HSIC}(\mathcal{R}, \mathcal{R})}}. \tag{19}$$

The estimation of $\mathrm{R}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R})$ requires a n_1-sample $(\mathbb{P}^{(i)}_\mathbf{X}, \mathcal{R}^{(i)})_{1 \leq i \leq n_1}$ of $(\mathbb{P}_\mathbf{X}, \mathcal{R})$. However, the quantities of interest $\mathcal{R}^{(i)}$ are not directly observable, they need to be estimated. To do so, a straightforward double-loop approach could be considered. The outer loop entails to generate the n_1-sized sample of input distribution. On the flip side, the inner loop involves two steps. A n_2-sized sample $(X_1^{(i,j)}, \ldots, X_d^{(i,j)})_{1 \leq j \leq n_2}$ is first generated according to each distribution $\mathbb{P}^{(i)}_\mathbf{X}$, before computing the corresponding outputs $(Y^{(i,j)})_{1 \leq j \leq n_2}$. This allows to estimate the quantity of interest $\mathcal{R}^{(i)}$ associated to each input distribution $\mathbb{P}^{(i)}_\mathbf{X}$. At the end, 2^{nd}-level HSIC can be estimated by

$$\widehat{\mathrm{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R}) = \frac{1}{n_1^2} \mathrm{Tr}(L_{\mathcal{D}_\mathrm{k}} H L_\mathcal{R} H), \tag{20}$$

where $L_{\mathcal{D}_\mathrm{k}}$ and $L_\mathcal{R}$ are the matrices defined for all (i, j) in $\{1, \ldots, n_1\}^2$ as $(L_{\mathcal{D}_\mathrm{k}})_{i,j} = l_{\mathcal{D}_\mathrm{k}}(\mathbb{P}^{(i)}_{X_k}, \mathbb{P}^{(j)}_{X_k})$ and $(L_\mathcal{R})_{i,j} = l_\mathcal{R}(\mathcal{R}^{(i)}, \mathcal{R}^{(j)})$. In addition, the matrix H is defined as in Equation (5). Finally, according to (19), 2^{nd}-level $\mathrm{R}^2_{\mathrm{HSIC}}$ indices can be estimated by

$$\widehat{\mathrm{R}}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R}) = \frac{\widehat{\mathrm{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R})}{\sqrt{\widehat{\mathrm{HSIC}}(\mathbb{P}_{X_k}, \mathbb{P}_{X_k}) \widehat{\mathrm{HSIC}}(\mathcal{R}, \mathcal{R})}}. \tag{21}$$

Consequently, the Monte Carlo double-loop approach requires a total of $n_1 n_2$ simulations. This approach is therefore not tractable for CPU-time expensive simulators, even for reasonable sample sizes n_1 and n_2. To overcome this issue and reduce the number of simulator-calls, we propose in the following a single-loop approach only requiring n_2 simulations, and allowing to consider a larger sample of input distribution.

3.2 Algorithm for computing 2^{nd}-level sensitivity indices with a Monte Carlo single-loop

We provide here a single-loop algorithm for estimating the 2^{nd}-level HSIC measures and indices, respectively defined in Equations (20) and (21). To do so, the inputs are generated according to a unique and known probability distribution, denoted $\bar{\mathbb{P}}_\mathbf{X} = \bar{\mathbb{P}}_{X_1} \times \ldots \times \bar{\mathbb{P}}_{X_d}$. We assume that this distribution has a density $\bar{f} : (x_1, \ldots, x_d) \mapsto \bar{f}_1(x_1) \times \ldots \times \bar{f}_d(x_d)$, and that all possible input distributions also have densities. The procedure is detailed in Algorithm 1.

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

Step 3 of Algorithm 1 involves a choice of kernel, according to the quantities of interest $\mathcal{R}$. A kernel for probability distributions is also required. Some examples of suitable characteristic RKHS kernels are provided in the following.

Characteristic RKHS kernel for probability distributions. The definition of RKHS kernels between distributions is based on the Maximum Mean Discrepancy (MMD), introduced in [19]. Let $\mathbb{Q}_1$ and $\mathbb{Q}_2$ be two distributions with a common support and K be a RKHS kernel on this support. The distance MMD between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ is defined as

$$\mathrm{MMD}_K(\mathbb{Q}_1, \mathbb{Q}_2) = \mathbb{E}\left[K(Z_1, Z_1')\right] - 2\mathbb{E}\left[K(Z_1, Z_2)\right] + \mathbb{E}\left[K(Z_2, Z_2')\right], \tag{24}$$

Algorithm 1 *GSA2 with a Monte Carlo single-loop*

Input: The probability density $\bar{f}$ and an observed n_2-sized sample $\bar{\mathbf{X}} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n_2)}\right)$.

