ESAIM: PS 30 (2026) 285–341
https://doi.org/10.1051/ps/2026004

ESAIM: Probability and Statistics
www.esaim-ps.org

KERNEL-BASED PARAMETER SCREENING FOR CONDITIONAL BAYESIAN CALIBRATION OF CHAINED NUMERICAL MODELS: APPLICATION TO FUEL PERFORMANCE SIMULATION OF PRESSURIZED WATER REACTORS

Oumar Baldé[1,*], Gabriel Sarazin[2], Amandine Marrel[1,3], Guillaume Damblin[2] and Antoine Bouloré[4]

Abstract. Numerical simulation is widely used in many fields of engineering to study complex physical systems. The numerical models, designed to faithfully represent the underlying physical phenomena, are subject to uncertainties of different natures (either numerical, stochastic or epistemic) that degrade the accuracy of the simulated outputs. Part of epistemic uncertainty arises from limited knowledge regarding some input model parameters $\boldsymbol{\theta}$. This component can be reduced through Bayesian calibration of the model against experimental data. Before calibration itself, sensitivity analysis can be used to better understand how parameter uncertainties impact the model output, and this may help confine calibration to the most impactful parameters. In this work, we show that kernel methods, especially those based on the Hilbert–Schmidt independence criterion (HSIC), are effective tools in support of Bayesian calibration, both for a single model and for two chained models. In the latter case, our main contribution is a screening methodology for the parameters $\boldsymbol{\theta}$ of the downstream model, which accounts for the posterior distribution of the upstream model parameters λ. By taking the expectation of the HSIC over λ, we define a new sensitivity measure that is able to incorporate the residual uncertainty due to the upstream model calibration. We show that the resulting sensitivity indices can be estimated from the same data used for conditional Bayesian calibration. We further demonstrate that the corresponding estimators are consistent and achieve convergence rates comparable to those of classical Monte Carlo estimators. Importantly, we construct two test procedures that enable rigorous decisions on which parameters among $\boldsymbol{\theta}$ should be selected. Finally, we apply the proposed approach to nuclear fuel simulation to screen the calibration parameters $\boldsymbol{\theta}$ of a fission gas behavior model which follows an upstream thermal model whose thermal conductivity λ was calibrated in previous work.

Mathematics Subject Classification. 62H20, 62F15, 62G05, 62G10, 62F12, 62P30.

Received September 24, 2025. Accepted February 26, 2026.

Keywords and phrases: Bayesian calibration, global sensitivity analysis, chained numerical models, dependence measures, Hilbert–Schmidt independence criterion, independence testing.

[1] CEA, DES, IRESNE, DER, SESI, Cadarache, 13108 Saint-Paul-Lez-Durance, France.
[2] Université Paris-Saclay, CEA, DES, ISAS, DM2S, SGLS, 91191 Gif-sur-Yvette, France.
[3] Avignon Université, LMA UPR 2151, 84140 Avignon, France.
[4] CEA, DES, IRESNE, DEC, SESC, Cadarache, 13108 Saint-Paul-Lez-Durance, France.

* Corresponding author: oumar.balde@cea.fr

1. Introduction

Like the power generation industry, many high-impact industrial sectors rely on numerical simulations to investigate complex physical phenomena. The numerical models used to mimic the behavior of systems or facilities often involve a large number of control variables, such as experimental conditions or geometric design specifications [1]. Additionally, simulations often involve physical parameters that cannot be measured experimentally, thus carry epistemic uncertainty, and therefore require calibration [2, 3]. Mathematically, a numerical model is a function $y : (\boldsymbol{x}, \boldsymbol{\theta}) \in \mathcal{D}_{\boldsymbol{X}} \times \mathcal{D}_{\boldsymbol{\Theta}} \subseteq \mathbb{R}^d \times \mathbb{R}^p \to \mathcal{D}_Y \subseteq \mathbb{R}$ where $\boldsymbol{x}$ denotes the control variables and $\boldsymbol{\theta}$ the calibration parameters.

In the context of uncertainty quantification [4], model calibration is a crucial step to improve the accuracy and robustness of numerical simulations [2, 5]. The calibration step aims to set the optimal parameter value(s) for $\boldsymbol{\theta}$ so that the simulator y sticks as closely as possible to the underlying physical reality. To do so, some experimental data of the system must be available. There are basically two types of calibration: deterministic calibration and Bayesian calibration. The first one tries to find the optimal value(s) of $\boldsymbol{\theta}$ by minimizing a loss function (most often based on the least-squares criterion) which quantifies the discrepancy between the available experimental data and the corresponding simulation data [6, 7]. In contrast, the second approach treats uncertainty in the calibration parameters $\boldsymbol{\theta}$ within a fully probabilistic framework. Bayesian inference then leverages both experimental data and expert judgment to extract maximal information [8, 9].

Bayesian calibration improves the accuracy of numerical models by estimating the parameters $\boldsymbol{\theta}$ from prior knowledge, available experimental data $\boldsymbol{z} = \{\boldsymbol{z}_i\}_{i=1}^{n_{\text{exp}}}$ and a probabilistic model. The prior distribution $\pi(\boldsymbol{\theta})$ encodes prior knowledge about $\boldsymbol{\theta}$ typically obtained through expert elicitation or statistical considerations [10]. The experimental data $\boldsymbol{z}$, reflecting the physical system with inherent uncertainty, are then used to update this prior with Bayes' rule, which leads to the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$. In practice, the maximum *a posteriori* (MAP) estimate of $\boldsymbol{\theta}$ (*i.e.*, the mode of the posterior density) is most commonly used to run the calibrated model.

For complex models, the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$ is almost always intractable analytically, but it can be approximated via simulation. In particular, Markov Chain Monte Carlo (MCMC) methods, such as the Metropolis-Hastings (MH) algorithm [11], allow simulating samples from $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$ by constructing a Markov chain whose stationary distribution is exactly $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$. However, due to the burn-in period required to reach stationarity, MCMC typically demands thousands of model evaluations, which is prohibitively expensive in many applications. In practice, to mitigate this cost, the model y is replaced by an inexpensive metamodel[1] (also called a surrogate model) $\widehat{y}$ [12, 13]. When the number p of parameters to be calibrated increases, additional difficulties arise. The most significant of these is the slow convergence of the MH algorithm. Adaptive MH variants [14] overcome this limitation by iteratively updating the proposal distribution, which improves convergence. However, such adaptive MH algorithms are difficult to tune efficiently. For this reason, in many recent works dedicated to high-dimensional Bayesian calibration, a preliminary step of dimension reduction is completed before using the MH algorithm. Bayesian calibration is thus limited to a subset of parameters which are selected by means of a preliminary sensitivity analysis [15–18].

Sensitivity analysis (SA) may be used to rank or screen model inputs according to their influence on the output(s) [19, 20]. Unlike local methods that measure sensitivity in the neighborhood of a nominal point, global sensitivity analysis (GSA) endeavors to account for the entire joint distribution of all sources of uncertainty [21]. In the 40 years that SA has been around, very different approaches have emerged, such as those based on the decomposition of the output variance [22, 23], the output gradient [24], linear regression [25], cooperative game theory [26, 27] or kernel methods [28, 29]. For a given problem, the most suitable SA method depends on many factors, including the dimensionality of the input space (*i.e.*, the number of input variables), the simulation budget (*i.e.*, the maximum number of model evaluations), correlations between the inputs, the mathematical nature of the output object, and specific data-related constraints. In particular, the given-data context arises when GSA must be performed using a provided set of input-output data originally generated for another purpose (*e.g.*, building a metamodel). This typically corresponds to a small dataset obtained from a Monte

[1] Mathematical function $\widehat{y}$ that is trained on input-output simulation data to provide a sufficiently accurate approximation of y.

Carlo or space-filling design. Many GSA methods have been proposed to tackle these challenges. Among them, the Hilbert–Schmidt independence criterion (HSIC) approach [28, 30] handles these difficulties effectively and has become widely used. In this work, the main technical challenge stems from a specific engineering context, involving the chaining of multiple numerical models within a multiphysics application.

The ALCYONE fuel application is composed of interdependent physical models that imitate the mechanical, thermal and chemical behaviors of fuel rods in the core of pressurized water reactors [31]. Here, we focus on the chaining of two models: the thermal model (upstream model) and the fission gas behavior model (downstream model). The thermal model simulates the evolution of the temperature within the fuel rod during the fission reaction and returns the associated temperature field as output. This field is subject to uncertainty because it is computed from an uncertain quantity λ called the thermal conductivity. In previous work, the posterior distribution of λ was inferred from experimental data of thermal component [32]. Then, the temperature field produced by the thermal model is used by the fission gas behavior model to simulate the evolution of the fission products (fuel swelling and fission gas release). This model returns several scalar outputs, among which the released gas fraction (RGF) will be the only output of interest in this work. It also involves parameters $\boldsymbol{\theta}$ to be calibrated from experimental RGF data $\boldsymbol{z}$. The uncertainty in the RGF results from both the uncertainty in $\boldsymbol{\theta}$ and that in λ, the latter being propagated through the temperature field. Estimating the posterior distribution of $\boldsymbol{\theta}$ conditional on λ (which is known as conditional Bayesian calibration) can be difficult when the dimension of $\boldsymbol{\theta}$ is high (beyond a dozen parameters).

The goal of this paper is to develop a screening methodology to reduce the dimension of $\boldsymbol{\theta}$, while accounting for the additional uncertainty brought by λ. To this end, a dedicated sensitivity measure must be defined to handle the bi-level uncertainty framework. Moreover, given the computational cost of the two models, the associated sensitivity indices must be estimable from a reasonable number of evaluations of the chained model, or from the same data used for the conditional calibration phase. For these reasons, we turned to the HSIC approach, as it converges rapidly, often bypasses problem-specific constraints, and relies on independence testing to control the risk of selecting weakly influential inputs. We further tailored the HSIC methodology to our problem, and we demonstrated that the customization retains its key advantages, particularly rapid convergence and reliable decision-making.