1. **Build a unique n_2-sized sample $\mathcal{E}$ of inputs/output.**
   We compute the output sample $\bar{Y} = \left(Y^{(1)}, \ldots, Y^{(n_2)}\right)$ associated to $\bar{\mathbf{X}}$. The inputs/output sample is denoted $\mathcal{E} = \left(\bar{\mathbf{X}}, \bar{Y}\right)$.
2. **Perform n_1 GSA1 using only $\mathcal{E}$.**
   We draw a n_1-sized sample $\mathbb{P}_{\mathbf{X}}^{(1)}, \ldots, \mathbb{P}_{\mathbf{X}}^{(n_1)}$ of input distributions. Then, we estimate all 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 3.1 are distinguished:
   (a) **Estimate the vector $\mathcal{R}^{(i)} = (\mathrm{R}^{2,(i)}_{\mathrm{HSIC},1}, \ldots, \mathrm{R}^{2,(i)}_{\mathrm{HSIC},d})$ of sensitivity indices.**
   The vector coordinates are estimated using Equation (14), with the alternative sample $\mathcal{E} = \left(\bar{\mathbf{X}}, \bar{Y}\right)$.
   (b) **Rank the inputs $X_1, \ldots, X_d$ using the indices $\mathrm{R}^2_{\mathrm{HSIC},1}, \ldots, \mathrm{R}^2_{\mathrm{HSIC},d}$.**
   The ranking is obtained by ordering the coordinates of the vectors estimated in Option 2a.
   (c) **Estimate the vector $\mathcal{R}^{(i)} = \left(\mathrm{P}_1^{(i)}, \ldots, \mathrm{P}_d^{(i)}\right)$ of p-values associated with asymptotic independence tests.**
   Each $\mathrm{P}_k^{(i)}$ is estimated using the properties of the modified estimators:
   $$\widetilde{\mathrm{P}}_k^{(i)} \simeq 1 - \widetilde{\mathrm{F}}_{\mathrm{G_k}}\left(n_2 \times \widetilde{\mathrm{HSIC}}(X_k^{(i)}, Y)_{obs}\right), \tag{22}$$
   where $\tilde{\mathrm{F}}_{\mathrm{G_k}}$ denotes the Gamma distribution approximating the asymptotic distribution of $n_2 \times \widetilde{\mathrm{HSIC}}(X_k^{(i)}, Y)$.
   (d) **Estimate the vector $\mathcal{R}^{(i)} = \left(\mathrm{p}_1^{(i)}, \ldots, \mathrm{p}_d^{(i)}\right)$ of p-values associated with permutation independence tests.**
   Keeping the same notations of Equation (9), each $\mathrm{p}_k^{(i)}$ is estimated as
   $$\widetilde{\mathrm{p}}_k^{(i)} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\widetilde{\mathrm{HSIC}}^{[b]}(X_k^{(i)}, Y) > \widetilde{\mathrm{HSIC}}(X_k^{(i)}, Y)}. \tag{23}$$
3. **Compute 2nd-level sensitivity indices.**
   Each indice $\mathrm{R}^2_{\mathrm{HSIC}}\left(\mathbb{P}_{X_k}, \mathcal{R}\right)$ is estimated from the sample $\left(\mathbb{P}_{\mathbf{X}}^{(i)}, \widetilde{\mathcal{R}}^{(i)}\right)_{1 \leq i \leq n_1}$ and using Equations (20) and (21).

where Z_1 and Z_1' (respectively Z_2 and Z_2') are independent random variables with common distribution $\mathbb{Q}_1$ (respectively $\mathbb{Q}_2$). From the MMD distance, we consider the radial RKHS distribution kernel defined as

$$l_{\mathcal{D}}\left(\mathbb{Q}_1, \mathbb{Q}_2\right) = \exp\left(-\lambda\, \mathrm{MMD}_K^2\left(\mathbb{Q}_1, \mathbb{Q}_2\right)\right), \tag{25}$$

where λ is a positive real parameter (fixed). The latter RKHS kernel is characteristic, regardless of λ value. This property results from [38, Theorem 4]. However, the parameter λ needs to be properly calibrated for well behaved estimators. Back to our case, we define the kernel $l_{\mathcal{D}_k}$ in Equation (20) by choosing $K = l_k$ and $\lambda = 1/s_k^2$ where

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

with $\mathbb{P}_{X_k}^{\dagger} = 1/n_1 \sum_{i=1}^{n_1} \mathbb{P}_{X_k}^{(i)}$ is the uniformly weighted mixture distribution of $\mathbb{P}_{X_k}^{(1)}, \ldots, \mathbb{P}_{X_k}^{(n_1)}$.

Characteristic RKHS kernel for permutations as quantity of interest. When considering Option 2b in Algorithm 1, we propose the use of the Mallows kernel, brought to light by [25]. The

Mallows kernel is shown to be universal (and characteristic) by [29]. To define this kernel, we first introduce the *number of discordant pairs* between two $\{1,\dots,d\}$-permutations σ_1 and σ_2 as

$$n_D(\sigma_1,\sigma_2) = \sum_{1\leq r<s\leq d} \left[\mathbb{1}_{\{\sigma_1(r)<\sigma_1(s)\}}\mathbb{1}_{\{\sigma_2(r)>\sigma_2(s)\}} + \mathbb{1}_{\{\sigma_1(r)>\sigma_1(s)\}}\mathbb{1}_{\{\sigma_2(r)<\sigma_2(s)\}}\right]. \tag{26}$$

The Mallows kernel K_M between σ_1 and σ_2 is then defined as

$$K_M(\sigma_1,\sigma_2) = \exp\left(-\lambda n_D(\sigma_1,\sigma_2)\right), \tag{27}$$

where λ is a positive real. From a numerical standpoint and based on a sample $\sigma^{(1)},\dots,\sigma^{(n_1)}$, we suggest to take λ as the inverse of the arithmetic mean of $\left\{n_D(\sigma^{(i)},\sigma^{(j)}) \text{ with } 1\leq i<j\leq n_1\right\}$.

Characteristic RKHS kernel for real vectors as quantities of interest. When either Option 2a, 2c or 2d is selected, we can simply use the Standardized Gaussian kernel (see Remark 2).

3.4 Possibilities for the unique sampling distribution

This section deals with the choice of the drawing density $\bar{f}$ in Algorithm 1. Since the inputs are independent, it boils down to choosing each martingale density $\bar{f}_k$. We recall that all possible densities of each input X_k have the same support $\mathcal{X}_k$. The main objective is to choose $\bar{f}_k$ as "close" as possible to the set all potential densities $\mathcal{F}_k$, while accommodating the probability distribution over $\mathcal{F}_k$. Three possibilities for this choice are detailed below: the mixture distribution, the Wasserstein barycenter and the Symmetrical Kullback-Leibler barycenter. For this, we consider here and only here, the following generic notations. We designate by h a random one-dimensional density of distribution $\mathbb{H}$, of which the support is denoted $\mathcal{H}$. We also designate by $\mathcal{S}$ the common support of all realizations of h.

The mixture distribution. We recall that the mixture distribution [16, 44] of h is defined as

$$\bar{h}_M = \mathbb{E}_{\mathbb{H}}[h] = \int_{\mathcal{H}} h \,\mathrm{d}\mathbb{H}(h). \tag{28}$$

In particular, when $\mathbb{H}$ is discrete with support $\mathcal{H}=\{h_1,\dots,h_m\}$, we obtain $\bar{h}_M = \sum_{r=1}^m h_r\,\mathbb{H}(h_r)$. Moreover, if the support $\mathcal{H}$ is parameterizable, i.e. $\mathcal{H}=\{h_\theta,\theta\in\Theta\}$, we have $\bar{h}_M = \int_\Theta h_\theta \pi(\theta)\,\mathrm{d}\theta$, where π is the distribution of the parameter θ.

The Symmetrical Kullback-Leibler barycenter. This barycenter is computed with respect to the so-called *Symmetrical Kullback-Leibler divergence*, which is obtained by symmetrizing the usual Kullback-Leibler divergence [27]. It is defined for two probability measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$ as

$$D_{\mathrm{SKL}}(\mathbb{Q}_1,\mathbb{Q}_2) = \frac{1}{2}\left[\mathbf{KL}(\mathbb{Q}_1||\mathbb{Q}_2) + \mathbf{KL}(\mathbb{Q}_2||\mathbb{Q}_1)\right], \tag{29}$$

where $\mathbf{KL}(\mathbb{Q}_1||\mathbb{Q}_2) = \mathbb{E}_{\mathbb{Q}_1}\left[\log(\mathrm{d}\mathbb{Q}_1/\mathrm{d}\mathbb{Q}_2)\right]$, with $\mathrm{d}\mathbb{Q}_1/\mathrm{d}\mathbb{Q}_2$ refers to the Radon–Nikodym derivative. The explicit formula of the Symmetrical Kullback-Leibler barycenter is unknown. In the specific case where $\mathbb{H}$ is uniform over a finite set $\{h_1,\dots,h_m\}$, a good approximation of this barycenter is shown in [46] and given by

$$\bar{h}_K \simeq \frac{1}{2}\left[\bar{h} + \frac{\tilde{h}}{\int_{\mathcal{S}}\tilde{h}}\right], \tag{30}$$

where $\bar{h} = 1/m\sum_{r=1}^m h_r$ and $\tilde{h} = \prod_{r=1}^m h_r^{1/m}$ are respectively the arithmetic and geometric means of $\{h_1,\dots,h_m\}$. In the general case, we propose to approximate the Symmetrical Kullback-Leibler barycenter as

$$\bar{h}_K \simeq \frac{1}{2}\left[\bar{h}_M + \frac{e^{\overline{\log(h)}_M}}{\int_{\mathcal{S}} e^{\overline{\log(h)}_M}}\right], \tag{31}$$

where $\bar{h}_M$ and $\overline{\log(h)}_M$ are respectively the mixture distributions of the random functions h and $\log(h)$.

The Wasserstein barycenter distribution. We remind that the Wassertein distance [17, 47] between two distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ is defined as

$$\mathrm{W}(\mathbb{Q}_1,\mathbb{Q}_2) = \inf_{\gamma\in\Gamma(\mathbb{Q}_1,\mathbb{Q}_2)}\left(\mathbb{E}_\gamma\left[(X-Y)^2\right]\right)^{1/2}, \tag{32}$$

where $\Gamma(\mathbb{Q}_1, \mathbb{Q}_2)$ is the set of probabilities of (X, Y) with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$. The quantile function of the Wasserstein barycenter [1] of a finite uniformly weighted set $\{h_1, \ldots, h_m\}$ is defined as

$$\bar{q}_{\mathrm{W}} = \frac{1}{m} \sum_{r=1}^{m} q_r, \tag{33}$$

where q_r is the quantile function associated to h_r. In the general case, we extend Equation (33) when h is generated according to a distribution $\mathbb{H}$ as

$$\bar{q}_{\mathrm{W}} = \mathbb{E}_{\mathbb{H}}[q_h], \tag{34}$$

where q_h is the quantile function associated to h.

4 Application of GSA2 methodology

First, the performance of our methodology is studied through simulated data. More specifically, the drawing density options presented in Section 3.4 are studied and compared. Moreover, we shed light on the benefit of this approach compared to the "double-loop" one. Secondly, the methodology is applied on a nuclear case study simulating a severe nuclear reactor accident.

4.1 Analytical example

To assess the efficiency of the "single-loop" methodology, we consider the analytical model $\mathcal{M}$ defined in Equation (17). The inputs X_1, X_2 and X_3 are assumed to be independent. Moreover, their probability distributions $\mathbb{P}_{X_1}$, $\mathbb{P}_{X_2}$ and $\mathbb{P}_{X_3}$ can equiprobably be $\mathbb{P}_U$, $\mathbb{P}_T$ or $\mathbb{P}_N$, 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, and $\mathbb{P}_N$ is the truncated normal distribution on $[0, 1]$ with mean 0.6 and standard deviation 0.2.