The outline of the paper is as follows. Section 2 provides the unified background material necessary for this work. First, Section 2.1 recalls the main principles of Bayesian calibration and highlights why parameter screening is crucial in high-dimensional settings. Then, Section 2.2 discusses the specific challenges of parameter screening in the context of Bayesian calibration, describes the limitations of classical methods, and shows that the HSIC methdodology fulfills all requirements. In Section 3, to illustrate the previous section in the simplest possible context, the HSIC approach is applied to the fission gas behavior model of the ALCYONE application, with the conductivity λ fixed at its nominal value. Parameter screening is subsequently performed in three different ways, still using HSIC indices, but examining different variants of the output of interest. Section 4, where it is shown how to adapt the HSIC approach to the particular context of chained numerical models, is the heart of this work. An integrated version of the HSIC is notably introduced in Section 4.1 to incorporate the uncertainty coming from the upstream numerical model. It is demonstrated that the resulting sensitivity measure properly characterizes independence in this bi-level uncertainty framework, which underlines its usefulness from a screening perspective. Then, four different estimators are proposed and their statistical properties are deeply investigated. Since the values taken by such kernel-based indices are always difficult to interpret, the final decision (regarding which parameters to select) is only trustworthy if it is supported by independence tests. For this reason, two test procedures, which are much inspired from what is done for standard HSIC indices, are developed in Section 4.2. Using these procedures, one can detect statistical dependence between the output and a given downstream model parameter, even when it only appears for certain values of the upstream model parameters. In Section 5, the proposed methodology is implemented on the fission gas behavior model of the ALCYONE application. Unlike what is done in Section 3, the posterior distribution of the thermal conductivity is this time considered in the selection process. Section 6 concludes the paper.

2. Parameter screening for Bayesian calibration

2.1. Bayesian calibration for black-box numerical models

The physical system $r : \mathcal{D}_{\boldsymbol{X}} \to \mathcal{D}_Y$ is represented as a function that maps inputs $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}} \subseteq \mathbb{R}^d$ to an output $y = r(\boldsymbol{x}) \in \mathcal{D}_Y \subseteq \mathbb{R}$. The probabilistic equation linking the experimental data $\boldsymbol{z}$ to the model outputs $\{y_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\}_{i=1}^{n_{\text{exp}}}$ for a set of experimental input configurations $\{\boldsymbol{x}_i\}_{i=1}^{n_{\text{exp}}}$ is given by:

$$\forall\, 1 \leq i \leq n_{\text{exp}}, \quad z_i = y_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + b(\boldsymbol{x}_i) + \epsilon_i. \tag{2.1}$$

For a given input configuration $\boldsymbol{x}_i$, the term ϵ_i represents the experimental uncertainty, mainly due to measurement error. The vector $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^{n_{\text{exp}}}$ is typically assumed to follow a multivariate Gaussian distribution with covariance matrice $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. The function $b(\boldsymbol{x})$, referred to as the model discrepancy in [8], quantifies the systematic mismatch between $y_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $r(\boldsymbol{x})$ when the model is operated at the optimal but unknown calibration parameter value $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\Theta}} \subseteq \mathbb{R}^p$. The discrepancy $b(\boldsymbol{x})$ is generally modeled by a Gaussian process and its hyperparameters are estimated jointly with the calibration parameters. In practice, this component may be neglected and merged with experimental uncertainty [33]. We will make this assumption in this work. The probabilistic model then simplifies to:

$$\forall\, 1 \leq i \leq n_{\text{exp}}, \quad z_i = y_{\boldsymbol{\theta}}(\boldsymbol{x}_i) + \epsilon_i. \tag{2.2}$$

Then, the posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$ is computed from Bayes' formula:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{z}) \propto \mathcal{L}(\boldsymbol{z} \mid \boldsymbol{\theta})\, \pi(\boldsymbol{\theta}), \tag{2.3}$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution quantifying the uncertainty of the calibration parameters before collecting $\boldsymbol{z}$. The likelihood $\mathcal{L}(\boldsymbol{z} \mid \boldsymbol{\theta})$ is a measure of the agreement between the observed data $\boldsymbol{z}$ and the outputs of the model $y_{\boldsymbol{\theta}}(\boldsymbol{x})$ for a given value of the parameters $\boldsymbol{\theta}$.

The posterior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$ is often estimated using MCMC algorithms [34], particularly the MH algorithm [11]. Once its Markov chain has reached stationarity, the algorithm generates samples from $\pi(\boldsymbol{\theta} \mid \boldsymbol{z})$. Achieving stationarity in the Markov chain can demand thousands of model evaluations, which makes this approach impractical when the model is computationally expensive. To overcome this limitation, the model $y_\theta(\boldsymbol{x})$ is often replaced by a metamodel $\widehat{y}(\boldsymbol{x}, \boldsymbol{\theta})$ [12]. After being trained on simulation data, this metamodel provides predictions of the output $y_{\boldsymbol{\theta}}(\boldsymbol{x})$ for any input pair $(\boldsymbol{x}, \boldsymbol{\theta})$. Sometimes, a metamodel $\widehat{y}_i(\boldsymbol{\theta})$ is built for each function $y_i(\boldsymbol{\theta}) = y_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ obtained by restricting y to a specific experimental configuration $\boldsymbol{x}_i$. This metamodeling strategy is particularly relevant when the experimental configurations are significantly different from each other, such as for example in nuclear engineering [35]. However, a drawback of this approach is that it requires training a separate metamodel for each experimental configuration. For Bayesian calibration, Gaussian process (GP) regression is by far the most widely employed technique [33, 36], while polynomial chaos expansion (PCE) is also frequently advocated in the literature [15, 17].

Unfortunately, everything becomes much harder when the number of calibration parameters becomes large. This difficulty is encountered in many application fields, as illustrated in [15, 37, 38] where $p \in \{36, 48, 11\}$ respectively. Indeed, increasing the number of calibration parameters raises two majors issues:

- **Poor metamodel accuracy.** Building an accurate metamodel becomes increasingly difficult as the number of hyperparameters to optimize grows. For GP metamodels, recent strategies to address this problem include advanced screening techniques [39], multi-objective optimization [40], or active subspace construction [41], among many others. A comprehensive review dedicated to GP surrogate modeling in high-dimensional input spaces can be found in [42].
- **Poor MCMC mixing.** Working in a high-dimensional parameter space often leads to dysfunction of the MH algorithm. Suboptimal tuning of the proposal distribution can severely impair the mixing of the

Markov chain, resulting in very slow convergence to the stationary distribution [43]. This may create identifiability issues and yield uninformative posterior distributions with highly correlated marginals [44]. Adaptive MH variants have been proposed to mitigate this phenomenon [14].

Both problems are classic manifestations of the well-known curse of dimensionality. The simplest solution is to restrict calibration to the most influential parameters, while setting all other parameters to their nominal values. This reduces the dimensionality without losing significant information about the output distribution. GSA provides a rigorous framework for making such decisions.

2.2. HSIC-based screening for efficient Bayesian calibration

Whether used to facilitate surrogate modeling or to improve Markov chain mixing, GSA is a fairly standard technique in model calibration [37, 38, 45–47]. When performed as a preliminary step to calibration, GSA has some specific features.

- **GSA can be limited to screening.** There is no need to rank the inputs according to their influence on the output, since the goal is only dimension reduction. This is actually an advantage, as screening methods [20, 24, 28] are computationally cheaper than ranking methods [23, 48].
- **GSA must be performed from given data.** Most often, the design of experiments is built to facilitate the (forthcoming) construction of a metamodel. No specific computational budget is allocated to GSA, which must therefore work with the available simulation data. These data typically take the form of a training set of input-output observations. Here, they will be assumed to come from a Monte Carlo design, even though in practice they are more likely to originate from a space-filling design.
- **GSA may target different output quantities.** Depending on the objective, it is not always the same output object that is analyzed. Three common situations can be distinguished. Let $\boldsymbol{\Theta}$ denote the random vector modeling the prior uncertainty on the calibration parameters $\boldsymbol{\theta}$.
  - **Approach A.** The output of interest is the model output for one specific input configuration $\boldsymbol{x}_i$:

$$Y_i\left(\boldsymbol{\Theta}\right) = y\left(\boldsymbol{x}_i, \boldsymbol{\Theta}\right). \tag{2.4}$$

    For an exhaustive study over all configurations, the GSA procedure must be repeated n_{exp} times, resulting in n_{exp} sets of sensitivity indices (one set for each configuration).
  - **Approach B.** The output of interest is the full output vector:

$$\boldsymbol{Y}\left(\boldsymbol{\Theta}\right) = \left[Y_i\left(\boldsymbol{\Theta}\right)\right]_{1\leq i\leq n_{\text{exp}}}. \tag{2.5}$$

    Here, the output object is a random vector with n_{exp} components. This constitutes a genuine methodological constraint, as many GSA methods do not naturally handle vector outputs.
  - **Approach C.** The output of interest is the (generalized) least-squares criterion:

$$L(\boldsymbol{\Theta}) = \Big(\boldsymbol{z} - \boldsymbol{Y}\left(\boldsymbol{\Theta}\right)\Big)^t \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\Big(\boldsymbol{z} - \boldsymbol{Y}\left(\boldsymbol{\Theta}\right)\Big). \tag{2.6}$$

These three approaches provide complementary insights, as each has its own scope of application. Approaches A and B are particularly suitable when a metamodel needs to be constructed. Screening is carried out with the goal of selecting the parameters that offer the best predictive performance. By contrast, in scenarios where no metamodel is required, Approach C should be preferred because the analyzed output is more directly aligned with the final objective of Bayesian calibration. Indeed, under the assumption of Gaussian noise, minimizing the least-squares criterion is equivalent to maximizing the likelihood which is a central component of Bayesian inference. Consequently, it will be interesting to compare the results of Approaches A and B against those of Approach C.