In practice, this configuration may occurs when for example three experts agree on the input variation ranges but, have different opinions on the nature of the probability distribution. More precisely:

- the first expert claims that except the range of variation, no other information can be assumed on the uncertain variable;
- the second adds that the most likely value is 0.4;
- the third thinks that the mean and the standard deviation can respectively be assumed equal to 0.6 and 0.2.

According to the principle of maximum entropy for expert elicitation [30, 32], the information provided by these experts are respectively modeled by the distributions $\mathbb{P}_U$, $\mathbb{P}_T$ and $\mathbb{P}_N$. By assigning equal importance to these three opinions, the uniform distribution on the set $\{\mathbb{P}_U, \mathbb{P}_T, \mathbb{P}_N\}$ seems to be here the most reasonable choice for the second-level uncertainty.

As a first step, we well approximate the theoretical values of 2^{nd}-level GSA indices for the model $\mathcal{M}$. Subsequently, to study the convergence rates of "single-loop" estimators, we apply Algorithm 1 for different sample sizes. To define 1^{st} and 2^{nd}-level HSIC measures, we use Standardized Gaussian kernel (see Remark 2) for all vector quantities and the kernels presented in Section 3.3 for non-vectorial quantities of interest.

4.1.1 Computation of theorical values

We focus here on Option 2a of Algorithm 1, the other quantities of interest are studied in Section 4.1.4. To approximate the theoretical values of 2^{nd}-level HSIC measures and indices, we consider the set of the $n_1 = 27$ possible 3-tuples of input probability distributions. For each input distribution, the 1^{st}-level HSIC measures and indices are computed using a sample of size $n_2 = 1000$, generated according to the prior input density. The theoretical 2^{nd}-level HSIC measures are computed:

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

The theoretical 2^{nd}-level HSIC indices are also computed:

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

In this example, we observe that $\mathrm{R}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_1}, \mathcal{R})$ is significantly larger than the other two indices, while $\mathrm{R}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_3}, \mathcal{R})$ is negligible. Based on these results, the lack of knowledge on $\mathbb{P}_{X_3}$ is not

responsible for the variability of 1st-level HSIC indices. This distribution can simply be set to a reference one. Furthermore, the impact of $\mathbb{P}_{X_1}$ uncertainty is by far the largest and the one of $\mathbb{P}_{X_2}$ remains non-negligible. Therefore, characterization efforts should 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 sequel, 2nd-level HSIC estimators using the mixture distribution, the Wasserstein barycenter and the Symmetrical Kullback-Leibler barycenters are respectively denoted $\widetilde{\text{HSIC}}_M(\mathbb{P}_{X_k}, \mathcal{R})$, $\widetilde{\text{HSIC}}_W(\mathbb{P}_{X_k}, \mathcal{R})$ and $\widetilde{\text{HSIC}}_K(\mathbb{P}_{X_k}, \mathcal{R})$. Similarly, the 2nd-level indices are denoted $\widetilde{\text{R}}^2_{\text{HSIC},M}(\mathbb{P}_{X_k}, \mathcal{R})$, $\widetilde{\text{R}}^2_{\text{HSIC},W}(\mathbb{P}_{X_k}, \mathcal{R})$ and $\widetilde{\text{R}}^2_{\text{HSIC},K}(\mathbb{P}_{X_k}, \mathcal{R})$.

To study the convergence rate of the "single-loop" estimators, we apply Algorithm 1 from samples with sizes ranging from $n_2 = 100$ to $n_2 = 1500$. For each sample size, the estimations are repeated independently 200 times using independent samples. Results are given by Figure 2, where the theoretical values of $\text{R}^2_{\text{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R})$ are represented in dotted lines. Visually, the estimators based on the mixture distribution and the Symmetrical Kullback-Leibler barycenter seem to perform similarly both for small and large sample sizes. In particular, the dispersion of these estimators are satisfying from $n_2 = 700$. In contrast, the Wasserstein barycenter estimators are less accurate (higher dispersion) compared to the previous estimators, especially for small and medium size samples (i.e. n_2 in $[300, 700]$).