Only few GSA methods are able to accommodate all of these constraints. Derivate-based techniques [20, 24] are not suitable, since they require specific experimental designs to estimate model derivatives. Linear methods [25] are compatible with a prescribed Monte Carlo design, but they fail to detect nonlinear effects and interactions, and therefore cannot be regarded as reliable screening alternatives.

The Sobol' approach [23], also known as ANOVA (ANalysis Of VAriance) deserves a more thorough discussion. For a given input variable, the total-order Sobol' index equals zero if and only if the model output does not functionally depend on that input [49]. This property makes the total-order Sobol' indices particularly well-suited for screening. Moreover, extensions to vector and functional outputs have been proposed [50, 51], thereby broadening their scope. However, their estimation is not straightforward in the given-data context. In fact, the most natural way of estimating Sobol' indices is to generate a pick-freeze design [52] and this often results in a prohibitive simulation budget.

From a given Monte Carlo sample, two strategies can be considered to estimate the total-order Sobol' indices. One option is to train, validate, and emulate a metamodel, for instance a GP model as in [53]. Of course, the accuracy of the resulting estimates strongly depends on the predictive performance of the metamodel, which can be challenging to achieve in high dimensions, as previously discussed. Other option is to employ the kernel-based estimator [54], obtained by plugging the Nadaraya–Watson regression estimator into the naive Sobol' estimator. Again, this approach is difficult to apply in high dimensions, this time because the assumptions made on the smoothing kernels become increasingly restrictive and often unrealistic. In the Bayesian calibration literature, parameter screening using Sobol' indices is typically restricted to applications where either the computational cost or the dimensionality of the parameter space is low.

Dependence measures form another important family of GSA methods [28, 55, 56]. For each input-output pair, dependence is measured by quantifying the discrepancy between the joint distribution (capturing the true dependence) and the product of the marginal distributions (which would arise under independence). This framework brings together sensitivity measures that may target fundamentally different aspects of the input-output dependence. Rank-based measures such as Spearman's ρ or Kendall's τ are easy to estimate but do not characterize independence, whereas divergence-based measures can detect any form of dependence but rely on density estimation and are poorly suited to multidimensional outputs.

These limitations are overcome by the Hilbert-Schmidt independence criterion (HSIC) [30], a kernel-based dependence measure that transports the distributions into a reproducing kernel Hilbert space (RKHS) and computes the squared norm between their image functions. The use of HSIC indices for screening purposes was popularized by [28]. The method was subsequently extended in [29, 57, 58] and has now become a widely adopted standard practice [59–61]. Several HSIC-based methods have recently been developed for a variety of contexts, including fuzzy systems, knowledge-based systems, and chemometrics [62–64]. However, to the best of our knowledge, HSIC indices have never been applied to parameter screening in Bayesian calibration. This is rather surprising given that they meet all the necessary requirements.

- Under some mild conditions on the kernels, the HSIC is equal to zero if and only if the two random objects under study are independent of each other. However, it is not straightforward to determine which inputs have low influence on the output because each HSIC index follows its own maximum mean discrepancy (MMD) metric, which depends on the two kernels and the two marginal distributions. Consequently, for each input-output pair, a test of independence must be run to decide whether the HSIC value is sufficiently close to zero for accepting input-output independence. The test returns a p-value that allows controlling Type-I error (*i.e.*, the probability of incorrectly selecting an input whose influence should be neglected).
- The HSIC can be accurately estimated from a small amount of input-output data.
- The HSIC is well-defined for any type[2] of output objects, provided that an appropriate kernel is available. Moreover, the estimation procedure remains unchanged.

Further details on the HSIC methodology are provided in Appendix C. The basic concepts are presented in Appendix C.1. This includes the definition via the MMD, the alternative expression in terms of kernel-based

[2]In the seminal paper [65], it is only assumed that the two random objects lie in (possibly different) separable spaces.

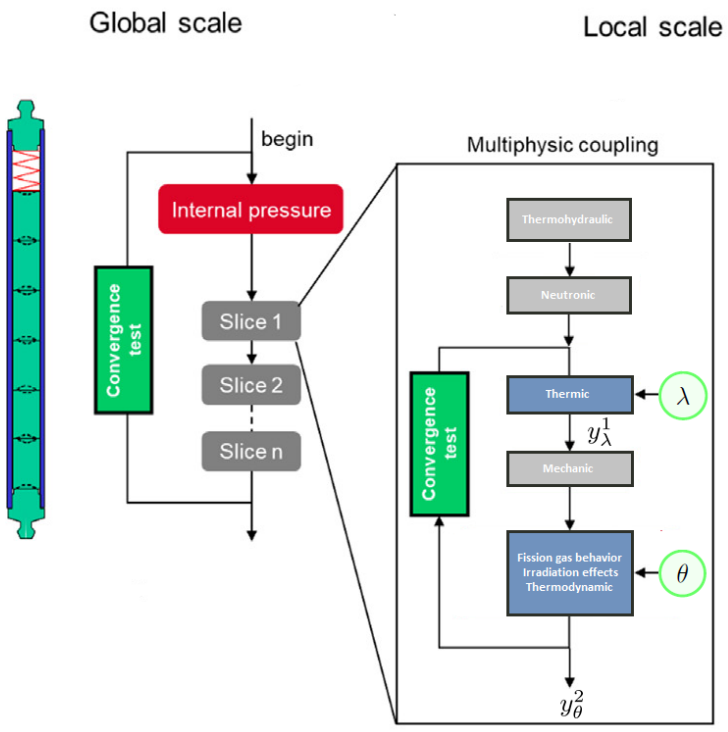

FIGURE 1. Simplified representation of the multiphysics coupling at work in the ALCYONE application [31]. The thermal model y^1_λ feeds into the fission gas behavior model $y^2_{\boldsymbol{\theta}}$. Both are depicted as blue boxes.

moments, and estimation using either a U- or V-statistic estimator. Then, HSIC-based independence testing is outlined in Appendix C.3, with emphasis on the three possible test procedures. Finally, Appendix C.5 discusses implementation aspects in the specific context of Bayesian calibration. In particular, guidelines are provided on how to choose kernels depending on the selected approach.

3. Application to the fission gas behavior model

In this section, we apply the HSIC methodology to a numerical model from the nuclear industry whose parameters need to be calibrated. More specifically, the goal is to perform parameter screening for the fission gas behavior model of the ALCYONE application [31].

3.1. Presentation of the ALCYONE application

ALCYONE is a multidimensional fuel performance simulation code developed at CEA within the PLEIADES software environment [66]. It is specifically designed to simulate the thermo-mechanical behavior of nuclear fuel rods under irradiation in pressurized water reactors (PWRs). A simplified diagram is shown in Figure 1. Although the application integrates numerical models from five physical disciplines, we focus here only on the chaining between two models (represented in the diagram as blue rectangular boxes): the thermal model (upstream model) and the fission gas behavior model (downstream model). The thermal model simulates how the temperature evolves inside the fuel rod during fission and outputs the corresponding temperature field throughout the rod. This temperature field feeds the fission gas behavior model which predicts the release and redistribution of fission gases within the fuel rod.

Fission gases produced during irradiation are distributed in three forms: dissolved in the fuel grains, trapped in intragranular bubbles, and trapped in intergranular bubbles. When the intergranular bubbles at the grain boundaries of uranium dioxide connect to form continuous pathways, a phenomenon known as fission gas release occurs into the free volume of the fuel rod. Under high burnup conditions, this process may induce fuel restructuring (finer grain structure, increased porosity), which strongly affects gas behavior. The fission gas behavior model simulates these phenomena and provides a scalar output of interest, the released gas fraction (RGF), which quantifies the proportion of gas released from the fuel.

The thermal model will be denoted as y^1_λ where $\lambda \in \mathcal{D}_\Lambda \subseteq \mathbb{R}$ is the thermal conductivity. Experimental data were used to calibrate this parameter in previous work [32]. The fission gas behavior model will be denoted by $y^2_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} \in \mathcal{D}_{\boldsymbol{\Theta}} \subseteq \mathbb{R}^p$ are the calibration parameters. Accurate calibration of these $p = 11$ parameters, while rigorously accounting for the posterior uncertainty on λ, is a critical step to guarantee the reliability of ALCYONE simulations.

For fixed λ, the model y^1_λ takes a control variable $\boldsymbol{x} \in \mathcal{D}_{\boldsymbol{X}}$ describing the fuel rod configuration (linear heat rate, burnup rate, geometry) and returns the temperature field throughout the rod. For fixed $\boldsymbol{\theta}$, the model $y^2_{\boldsymbol{\theta}}$ takes this temperature field as input and returns the RGF. The chaining between the two models can be summarized by the following deterministic function:

$$\begin{array}{rccc} y: & \mathcal{D}_{\boldsymbol{X}} \times \mathcal{D}_\Lambda \times \mathcal{D}_{\boldsymbol{\Theta}} & \longrightarrow & \mathcal{D}_Y \\ & (\boldsymbol{x}, \lambda, \boldsymbol{\theta}) & \longmapsto & y(\boldsymbol{x}, \lambda, \boldsymbol{\theta}) = \left[y^2_{\boldsymbol{\theta}} \circ y^1_\lambda\right](\boldsymbol{x}). \end{array} \tag{3.1}$$

At this step, the epistemic uncertainty affecting λ is voluntarily ignored. The conductivity is fixed at its nominal value $\lambda_{\text{nom}} = 1$. This reduces the chained model to:

$$\begin{array}{rccc} y_s: & \mathcal{D}_{\boldsymbol{X}} \times \mathcal{D}_{\boldsymbol{\Theta}} & \longrightarrow & \mathcal{D}_Y \\ & (\boldsymbol{x}, \boldsymbol{\theta}) & \longmapsto & y_s(\boldsymbol{x}, \boldsymbol{\theta}) = \left[y^2_{\boldsymbol{\theta}} \circ y^1_{\lambda_{\text{nom}}}\right](\boldsymbol{x}). \end{array} \tag{3.2}$$

The uncertainty on λ will be reintroduced in the study from Section 3.4 onwards. For now, the goal is to identify the components of $\boldsymbol{\theta}$ that have the greatest influence on the RGF $y_s(\boldsymbol{x}, \boldsymbol{\theta})$.

3.2. Strategies considered for parameter screening

In the existing literature, the HSIC approach is mostly used to reduce dimensionality before building a metamodel [39, 58, 67]. Here, the presence of control variables $\boldsymbol{x}$ encoding multiple experimental configurations introduces an additional layer of complexity, which motivates the use of Approaches A, B and C presented in Section 2.2.

As the calibration parameters $\boldsymbol{\theta}$ of the fission gas behavior model are uncertain, they are modeled by a random vector $\boldsymbol{\Theta}$ following a distribution $\mathbb{P}_{\boldsymbol{\Theta}}$, with support $\mathcal{D}_{\boldsymbol{\Theta}} \subseteq \mathbb{R}^p$ and density $\pi(\boldsymbol{\theta})$. The components Θ_j of $\boldsymbol{\Theta}$ are assumed to be independent. As parameter screening precedes model calibration, $\mathbb{P}_{\boldsymbol{\Theta}}$ corresponds to the prior distribution of $\boldsymbol{\theta}$, informed by expert knowledge.

One key advantage of the HSIC approach is that the related indices can be estimated using a Monte Carlo design of (relatively small) size n. In the parameter space, the input data can be written as:

$$D_{\boldsymbol{\Theta},n} := \left\{\boldsymbol{\Theta}^{(l)}\right\}_{1 \le l \le n_{\text{exp}}} \in \mathbb{R}^{n \times p} \quad \text{with} \quad \boldsymbol{\Theta}^{(l)} = \left[\boldsymbol{\Theta}^{(l)}_1, \ldots, \boldsymbol{\Theta}^{(l)}_p\right]^t \in \mathbb{R}^p. \tag{3.3}$$

The fission gas behavior model is then evaluated for all pairs $\left(\boldsymbol{x}_i, \boldsymbol{\Theta}^{(l)}\right)$ where $\boldsymbol{x}_i$ is the i-th experimental configuration and $\boldsymbol{\Theta}^{(l)}$ is the l-th parameter value. For a fixed configuration $\boldsymbol{x}_i$, the n-sample obtained by evaluating $y_s(\boldsymbol{x}_i, \cdot)$ on all parameter values $\boldsymbol{\Theta}^{(l)}$ yields:

$$D_{Y_i,n} := \left\{Y_i^{(l)}\right\}_{1 \le l \le n} \quad \text{with} \quad Y_i^{(l)} = y_s\left(\boldsymbol{x}_i, \boldsymbol{\Theta}^{(l)}\right). \tag{3.4}$$

Combining all such samples $D_{Y_i,n}$ forms the complete output dataset:

$$\begin{aligned} D_{\boldsymbol{Y},n} &:= \{D_{Y_i,n}\}_{1\le i\le n_{\text{exp}}} \\ &= \left\{\boldsymbol{Y}^{(l)}\right\}_{1\le l\le n} \quad \text{with} \quad \boldsymbol{Y}^{(l)} = \left[Y_i^{(l)}\right]_{1\le i\le n_{\text{exp}}} \end{aligned} \tag{3.5}$$

The total simulation budget amounts to $n_{\text{exp}} \times n$ evaluations of y_s. At this stage, all the data required to implement the three screening approaches are available. The next steps depend on the selected screening strategy.

- **Approach A.** GSA is performed on the RGF of each fuel rod, that is on the output of y_s for a given experimental configuration $\boldsymbol{x}_i$:

$$Y_i = Y_i(\boldsymbol{\Theta}) = y(\boldsymbol{x}_i, \lambda_{\text{nom}}, \boldsymbol{\Theta}) =: f_i^A(\boldsymbol{\Theta}) =: \widetilde{f}_i^A(\boldsymbol{\Theta}, \lambda_{\text{nom}}). \tag{3.6}$$

  The function f_i^A, which encapsulates the chained model y, is introduced to establish a direct link between the analyzed output and the parameters $\boldsymbol{\Theta}$. The function $\widetilde{f}_i^A$ is simply a rewriting of f_i^A that makes the role of λ explicit. To estimate HSIC indices, the following data are required:

$$D_{i,n}^A := \{D_{\boldsymbol{\Theta},n}; D_{Y_i,n}\} \ . \tag{3.7}$$

  The resulting HSIC estimates are denoted by $\widehat{H}_{ij}$ for a fixed i and $1 \le j \le p$. These quantities are then used in independence tests. For each parameter Θ_j, the null hypothesis $(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp Y_i$ is tested. For this, a test procedure (chosen depending on the sample size) is run and provides a p-value $\widehat{p}_{ij}$. The significance level (controlling Type-I error) is set to $\alpha = 5\%$. If $\widehat{p}_{ij} < \alpha$, the null hypothesis is rejected and Θ_j is considered influential on Y_i. Repeating the procedure across all rods of the fuel assembly produces n_{exp} sets of p-values, which may lead to as many distinct conclusions. To make a global decision, one may compute the detection rate of each parameter Θ_j, defined as the percentage of rods for which that parameter is detected as influential:

$$\rho_j := \sum_{i=1}^{n_{\text{exp}}} \mathbf{1}_{\{\widehat{p}_{ij}<\alpha\}} \ . \tag{3.8}$$

  A parameter Θ_j is then selected if its detection rate ρ_j exceeds a (user-chosen) threshold, for example $\beta = 80\%$. In other words, a parameter is considered influential on the fuel assembly if it is influential on at least a fraction β of the fuel rods.
- **Approach B.** GSA is performed on the vector of all RGFs, that is on the vector:

$$\boldsymbol{Y} = \boldsymbol{Y}(\boldsymbol{\Theta}) = [Y_i]_{1\le i\le n_{\text{exp}}} =: f^B(\boldsymbol{\Theta}) =: \widetilde{f}^B(\boldsymbol{\Theta}, \lambda_{\text{nom}}). \tag{3.9}$$

  To estimate HSIC indices, the following data are required:

$$D_n^B := \{D_{\boldsymbol{\Theta},n}; D_{\boldsymbol{Y},n}\} \ . \tag{3.10}$$

  The null hypothesis becomes $(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp \boldsymbol{Y}$. The HSIC methodology naturally extends to the case of a vector output, provided that a characteristic kernel on $\mathbb{R}^{n_{\text{exp}}}$ is used to handle $\boldsymbol{Y}$ [68, 69]. A first option is to take the multivariate generalization of standard kernels (such as Gaussian and Matérn kernels). A second option is to equip each component Y_i with a characteristic kernel on $\mathbb{R}$ (for instance a Gaussian kernel) and then take the tensor product of these kernels. A third option is to consider kernel functions specifically designed for high-dimensional data, such as PCA kernels [58]. We adopted the second option,

as tensor-product kernels are characteristic in most cases (see Appendix A.3) while being very flexible (with a separate bandwidth parameter for each component Y_i). A more detailed discussion of the pros and cons of the different kernels is provided in Appendix C.5. Once an appropriate kernel is chosen for $\boldsymbol{Y}$, the standard HSIC procedure can be applied (see Appendix C). This produces HSIC estimates $\widehat{H}_j$ and corresponding p-values $\widehat{p}_j$, from which a global decision can ultimately be made.

- **Approach C.** GSA is performed on the least-squares criterion (measuring the discrepancy between the observed and simulated RGFs):

$$L = L(\boldsymbol{\Theta}) = \left(\boldsymbol{z} - \boldsymbol{Y}\right)^t \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \left(\boldsymbol{z} - \boldsymbol{Y}\right) =: f^C(\boldsymbol{\Theta}) =: \widetilde{f}^C(\boldsymbol{\Theta}, \lambda_{\text{nom}}). \tag{3.11}$$

To estimate HSIC indices, the following data are required:

$$D_n^C := \{D_{\boldsymbol{\Theta},n}; D_{L,n}\} \quad \text{with} \quad D_{L,n} = \left\{L^{(l)}\right\}_{1 \le l \le n} \quad \text{and} \quad L^{(l)} = L\left(\boldsymbol{\Theta}^{(l)}\right). \tag{3.12}$$

Since L is scalar, applying the HSIC methodology is straightforward. This time, the null hypothesis is $(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp L$. As in Approach B, one obtains HSIC estimates $\widehat{H}_j$, and then p-values $\widehat{p}_j$ from which the influential parameters are selected.

Remark 3.1. All three approaches can be carried out without much extra computational effort. Indeed, once $D_{\boldsymbol{Y},n}$ is obtained, at the cost of $n_{\text{exp}} \times n$ evaluations of y_s, everything that follows may be viewed as inexpensive post-processing. In particular, the datasets built at equations (3.7), (3.10) and (3.12) are directly derived from $D_{\boldsymbol{\Theta},n}$ and $D_{\boldsymbol{Y},n}$.

Remark 3.2. Rank-based dependence measures also offer a robust and distribution-free alternative to kernel-based dependence measures. However, two important limitations motivated our decision not to use them. First, unlike the HSIC built on two characteristic kernels, rank correlations (such as Spearman's ρ or Kendall's τ) do not characterize independence, as they mainly capture monotonic forms of dependence. Second, extending rank-based measures to vectors or non-Euclidean objects is non-trivial, whereas this is immediate for the HSIC, provided that appropriate kernels are available. For these two reasons, the HSIC appears better suited for parameter screening, although rank-based measures could in principle be used as complementary diagnostics.

Remark 3.3. Due to the strong similarities between the three approaches, developments will be conducted within a unified framework where the output of interest is denoted as:

$$G_{\text{nom}} := f(\boldsymbol{\Theta}) = \widetilde{f}(\boldsymbol{\Theta}, \lambda_{\text{nom}}).$$

The superscripts A, B and C will be omitted, except when needed to avoid ambiguity.