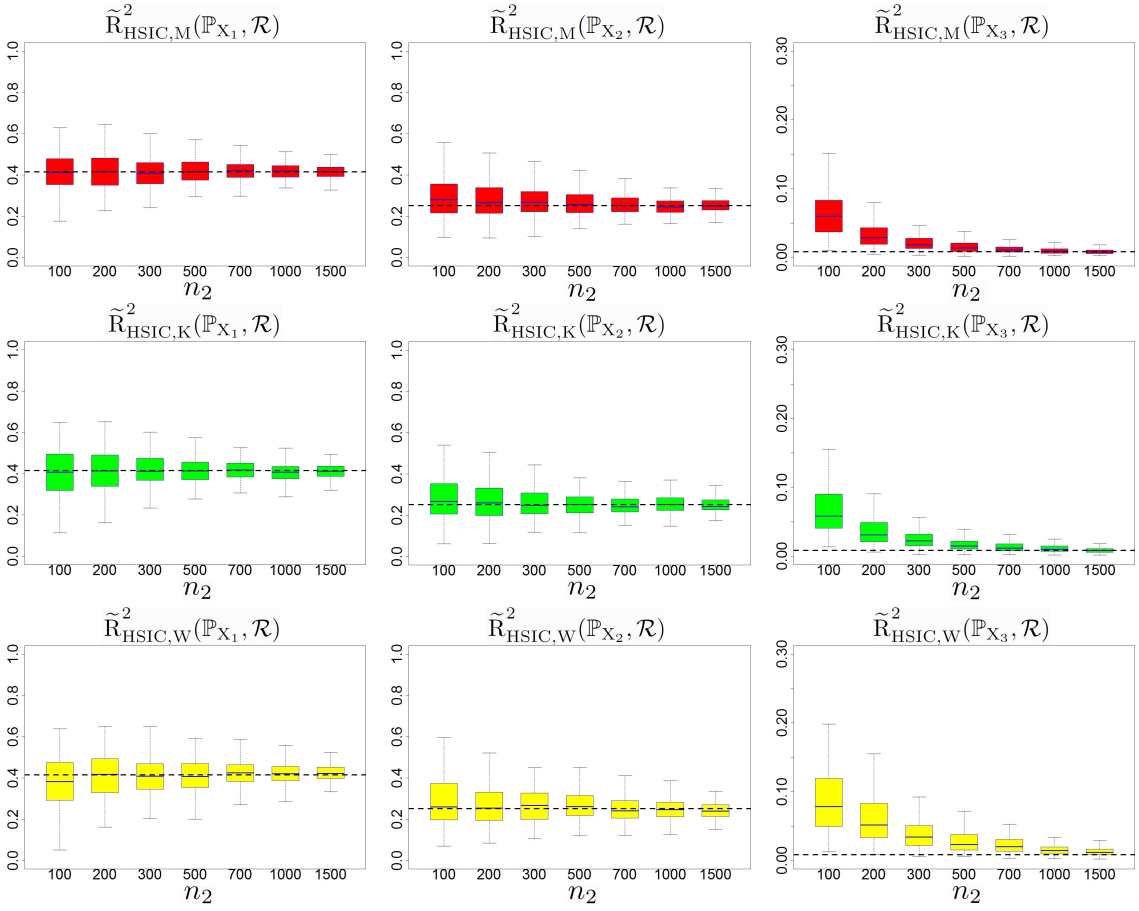

Figure 2: Model $\mathcal{M}$ – Convergence plots of the estimators $\widetilde{\text{R}}^2_{\text{HSIC},M}(\mathbb{P}_{X_k}, \mathcal{R})$, $\widetilde{\text{R}}^2_{\text{HSIC},W}(\mathbb{P}_{X_k}, \mathcal{R})$ and $\widetilde{\text{R}}^2_{\text{HSIC},K}(\mathbb{P}_{X_k}, \mathcal{R})$, with respect to the sample size n_2. Theoretical values are represented in dotted lines.

A more pragmatic way to compare these drawing densities, is to compare the estimators ability to correctly rank the input distributions based on their influence level. To do so, we compute for each sample size, the percentage of times the theoretical ranking and the one given by the

estimators match. The results are presented in Table 2. As expected, the estimators based on the mixture distribution and the Symmetrical Kullback-Leibler barycenter outperform those based on the Wasserstein barycenter, and this regardless of the sample size. This can be explained by the fact that the likelihood ratio f/f_W is very high in the neighborhoods of 0 and 1. Furthermore, the Kullback-Leibler barycenter seems to give slightly better results for small samples $n_2 \leq 300$, the reverse is true from $n_2 = 500$.

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

Table 2: Model $\mathcal{M}$ – Good ranking rates of $(\mathbb{P}_{X_1}, \mathbb{P}_{X_2}, \mathbb{P}_{X_3})$ using the estimators $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},M}(\mathbb{P}_{X_k}, \mathcal{R})$, $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},K}(\mathbb{P}_{X_k}, \mathcal{R})$ and $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},W}(\mathbb{P}_{X_k}, \mathcal{R})$, with respect to the sample size n_2.

4.1.3 Comparison with Monte Carlo "double-loop" approach

We compare now the performance of the "single-loop" and "double-loop" approaches, in terms of convergence rates of estimators. To do so, we consider a total simulation budget of $n = 1026$ for both approaches. More precisely for the "double-loop" approach, a sample of size $n_2 = 38$ is generated for each possible 3-tuple of input distributions (for a total number of $n = n_1 \times n_2 = 1026$ simulations). The associated estimators are denoted $\widehat{\mathrm{R}}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R})$ with $k \in \{1, 2, 3\}$. Concerning the "single-loop" approach, we apply Algorithm 1 with $n_2 = 1026$ and we compute the estimators $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},M}(\mathbb{P}_{X_k}, \mathcal{R})$ and $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},K}(\mathbb{P}_{X_k}, \mathcal{R})$ with $k \in \{1, 2, 3\}$.

Each estimation is repeated 200 times with independent Monte Carlo samples. The estimator boxplots are shown by Figure 3, where the theoretical values are represented in dotted lines. The "double-loop" estimators show much more variability than the "single-loop" ones. Also, notice that the "single-loop" estimators are much less biased than the "double-loop" ones. Our approach significantly outperforms the "double loop". This conclusion is also supported by the *good ranking rates* presented in Table 3. A reasonable explanation for the benefit of the "single-loop" approach, may be the simulation budget for 1st-level HSIC. Indeed, given a total budget of n simulations, each 1st-level HSIC is computed using $n_2 = n$ for the "single-loop" approach, against $n_2 = n/n_1$ for the "double loop" approach. Although the prior estimators converge faster than the alternative ones, the total simulation number is drastically reduced when using the "double-loop" approach.

For this same model $\mathcal{M}$, other numerical studies with different hypothesis on input distribution uncertainty have been performed and yield similar results and conclusions.

Double loop	Single loop	
$\widehat{\mathrm{R}}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R})$	$\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},M}(\mathbb{P}_{X_k}, \mathcal{R})$	$\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},K}(\mathbb{P}_{X_k}, \mathcal{R})$
67.5%	100%	99%

Table 3: Model $\mathcal{M}$ – Comparison of good ranking rates of "single-loop" and "double-loop" estimators, for $n = 1026$.

4.1.4 GSA2 with other quantities of interest

It is fair to wonder whether GSA2 conclusions vary if we decide to choose other quantities of interest. To answer that, we deal with Options 2b, 2c and 2d of Algorithm 1. In all cases, we keep the same kernel choices as described at the beginning of Section 4.1. Let us examine these possibilities one-by-one.

• **Ranking by $\mathrm{R}^2_{\mathrm{HSIC}}$.** Before looking closely at the simulation results, one can notice that the convergence of 1st-level $\mathrm{R}^2_{\mathrm{HSIC}}$ estimators systematically implies the convergence of those by ranking. Therefore, the estimators of GSA2 indices of Option 2b converge faster than those of Option 2a.

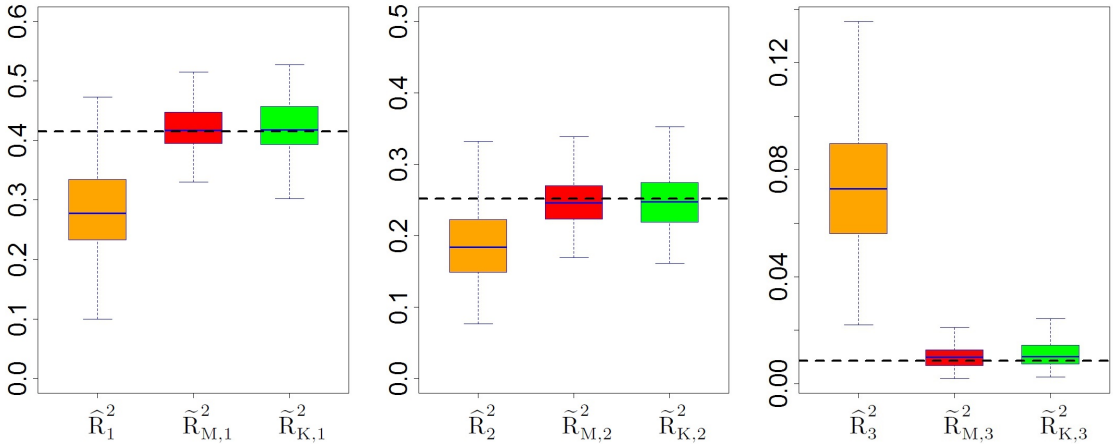

Figure 3: Model $\mathcal{M}$ – Comparison of 2^{nd}-level HSIC indices estimated by the "single-loop" and "double-loop" approaches for $n = 1026$. The estimators are denoted, $\widehat{\mathrm{R}}^2_k$ for $\widehat{\mathrm{R}}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_k}, \mathcal{R})$, $\widetilde{\mathrm{R}}^2_{M,k}$ for $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},M}(\mathbb{P}_{X_k}, \mathcal{R})$ and $\widetilde{\mathrm{R}}^2_{K,k}$ for $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},K}(\mathbb{P}_{X_k}, \mathcal{R})$. Theoretical values are represented in dotted lines.

Moreover, according to Section 4.1.2 results, a sample of size $n_2 = 1000$ of the drawing density is sufficient to accurately estimate the indices. We thus obtain:

$$\mathrm{R}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_1}, \mathcal{R}) = 0.3830,\ \mathrm{R}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_2}, \mathcal{R}) = 0.0958 \text{ and } \mathrm{R}^2_{\mathrm{HSIC}}(\mathbb{P}_{X_3}, \mathcal{R}) \simeq 0.$$

The gaps of theses values are more meaningful compared to those presented in Section 4.1.1. This is likely related to the stability of the ranking compared to GSA1 indices. Indeed, only significant variations of GSA1 indices contribute to GSA2 indices using the ranking. We safely conclude that $\mathbb{P}_{X_1}$ is the main contributor for the ranking uncertainty; less characterization efforts are required.

• **P-values vector.** When considering Options 2c or 2d as the quantity of interest, two points are highlighted. Firstly, the estimators of GSA2 indices show a large variance, regardless of the p-value estimation method (Gamma approximation or permutations), even for very large n_2 such as $n_2 = 5000$. In addition, the three estimated GSA2 indices are small (not exceeding 0.2). To help understanding these results, we focus on the estimated p-values for each possible input distribution. To do so, we use the permutation method with $B = 1000$ resamplings. The results show that the p-values associated to X_1 and X_2 are almost equal to zero (exactly zero numerically), regardless of the input distribution. Moreover, the p-values associated to X_3 are very low and in most cases below 10^{-5}. Therefore, the high variance of GSA2 indices is due to the difficulty of accurately estimating each p-value. In this case, Options 2c and 2d are not relevant: the independence hypothesis is not reliable and this, regardless of the input distribution. The 2^{nd}-level input uncertainties have no impact on these GSA1 results.