3.3. Results

In this section, we apply the three proposed approaches to the fission gas behavior model y_s and compare their results. The experimental input configurations $\{\boldsymbol{x}_i\}_{i=1}^{n_{\text{exp}}}$ correspond to the specifications of the $n_{\text{exp}} = 40$ fuel rods within the fuel assembly.

Due to computational constraints, the size of the Monte Carlo design $D_{\boldsymbol{\Theta},n}$ is restricted to $n = 200$. Although this sample size is already large, it is not yet sufficient to use the asymptotic test [30]. Instead, the permutation-based test [29] and the non-asymptotic Gamma test [58] are employed. The significance level is set to $\alpha = 5\%$.

The results of all independence tests performed in Approach A are shown in Figure 2. The parameters Θ_j are plotted on the y-axis, the outputs Y_i on the x-axis, sorted by decreasing burnup rates. Green points indicate that $(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp Y_i$ was rejected (statistical dependence detected), whereas gray points indicate that (H_0) was accepted (no influence). We specifically distinguish results obtained with U-statistics from those obtained with V-statistics. Decisions are made from the p-values returned by the permutation test. The non-asymptotic Gamma test yields identical decisions.

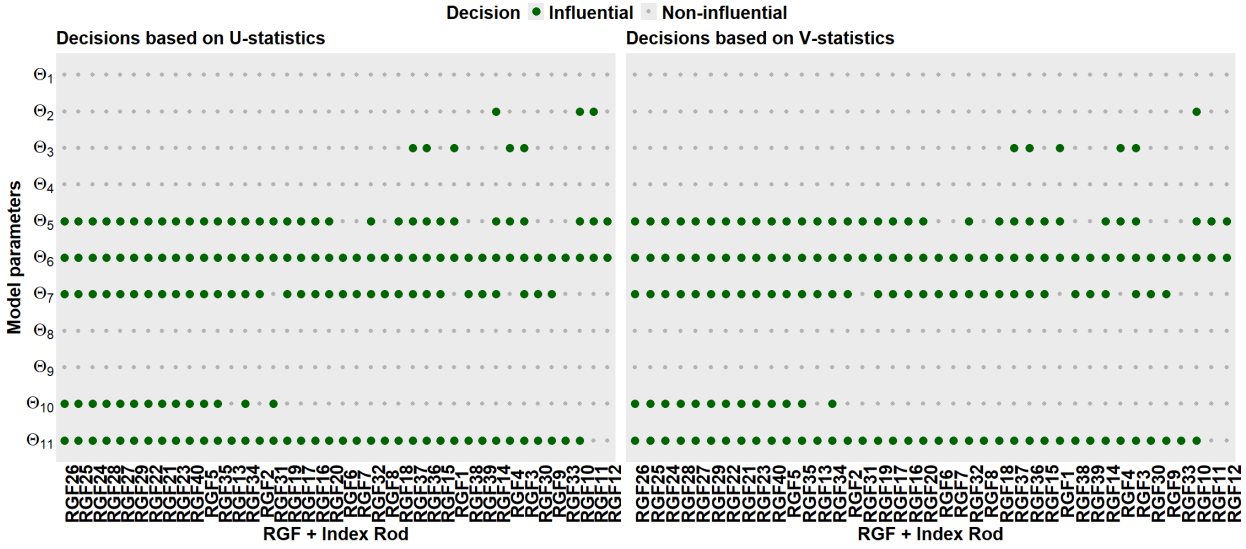

FIGURE 2. Results obtained for Approach A. For each fuel rod (characterized by the experimental configuration $\boldsymbol{x}_i$), the influence (weak or strong) of each parameter Θ_j is assessed using an HSIC test of independence. The test statistic is either the U-statistic estimator (left) or the V-statistic estimator (right). In both cases, a permutation-based test procedure is used. The significance level is set to $\alpha = 5\%$. The detection rates (percentages of rods where the parameters are detected as influential) are reported in Table 1.

TABLE 1. Comparison of the screening results obtained for the three considered approaches. Only the results associated with U-statistics are shown.

Results	Detection rates	p-values	
Parameters \ Approach	A	B	C
Θ_1	0%	0.972	0.538
Θ_2	7.5%	0.25	0.741
Θ_3	12.5%	0.372	0.127
Θ_4	0%	0.909	0.801
Θ_5	80%	0	0
Θ_6	100%	0	0.006
Θ_7	82.5%	0	0
Θ_8	0%	0.869	0.757
Θ_9	0%	0.856	0.729
Θ_{10}	35%	0.022	0.081
Θ_{11}	95%	0	0

The parameters Θ_5, Θ_6, Θ_7, and Θ_{11} are found to be influential on the RGF of all fuel rods. By contrast, the parameter Θ_{10} seems to be influential on a small number of RGFs, those corresponding to the highest burnup rates. The remaining parameters appear to have no influence on the RGFs, regardless of the fuel rod considered. To summarize these results into a single set of indices, the detection rates defined in equation (3.8)

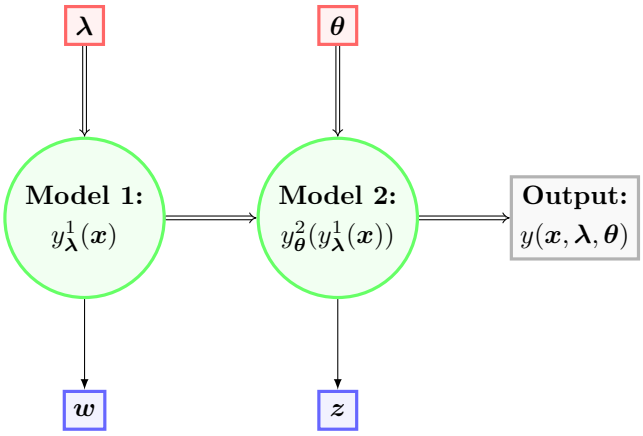

FIGURE 3. Diagram of the chaining between two numerical model: the upstream model $y^1_{\boldsymbol{\lambda}}$ and the downstream model $y^2_{\boldsymbol{\theta}}$. The calibration parameters are put above the models, and the available experimental data below. Double arrows indicate uncertainty propagation, while single arrows indicate data assimilation.

are computed and reported in Column A of Table 1. For the four most influential parameters, the detection rates exceed $\beta = 80\%$, whereas for Θ_{10}, it is only $\rho_{10} = 35\%$.

Table 1 also provides the p-values obtained for Approaches B and C. One can see that Approach B selects the five expected parameters, including Θ_{10}. Approach C selects the four predominant parameters, but rejects Θ_{10}. However, it should be noted that the related p-value is quite close to $\alpha = 5\%$, which confirms that Θ_{10} exerts only a minor influence. Overall, it is reassuring to see that the three screening strategies lead to the consistent conclusions. Furthermore, the results are largely insensitive to the HSIC estimator (U- vs. V-statistic) and the test procedure (permutations vs. Gamma approximation).

3.4. Accounting for the upstream model uncertainty

For a rigorous uncertainty treatment in the simulation chain, the thermal conductivity λ cannot be regarded as a constant parameter. Indeed, it is subject to an epistemic uncertainty resulting from the calibration of the thermal model. Naturally, this uncertainty must be incorporated into the study of the fission gas behavior model by adapting both the Bayesian calibration and GSA methodologies. This bi-level uncertainty framework is illustrated in Figure 3. For the sake of generality, the thermal and fission gas behavior models are labeled as Models 1 and 2. In addition, the calibration parameter(s) of Model 1 are denoted by $\boldsymbol{\lambda}$ to allow for a more general setting, where multiple parameters may be involved.

It is assumed that Model 1 has already been calibrated with some experimental data $\boldsymbol{w}$. For ALCYONE, this corresponds to what was done for the thermal model in [32]. To account for the posterior uncertainty on $\boldsymbol{\lambda}$, we introduce a random vector $\boldsymbol{\Lambda}$ with distribution $\mathbb{P}_{\boldsymbol{\Lambda}}$. The challenge in calibrating Model 2 is to make optimal use of the posterior knowledge on $\boldsymbol{\lambda}$, the prior knowledge on $\boldsymbol{\theta}$, and the experimental data $\boldsymbol{z}$. The main objective is to estimate the conditional density:

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}, \boldsymbol{z}) \quad \propto \quad \mathcal{L}(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})\, \pi(\boldsymbol{\theta}). \tag{3.13}$$

This inference task is called *conditional Bayesian calibration*.

Remark 3.4. Before conditional calibration, the random vectors $\boldsymbol{\Lambda}$ (posterior uncertainty on the parameters of Model 1) and $\boldsymbol{\Theta}$ (prior uncertainty on the parameters of Model 2) may reasonably be assumed independent.

After conditional calibration, the joint posterior density $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{z})$ generally does not factorize, meaning that $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ are no longer independent.

A conditional Bayesian calibration algorithm was developed in [70] to estimate the conditional posterior density for each value $\boldsymbol{\lambda}$ drawn from the posterior distribution $\mathbb{P}_{\boldsymbol{\Lambda}}$. This algorithm relies on experimental data $\boldsymbol{z}$ together with simulated data generated according to a nested Monte Carlo design. At the outer level, an m-sample of $\boldsymbol{\Lambda} \sim \mathbb{P}_{\boldsymbol{\Lambda}}$ is generated:

$$D_{\boldsymbol{\Lambda},m} := \left\{ \boldsymbol{\Lambda}^{(k)} \right\}_{1 \leq k \leq m} . \tag{3.14}$$

At the inner level, two sampling strategies can be considered.

(S1) An n-sample of $\boldsymbol{\Theta} \sim \mathbb{P}_{\boldsymbol{\Theta}}$ is generated for each value $\boldsymbol{\Lambda}^{(k)}$ in $D_{\boldsymbol{\Lambda},m}$. This amounts to introducing m independent copies $\left\{ \boldsymbol{\Theta}^{(k)} \right\}_{k=1}^{m}$ of $\boldsymbol{\Theta}$ and then generating an n-sample for each copy:

$$D_{\boldsymbol{\Theta}^{(k)},n} := \left\{ \boldsymbol{\Theta}^{(kl)} \right\}_{1 \leq l \leq n} . \tag{3.15}$$

The designs $D_{\boldsymbol{\Theta}^{(k)},n}$ are mutually independent. They are also independent of the design $D_{\boldsymbol{\Lambda},m}$. The chained model y is evaluated to compute all relevant model outputs:

$$D_{\boldsymbol{Y}^{(k)},n} := \left\{ \boldsymbol{Y}^{(kl)} \right\}_{1 \leq l \leq n} \quad \text{with} \quad \boldsymbol{Y}^{(kl)} = \left[y\left(\boldsymbol{x}_i, \boldsymbol{\Lambda}^{(k)}, \boldsymbol{\Theta}^{(kl)} \right) \right]_{1 \leq i \leq n_{\text{exp}}} . \tag{3.16}$$

$D_{\boldsymbol{Y}^{(k)},n}$ is an n-sample of the output vector $\boldsymbol{Y}^{(k)}$ obtained with $\boldsymbol{\Lambda}^{(k)}$:

$$\boldsymbol{Y}^{(k)} = \left[Y_i^{(k)} \right]_{1 \leq i \leq n_{\text{exp}}} \quad \text{with} \quad Y_i^{(k)} := y\left(\boldsymbol{x}_i, \boldsymbol{\Lambda}^{(k)}, \boldsymbol{\Theta}^{(k)} \right) . \tag{3.17}$$

For convenience, we also introduce notations for the final nested design:

$$D_{m,n}^{\text{S}_1} := \left\{ D_{\boldsymbol{\Theta}^{(k)}\boldsymbol{Y}^{(k)},n} \right\}_{1 \leq k \leq m} \quad \text{with} \quad D_{\boldsymbol{\Theta}^{(k)}\boldsymbol{Y}^{(k)},n} := \left\{ D_{\boldsymbol{\Theta}^{(k)},n}; D_{\boldsymbol{Y}^{(k)},n} \right\} . \tag{3.18}$$

(S2) A unique n-sample of $\boldsymbol{\Theta} \sim \mathbb{P}_{\boldsymbol{\Theta}}$ is generated and used for all values $\boldsymbol{\Lambda}^{(k)}$:

$$D_{\boldsymbol{\Theta},n} := \left\{ \boldsymbol{\Theta}^{(l)} \right\}_{1 \leq l \leq n} . \tag{3.19}$$

The designs $D_{\boldsymbol{\Lambda},m}$ and $D_{\boldsymbol{\Theta},n}$ remain independent. Evaluating the chained model y for $\boldsymbol{\Lambda}^{(k)}$ and all points in $D_{\boldsymbol{\Theta},n}$ produces the corresponding output data:

$$D_{\boldsymbol{Y}^{(k)},n} := \left\{ \boldsymbol{Y}^{(kl)} \right\}_{1 \leq l \leq n} \quad \text{with} \quad \boldsymbol{Y}^{(kl)} = \left[y\left(\boldsymbol{x}_i, \boldsymbol{\Lambda}^{(k)}, \boldsymbol{\Theta}^{(l)} \right) \right]_{1 \leq i \leq n_{\text{exp}}} . \tag{3.20}$$

The complete nested design is then given by:

$$D_{m,n}^{\text{S}_2} := \left\{ D_{\boldsymbol{\Theta}\boldsymbol{Y}^{(k)},n} \right\}_{1 \leq k \leq m} \quad \text{with} \quad D_{\boldsymbol{\Theta}\boldsymbol{Y}^{(k)},n} := \left\{ D_{\boldsymbol{\Theta},n}; D_{\boldsymbol{Y}^{(k)},n} \right\} . \tag{3.21}$$

Both strategies incur the same computational cost. More precisely, they require $N_1 = n_{\text{exp}} \times m$ evaluations of $y_{\boldsymbol{\lambda}}^1$ and $N_2 = n_{\text{exp}} \times m \times n$ evaluations of $y_{\boldsymbol{\theta}}^2$.

Conditional calibration performs well in low dimensions but is strongly affected by the curse of dimensionality. Therefore, the number of parameters must be reduced as much as possible. In this bi-level uncertainty framework, a screening method must satisfy two requirements.

1. The method must fully account for the uncertainty coming from the parameters of Model 1. In particular, a parameter Θ_j may be influential for some realizations of $\mathbf{\Lambda}$ while being non-influential for other ones. Consequently, Θ_j should only be discarded if it can be shown that it has a negligible impact on the output of Model 2 for all possible realizations of $\mathbf{\Lambda}$.
2. The method must accommodate the nested design used for conditional calibration.

The next section is dedicated to the development of such a methodology.

4. Screening methodology for conditional calibration

This section presents the main contribution of the paper. We develop a screening method that identifies the downstream model parameters that must be included in conditional Bayesian calibration. The section is divided into two parts, highlighting the analogy with the HSIC-based methodology (inference first, statistical testing second). Section 4.1 defines a sensitivity measure tailored to the constraints of our problem. In essence, this involves integrating the HSIC with respect to $\mathbb{P}_{\mathbf{\Lambda}}$. We show that the resulting criterion achieves the intended goal. First, it satisfies a mathematical property analogous to the HSIC: it vanishes if and only if the calibration parameter is independent of the output for almost all values of $\boldsymbol{\lambda}$. Second, it can be naturally estimated from the available data, with four possible estimators exhibiting attractive properties. However, these estimates alone are insufficient for screening, just as with HSIC indices. This motivates the use of hypothesis testing in Section 4.2. We explain how to formulate the problem and we show that some HSIC-based test procedures can be easily adapted to our bi-level uncertainty context.

4.1. Measuring sensitivity in the bi-level uncertainty framework

In Section 3, parameter screening was performed in a simplified context where $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\text{nom}}$. The HSIC was used to measure the dependence between any parameter Θ_j and any output of interest $G_{\text{nom}} = \widetilde{f}(\mathbf{\Theta}, \boldsymbol{\lambda}_{\text{nom}})$. One can write:

$$H_j(\boldsymbol{\lambda}_{\text{nom}}) := \text{HSIC}(\Theta_j, G_{\text{nom}}) \ . \tag{4.1}$$

Let $K_{\theta_j} : \mathcal{D}_{\Theta_j} \times \mathcal{D}_{\Theta_j} \to \mathbb{R}$ and $K_g : \mathcal{D}_G \times \mathcal{D}_G \to \mathbb{R}$ denote covariance kernels suitable for handling the variables Θ_j and G_{nom}. The quantity $H_j(\boldsymbol{\lambda}_{\text{nom}})$ can be expressed only in terms of kernel-based moments:

$$\begin{aligned} H_j(\boldsymbol{\lambda}_{\text{nom}}) = {} & \mathbb{E}_{\mathbf{\Theta}\mathbf{\Theta}'}\left[K_{\theta_j}\left(\Theta_j, \Theta_j'\right) K_g\left(\widetilde{f}\left(\mathbf{\Theta}, \boldsymbol{\lambda}_{\text{nom}}\right), \widetilde{f}\left(\mathbf{\Theta}', \boldsymbol{\lambda}_{\text{nom}}\right)\right)\right] \\ & + \mathbb{E}_{\mathbf{\Theta}\mathbf{\Theta}'}\left[K_{\theta_j}\left(\Theta_j, \Theta_j'\right)\right] \mathbb{E}_{\mathbf{\Theta}\mathbf{\Theta}'}\left[K_g\left(\widetilde{f}(\mathbf{\Theta}, \boldsymbol{\lambda}_{\text{nom}}), \widetilde{f}\left(\mathbf{\Theta}', \boldsymbol{\lambda}_{\text{nom}}\right)\right)\right] \\ & - 2\,\mathbb{E}_{\mathbf{\Theta}\mathbf{\Theta}''}\left[K_{\theta_j}\left(\Theta_j, \Theta_j'\right) K_g\left(\widetilde{f}\left(\mathbf{\Theta}, \boldsymbol{\lambda}_{\text{nom}}\right), \widetilde{f}\left(\mathbf{\Theta}'', \boldsymbol{\lambda}_{\text{nom}}\right)\right)\right]. \end{aligned} \tag{4.2}$$

where $\mathbf{\Theta}$, $\mathbf{\Theta}'$ and $\mathbf{\Theta}''$ are independent and identically distributed according to $\mathbb{P}_{\mathbf{\Theta}}$. We refer the reader to Appendix C for further details on the definition, reformulation, and estimation of HSIC indices. The definition of $H_j(\boldsymbol{\lambda}_{\text{nom}})$ naturally extends to any $\boldsymbol{\lambda} \in \mathcal{D}_{\mathbf{\Lambda}}$:

$$\begin{aligned} H_j(\boldsymbol{\lambda}) &= \text{HSIC}\left(\Theta_j, G_{\boldsymbol{\lambda}}\right) \\ &= \text{HSIC}\left(\Theta_j, G \mid \mathbf{\Lambda} = \boldsymbol{\lambda}\right) \end{aligned} \quad \text{with} \quad \begin{cases} G_{\boldsymbol{\lambda}} := \widetilde{f}\left(\mathbf{\Theta}, \boldsymbol{\lambda}\right) \\ G := \widetilde{f}\left(\mathbf{\Theta}, \mathbf{\Lambda}\right). \end{cases}$$