4.2 Nuclear safety application

Within the framework of 4^{th}-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 to model severe accident scenarios and assess the safety. Among them, a numerical tool called MACARENa (French: *Modélisation de l'ACcident d'Arrêt des pompes d'un REacteur refroidi au sodium*) developed by [13] 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 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. These studies show that only 3 input parameters mainly impact the accident transient predicted by MACARENa, namely:

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

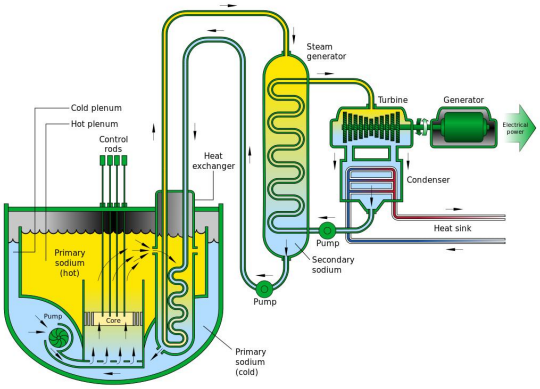

Figure 4: MACARENa application – Basic architecture of a Sodium-cooled Fast Reactor.

However, due to lack of data and knowledge, uncertainty remains on the distributions $\mathbb{P}_{X_1}$, $\mathbb{P}_{X_2}$ and $\mathbb{P}_{X_3}$. To take into account this uncertainty, the nature of each input distribution is assumed to be known, but with one uncertain parameter, as described in Table (4). The notations $\mathcal{N}_t(a, b, m, \sigma)$, $\mathrm{T}(a, b, c)$ and $\mathcal{U}(a, b)$ are respectively, the truncated normal distribution of mean m and standard deviation σ on $[a, b]$, the triangular law on $[a, b]$ with mode c and the uniform distribution on $[a, b]$. The identification of these uncertainties is based on expert advice. More specifically, the uncertainty on σ stems from a prior knowledge (no available data), while the uncertainties on c and m are due to their estimation using few existing partial data.

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

Table 4: MACARENa application – Uncertainties on $\mathbb{P}_{X_1}$,$\mathbb{P}_{X_2}$ and $\mathbb{P}_{X_3}$ distributions.

Among the outputs computed by MACARENa simulator to describe the ULOF accident, we focus on the first instant of sodium boiling denoted Y. To assess the impact of input distributions on GSA1 results of Y, we apply Algorithm 1 with Option 2a. We use the mixture density for the unique drawing and the same kernel choices as in Section 4.1. Moreover, we consider a Monte Carlo sample of size $n_2 = 1000$ for the unique drawing. This choice is motivated by two main reasons. Firstly, the MACARENa simulation cost (between 2 and 3 hours on average) which limits the total number of simulations. Secondly, the analytical three-dimensional example of Section 4.1 for which a budget of 1000 simulations give good results. In addition, we consider a Monte Carlo sample of $n_1 = 200$ 3-tuples of input distribution. These two choices for n_1 and n_2 will numerically be justified later, by studying the stability of estimators. Algorithm 1 gives the following GSA2 indices:

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

Consequently, the uncertainty on $\mathbb{P}_{X_1}$ mainly impacts GSA1 results, followed by $\mathbb{P}_{X_2}$, while the impact of $\mathbb{P}_{X_3}$ is negligible. To improve the robustness of GSA1 results, characterization efforts should then focus primarily on $\mathbb{P}_{X_1}$. A deeper analysis of the 200 results of GSA1 shows that the input X_2 is always the most predominant. Surprisingly, X_2 whose distribution is not the most influential on GSA1 results is the most influential on Y. This example illustrates, if necessary, that the information captured by GSA2 is different but complementary to that of GSA1.

To assess the accuracy of the estimation of GSA2 indices, we use a non-asymptotic bootstrapping approach [14]. For this, we first generate Monte Carlo subsamples with replacement from the initial sample (of 1000 simulations), then we re-estimate 2^{nd}-level $\mathrm{R}^2_{\mathrm{HSIC}}$ using these samples. More specifically, we consider 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 [26] 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 values estimated from the initial sample and the ones 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, (i.e. $n_2 \leq 400$) and small dispersions for medium and big sizes (i.e. $n_2 \geq 500$). Therefore, we conclude that the estimations of GSA2 indices with the sample of $n_2 = 1000$ simulations are consistent, the stability of estimations being satisfactory from $n_2 = 700$.

We also check the estimation consistency in terms of input distributions ranking. Table 5 gives for each subsample size, the rate of times that the ranking matches the one obtained with the initial sample. The results confirm the conclusions drawn from the stability plots.

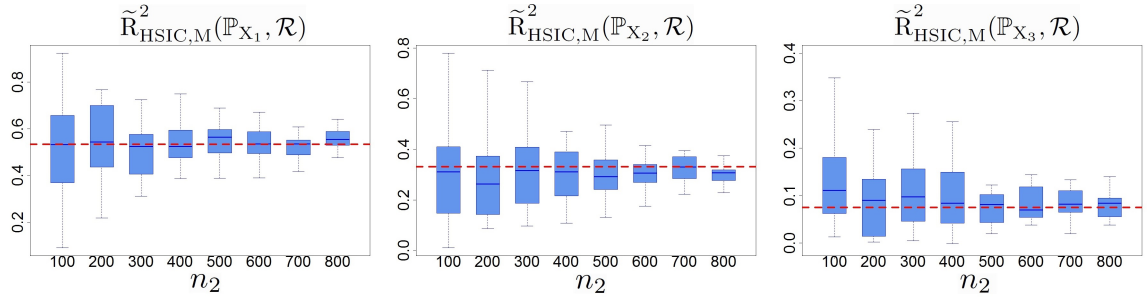

Figure 5: MACARENa application – Stability plots of the estimators $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},M}(\mathbb{P}_{X_k}, \mathcal{R})$, with respect to the sample size n_2. Reference values are represented in red dashed lines.

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

Table 5: MACARENa application – Good ranking rates of estimators $\widetilde{\mathrm{R}}^2_{\mathrm{HSIC},M}(\mathbb{P}_{X_k}, \mathcal{R})$, with respect to the size n_2 of the unique sample.

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, 2^{nd}-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 2^{nd}-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 with respect to the Symmetrical Kullback-Leibler distance or Wasserstein distance. The estimation of 2^{nd}-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 [33, 7, 48] 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 2^{nd}-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 [36, 3]. 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 2^{nd}-level of uncertainty.

An other potential prospect could be to make the connection between our approach and Perturbed-Law based Indices (PLI) [28, 39]. 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 2^{nd}-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] Mélisande Albert, Béatrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on hsic measures. *The Annals of Statistics*, 50(2):858–879, 2022.

[3] 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.

[4] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[5] Charles R Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

[6] Vincent Chabridon, Mathieu Balesdent, Jean-Marc Bourinet, Jérôme Morio, and Nicolas Gayton. Reliability-based sensitivity estimators of rare event probability in the presence of distribution parameter uncertainty. *Reliability Engineering & System Safety*, 178:164–178, 2018.

[7] Thomas M. Cioppa. *Efficient nearly orthogonal and space-filling experimental designs for high-dimensional complex models*. PhD thesis, Naval Postgraduate School, Monterey, California, United States, 2002.

[8] Imre Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4):191–213, 1972.

[9] Sebastien Da Veiga. Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, 85(7):1283–1305, 2015.

[10] 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.

[11] 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.

[12] Etienne de Rocquigny, Nicolas Devictor, and Stefano Tarantola. *Uncertainty in industrial practice: a guide to quantitative uncertainty management*. John Wiley & Sons, 2008.

[13] 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.

[14] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, CRC Press, 1993.

[15] Mohamed Reda El Amri and Amandine Marrel. Optimized hsic-based tests for sensitivity analysis: Application to thermalhydraulic simulation of accidental scenario on nuclear reactor. *Quality and Reliability Engineering International*, 38(3):1386–1403, 2022.

[16] B. Everitt and D.J. Hand. *Finite Mixture Distributions*. Springer Netherlands, 1981.

[17] Clark R Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

[18] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schoelkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*. MIT Press, 2008.

[19] 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.

[20] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, pages 63–77. Springer Berlin Heidelberg, 2005.

[21] John Michael Hammersley and David Christopher Handscomb. *Monte Carlo Methods*. Springer, 1964.

[22] 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):1175–1209, 2006.

[23] Bertrand Iooss and Paul Lemaître. *A Review on Global Sensitivity Analysis Methods*, pages 101–122. Springer, 2015.

[24] T. Ishigami and T. Homma. An importance quantification technique in uncertainty analysis for computer models. In *International Symposium on Uncertainty Modeling and Analysis*, pages 398–403. IEEE Computer Society, 1990.

[25] 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.

[26] V Roshan Joseph, Evren Gul, and Shan Ba. Maximum projection designs for computer experiments. *Biometrika*, 102(2):371–380, 2015.

[27] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[28] 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.

[29] Horia Mania, Aaditya Ramdas, Martin J Wainwright, Michael I Jordan, and Benjamin Recht. On kernel methods for covariates that are rankings. *Electronic Journal of Statistics*, 12(2):2537–2577, 2018.

[30] Mary A Meyer and Jane M Booker. *Eliciting and analyzing expert judgment: a practical guide*. Society for Industrial and Applied Mathematics, 2001.

[31] Anouar Meynaoui. *New developments around dependence measures for sensitivity analysis: application to severe accident studies for generation IV reactors*. PhD thesis, INSA Toulouse, Toulouse, France, 2019.

[32] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

[33] Luc Pronzato and Werner G Müller. Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22(3):681–701, 2012.

[34] 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.

[35] Robert J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 1980.

[36] Philippe Smets. What is dempster-shafer's model. *Advances in the Dempster-Shafer theory of evidence*, pages 5–34, 1994.

[37] I.M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1(4):407–414, 1993.

[38] B. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckriet, and B. Schoelkopf. Kernel choice and classifiability for rkhs embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22*. Curran Associates Inc., 2009.

[39] Roman Sueur, Bertrand Iooss, and Thibault Delage. Sensitivity analysis using perturbed-law based indices for quantiles and application to an industrial case. 10th International Conference on Mathematical Methods in Reliability (MMR 2017), 2017.

[40] Masashi Sugiyama and Taiji Suzuki. Least-squares independence test. *IEICE Transactions on Information and Systems*, E94.D(6):1333–1336, 2011.

[41] Masashi Sugiyama and Makoto Yamada. On kernel parameter selection in hilbert-schmidt independence criterion. *IEICE Transactions on Information and Systems*, E95.D(10):2564–2567, 2012.

[42] Zoltán Szabó and Bharath K Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

[43] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[44] D. Michael Titterington, Adrian F. M. Smith, and Udi E. Makov. *Statistical analysis of finite mixture distributions*. Wiley, 1985.

[45] R. v. Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947.

[46] Raymond Veldhuis. The centroid of the symmetrical kullback-leibler distance. *IEEE Signal Processing Letters*, 9(3):96–99, 2002.

[47] Cédric Villani. *Topics in optimal transportation*. American Mathematical Society, 2003.

[48] 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.

[49] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Uncertainty in Artificial Intelligence*, pages 804–813. AUAI Press, 2011.

A Proof of Proposition 1

We prove here that

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

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

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

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

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

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

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

Finally,

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

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

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

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