The notation introduced in the second line means that $\boldsymbol{\Lambda}$ must be set to the value $\boldsymbol{\lambda}$ before computing the HSIC. After introducing the function:

$$\begin{aligned} h\left(\boldsymbol{\theta}^1,\boldsymbol{\theta}^2,\boldsymbol{\theta}^3,\boldsymbol{\theta}^4,\boldsymbol{\lambda}\right) = & K_{\theta_j}\left(\theta_j^1,\theta_j^2\right) K_g\left(\widetilde{f}\left(\boldsymbol{\theta}^1,\boldsymbol{\lambda}\right),\widetilde{f}\left(\boldsymbol{\theta}^2,\boldsymbol{\lambda}\right)\right) \\ & + K_{\theta_j}\left(\theta_j^1,\theta_j^2\right) K_g\left(\widetilde{f}\left(\boldsymbol{\theta}^3,\boldsymbol{\lambda}\right),\widetilde{f}\left(\boldsymbol{\theta}^4,\boldsymbol{\lambda}\right)\right) \\ & - 2\, K_{\theta_j}\left(\theta_j^1,\theta_j^2\right) K_g\left(\widetilde{f}\left(\boldsymbol{\theta}^1,\boldsymbol{\lambda}\right),\widetilde{f}\left(\boldsymbol{\theta}^3,\boldsymbol{\lambda}\right)\right), \end{aligned} \tag{4.3}$$

$H_j(\boldsymbol{\lambda})$ can be rewritten as a single expectation:

$$H_j(\boldsymbol{\lambda}) = \mathbb{E}\big[h\big(\boldsymbol{\Theta}^1,\boldsymbol{\Theta}^2,\boldsymbol{\Theta}^3,\boldsymbol{\Theta}^4,\boldsymbol{\lambda}\big)\big] \quad \text{with} \quad \boldsymbol{\Theta}^1 \perp\!\!\!\perp \boldsymbol{\Theta}^2 \perp\!\!\!\perp \boldsymbol{\Theta}^3 \perp\!\!\!\perp \boldsymbol{\Theta}^4 \sim \mathbb{P}_{\boldsymbol{\Theta}}\,. \tag{4.4}$$

Assuming that K_{θ_j}, K_g and $\widetilde{f}$ are measurable, h is measurable as well, and Tonelli's theorem ensures that the function $H_j : \mathcal{D}_{\boldsymbol{\Lambda}} \to \mathbb{R}$ is measurable. This provides a sound theoretical framework for propagating $\mathbb{P}_{\boldsymbol{\Lambda}}$ to the HSIC index. The random variable $H_j(\boldsymbol{\Lambda})$ is therefore well-defined. As already pointed out in Remark 3.4, the vectors $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ are independent, which allows us to rewrite $H_j(\boldsymbol{\Lambda})$ in terms of conditional expectations:

$$H_j(\boldsymbol{\Lambda}) = \mathrm{HSIC}\left(\Theta_j, G \mid \boldsymbol{\Lambda}\right) \tag{4.5}$$

$$\begin{aligned} = & \mathbb{E}_{\boldsymbol{\Theta}\boldsymbol{\Theta}'}\left[K_{\theta_j}\left(\Theta_j,\Theta_j'\right) K_g\left(\widetilde{f}\left(\boldsymbol{\Theta},\boldsymbol{\Lambda}\right),\widetilde{f}\left(\boldsymbol{\Theta}',\boldsymbol{\Lambda}\right)\right)\middle|\boldsymbol{\Lambda}\right] \\ & + \mathbb{E}_{\boldsymbol{\Theta}\boldsymbol{\Theta}'}\left[K_{\theta_j}\left(\Theta_j,\Theta_j'\right)\right]\mathbb{E}_{\boldsymbol{\Theta}\boldsymbol{\Theta}'}\left[K_g\left(\widetilde{f}(\boldsymbol{\Theta},\boldsymbol{\Lambda}),\widetilde{f}\left(\boldsymbol{\Theta}',\boldsymbol{\Lambda}\right)\right)\middle|\boldsymbol{\Lambda}\right] \\ & - 2\,\mathbb{E}_{\boldsymbol{\Theta}\boldsymbol{\Theta}''}\left[K_{\theta_j}\left(\Theta_j,\Theta_j'\right) K_g\left(\widetilde{f}\left(\boldsymbol{\Theta},\boldsymbol{\Lambda}\right),\widetilde{f}\left(\boldsymbol{\Theta}'',\boldsymbol{\Lambda}\right)\right)\middle|\boldsymbol{\Lambda}\right]. \end{aligned} \tag{4.6}$$

Equation (4.5) emphasizes that some residual uncertainty, expressed as a measurable function of $\boldsymbol{\Lambda}$, remains after the HSIC is computed.

4.1.1. Definition of a $\mathbb{P}_\Lambda$-informed sensitivity measure

To account for the uncertainty coming from the upstream model parameters $\boldsymbol{\Lambda}$, a natural idea is to compute the expectation of $H_j(\boldsymbol{\Lambda})$. As a consequence, we define:

$$\forall\, 1 \le j \le p, \quad \mathscr{H}_j := \mathbb{E}_{\boldsymbol{\Lambda}}\left[H_j\left(\boldsymbol{\Lambda}\right)\right]. \tag{4.7}$$

Proposition 4.1 shows that $\mathscr{H}_j$ is a relevant sensitivity measure to detect almost surely (a.s.) any type of statistical dependence between Θ_j and $\widetilde{f}\left(\boldsymbol{\Theta},\boldsymbol{\Lambda}\right)$.

Proposition 4.1. *Let $j \in \{1,\dots,p\}$ be fixed. Assume that K_{θ_j} and K_g are two characteristic kernels. It holds that:*

$$\mathscr{H}_j = 0 \iff H_j\left(\boldsymbol{\Lambda}\right) = 0 \quad \mathbb{P}_{\boldsymbol{\Lambda}}\text{-a.s.} \iff \Theta_j \perp\!\!\!\perp \widetilde{f}\left(\boldsymbol{\Theta},\boldsymbol{\Lambda}\right). \tag{4.8}$$

The detailed proof is provided in Appendix E.1. Since the HSIC is non-negative, the first equivalence is immediate. The second equivalence is slightly more technical. The key argument is that the HSIC, when built using two characteristic kernels, vanishes if and only if the two random objects are independent. Hence, for any $\boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\Lambda}}$, if $H_j(\boldsymbol{\lambda}) = 0$, then $\Theta_j \perp\!\!\!\perp G_{\boldsymbol{\lambda}}$. To prove the direct implication, we verify the factorization property of expectations, while the converse is almost trivial.

Remark 4.2. Proposition 4.1 highlights that testing:

$$(\mathrm{H}_0): \quad \Theta_j \perp\!\!\!\perp \widetilde{f}\,(\boldsymbol{\Theta}, \boldsymbol{\Lambda}) \tag{4.9}$$

is not equivalent to testing:

$$(\widetilde{\mathrm{H}}_0): \quad \forall \boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\Lambda}}, \quad \Theta_j \perp\!\!\!\perp \widetilde{f}\,(\boldsymbol{\Theta}, \boldsymbol{\lambda}) \tag{4.10}$$

which is indeed much stronger. $(\widetilde{\mathrm{H}}_0)$ implies (H_0) but the converse is generally false. Having $\mathscr{H}_j = 0$ only ensures independence between Θ_j and $G_{\boldsymbol{\lambda}}$ for $\mathbb{P}_{\boldsymbol{\Lambda}}$-almost all $\boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\Lambda}}$. In practice, this property is sufficient and constitutes a solid theoretical guarantee. However, one should be aware that there may exist a $\mathbb{P}_{\boldsymbol{\Lambda}}$-null set $\mathcal{D}_{\boldsymbol{\Lambda}}^{\text{null}}$ containing values $\boldsymbol{\lambda}$ such that $\Theta_j \not\perp\!\!\!\perp G_{\boldsymbol{\lambda}}$.

Remark 4.3. Another probabilistic characteristic of $H_j(\boldsymbol{\Lambda})$ could be used to define a sensitivity index. For example, one could consider its essential supremum, defined by:

$$\mathscr{H}_j^{\sup} := \operatorname{ess\,sup} H_j(\boldsymbol{\Lambda}) = \inf\left(\{C > 0: \quad H_j(\boldsymbol{\Lambda}) \leq C \quad \mathbb{P}_{\boldsymbol{\Lambda}}\text{-a.s.}\}\right). \tag{4.11}$$

Just as $\mathscr{H}_j$, the index $\mathscr{H}_j^{\sup}$ is able to characterize independence between Θ_j and G. Following the same arguments as in the proof of Proposition 4.1, one can show that:

$$\mathscr{H}_j^{\sup} = 0 \iff H_j(\boldsymbol{\Lambda}) = 0 \quad \mathbb{P}_{\boldsymbol{\Lambda}}\text{-a.s.} \iff \Theta_j \perp\!\!\!\perp \widetilde{f}\,(\boldsymbol{\Theta}, \boldsymbol{\Lambda}). \tag{4.12}$$

However, intuitively, it seems harder to estimate accurately $\mathscr{H}_j^{\sup}$ from a reasonable amount of simulated data.

4.1.2. Estimation from nested simulation data

We now turn to the estimation of $\mathscr{H}_j$. Remember that one major constraint of our problem is that inference must be performed from the same data as conditional Bayesian calibration. Consequently, the sensitivity indices $\mathscr{H}_j$ must be estimable from these specific data. As explained in Section 3.4, the data are organized as a nested Monte Carlo design.

- The outer level $D_{\boldsymbol{\Lambda},m} = \left\{\boldsymbol{\Lambda}^{(k)}\right\}_{k=1}^{m}$ consists of an m-sample of $\boldsymbol{\Lambda}$.
- The inner level $D_{m,n}$ depends on the adopted sampling strategy. For (S1), it contains an n-sample of all pairs $\left(\boldsymbol{\Theta}^{(k)}, \boldsymbol{Y}^{(k)}\right)$ associated with the values $\boldsymbol{\Lambda}^{(k)}$:

$$D_{\boldsymbol{\Theta}^{(k)}\boldsymbol{Y}^{(k)},n} = \left\{D_{\boldsymbol{\Theta}^{(k)},n}; D_{\boldsymbol{Y}^{(k)},n}\right\} = \left\{\left(\boldsymbol{\Theta}^{(kl)}, \boldsymbol{Y}^{(kl)}\right)\right\}_{1\leq l\leq n}. \tag{4.13}$$

  For (S2), it contains an n-sample of all pairs $\left(\boldsymbol{\Theta}, \boldsymbol{Y}^{(k)}\right)$:

$$D_{\boldsymbol{\Theta}\boldsymbol{Y}^{(k)},n} = \left\{D_{\boldsymbol{\Theta},n}; D_{\boldsymbol{Y}^{(k)},n}\right\} = \left\{\left(\boldsymbol{\Theta}^{(l)}, \boldsymbol{Y}^{(kl)}\right)\right\}_{1\leq l\leq n}. \tag{4.14}$$

As already mentioned, a simple post-processing step allows transforming observations of the output vector $\boldsymbol{Y}$ into observations of the output of interest G (depending on the chosen screening approach). In the following,

all introduced notations apply to the post-processed data:

$$\begin{aligned} &\bullet \quad D^{\mathrm{S}_1}_{m,n} = \left\{D_{\boldsymbol{\Theta}^{(k)}G^{(k)},n}\right\}_{1\le k\le m} \quad \text{with} \quad D_{\boldsymbol{\Theta}^{(k)}G^{(k)},n} = \left\{\left(\boldsymbol{\Theta}^{(kl)}, G^{(kl)}\right)\right\}_{1\le l\le n} ; \\ &\bullet \quad D^{\mathrm{S}_2}_{m,n} = \left\{D_{\boldsymbol{\Theta}G^{(k)},n}\right\}_{1\le k\le m} \quad \text{with} \quad D_{\boldsymbol{\Theta}G^{(k)},n} = \left\{\left(\boldsymbol{\Theta}^{(l)}, G^{(kl)}\right)\right\} . \end{aligned} \tag{4.15}$$

Based on these data, the estimation of $\mathscr{H}_j = \mathbb{E}_{\boldsymbol{\Lambda}}\left[H_j(\boldsymbol{\Lambda})\right]$ is fairly natural. As a first step, the expectation over $\boldsymbol{\Lambda}$ can be approximated by the empirical average over the values $\boldsymbol{\Lambda}^{(k)}$:

$$\widehat{\mathscr{H}}_j := \widehat{\mathscr{H}}_j\left(D_{\boldsymbol{\Lambda},m}\right) = \frac{1}{m}\sum_{k=1}^{m} H_j\left(\boldsymbol{\Lambda}^{(k)}\right) \quad \text{with} \quad H_j\left(\boldsymbol{\Lambda}^{(k)}\right) = \mathrm{HSIC}\left(\Theta_j, G^{(k)} \mid \boldsymbol{\Lambda}^{(k)}\right). \tag{4.16}$$

As it embeds the HSIC, the function H_j is not analytically tractable. The second step therefore consists in estimating all values $H_j\left(\boldsymbol{\Lambda}^{(k)}\right)$ from the data provided in $D_{m,n}$. Assuming that this estimation step is feasible, the final estimator is given by:

$$\widehat{\widehat{\mathscr{H}}}_j\left(D_{m,n}\right) := \frac{1}{m}\sum_{k=1}^{m} \widehat{H}_j\left(\boldsymbol{\Lambda}^{(k)}\right) . \tag{4.17}$$

where $\widehat{H}_j\left(\boldsymbol{\Lambda}^{(k)}\right)$ denotes an estimator of $H_j\left(\boldsymbol{\Lambda}^{(k)}\right)$. Two standard estimators of the HSIC are commonly used: the U- and V-statistic estimators (see Appendix C.2). In both cases, the estimate can be computed from an n-sample of joint observations. In our context, for each quantity $H_j(\boldsymbol{\Lambda}^{(k)})$ to be estimated, this sample is provided by $D_{\boldsymbol{\Theta}^{(k)}G^{(k)},n}$ under (S1) and by $D_{\boldsymbol{\Theta}G^{(k)},n}$ under (S2).

Finally, since each quantity $H_j\left(\boldsymbol{\Lambda}^{(k)}\right)$ can be estimated using two different HSIC estimators on two different datasets, four distinct estimators of $\mathscr{H}_j$ can be constructed:

$$\begin{aligned} &\textbf{Estimator 1:} \quad \widehat{\widehat{\mathscr{H}}}_j^{\,\mathrm{U}}\left(D^{\mathrm{S}_1}_{m,n}\right) := \frac{1}{m}\sum_{k=1}^{m} \widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(k)}G^{(k)},n}\right) =: \widehat{\widehat{\mathscr{H}}}_j^{\,1} ; \\ &\textbf{Estimator 2:} \quad \widehat{\widehat{\mathscr{H}}}_j^{\,\mathrm{V}}\left(D^{\mathrm{S}_1}_{m,n}\right) := \frac{1}{m}\sum_{k=1}^{m} \widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}^{(k)}G^{(k)},n}\right) =: \widehat{\widehat{\mathscr{H}}}_j^{\,2} ; \\ &\textbf{Estimator 3:} \quad \widehat{\widehat{\mathscr{H}}}_j^{\,\mathrm{U}}\left(D^{\mathrm{S}_2}_{m,n}\right) := \frac{1}{m}\sum_{k=1}^{m} \widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G^{(k)},n}\right) =: \widehat{\widehat{\mathscr{H}}}_j^{\,3} ; \\ &\textbf{Estimator 4:} \quad \widehat{\widehat{\mathscr{H}}}_j^{\,\mathrm{V}}\left(D^{\mathrm{S}_2}_{m,n}\right) := \frac{1}{m}\sum_{k=1}^{m} \widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}G^{(k)},n}\right) =: \widehat{\widehat{\mathscr{H}}}_j^{\,4} . \end{aligned} \tag{4.18}$$

Above, the superscripts U and V indicate which HSIC estimator is used. Hereafter, $\widehat{\widehat{\mathscr{H}}}_j$ will refer to any of the four proposed estimators.

Remark 4.4. The estimator $\widehat{\widehat{\mathscr{H}}}_j$ in equation (4.17) is written with a double hat to emphasize the presence of two approximation levels. The estimator $\widehat{\mathscr{H}}_j$ in equation (4.16) is not tractable, but it plays a crucial role in the analysis of convergence speeds (see Prop. 4.5). In particular, it allows the two sources of error to be disentangled in the proofs of Appendix E.2, thereby reducing the analysis to classical arguments.

In order to find out which estimator(s) should be preferred, the statistical properties (bias, consistency, convergence rate) of all four estimators need to be investigated. For obvious reasons, a minimal requirement for

TABLE 2. Properties of the four proposed estimators.

Estimators \ Properties	Non-negativity	Unbiasedness	Consistency	MSE decay rate
Estimator 1	✗	✓	✓	$\mathcal{O}\left(\frac{1}{m}\right)$
Estimator 2	✓	✗	✓	$\mathcal{O}\left(\frac{1}{m}+\frac{1}{n^2}\right)$
Estimator 3	✗	✓	✓	$\mathcal{O}\left(\frac{1}{m}+\frac{1}{n}\right)$
Estimator 4	✓	✗	✓	$\mathcal{O}\left(\frac{1}{m}+\frac{1}{n}\right)$

the estimator is consistency. It must be proven that $\widehat{\widehat{\mathscr{H}_j}}$ converges in probability to the target quantity $\mathscr{H}_j$ as the sample sizes grow $(m, n \to \infty)$. To establish convergence in probability, it is often more convenient to prove convergence in mean square. This amounts to showing that the mean square error (MSE) tends to zero as the sample sizes become arbitrarily large. In our setting, this reads:

$$\mathrm{MSE}\left(\widehat{\widehat{\mathscr{H}_j}}\right) := \mathbb{E}\left[\left|\widehat{\widehat{\mathscr{H}_j}} - \mathscr{H}_j\right|^2\right] \xrightarrow[m,n\to\infty]{} 0. \tag{4.19}$$

Convergence in mean square is stronger than convergence in probability, but it is often easier to demonstrate, and it reveals convergence rates more readily. Proposition 4.5 states that all four estimators are consistent, and that two of them are unbiased. In addition, upper bounds on the decay rates of the MSE are provided for the four estimators. These results are particularly valuable, as they can guide the allocation of the simulation budget $(N_2 = n_{\text{exp}} \times m \times n)$ between the two levels of sampling.

Proposition 4.5. *Let $j \in \{1, \ldots, p\}$ be fixed. Assume that K_{θ_j} and K_g are two bounded kernels. The statistical properties of the four estimators of $\mathscr{H}_j$ are summarized in Table 2.*

The detailed proofs can be found in Appendix E.2. They are much inspired from ideas presented in a recent work by Fellmann et al. [71].

Remark 4.6. It is worth noting that the assumption of bounded kernels is pretty standard in the literature related to HSIC indices, especially to obtain concentration inequalities (see Thm. 4 in [72]). This assumption is not very stringent as it is verified by most translation-invariant kernels used in statistical applications (especially Gaussian and Matérn kernels).

Remark 4.7. It should be clearly understood that stating the MSE is $\mathcal{O}(1/m)$ does not imply that the MSE is asymptotically equivalent to $1/m$ as $m, n \to \infty$. Rather, this indicates that the MSE can be bounded above by C/m for some positive constant C, potentially very large. In a hypothetical scenario where the theoretical values $\mathscr{H}_j$ are known and the chained model y can be evaluated without limitation, one could have compared the convergence rates of Proposition 4.5 with the true rates, which might be much faster. Therefore, the rates from Proposition 4.5 should be interpreted as worst-case rates.

Proposition 4.5 provides valuable insights into the statistical properties of the four estimators. All of them are consistent and exhibit convergence rates comparable to those of classical Monte Carlo estimators. This guarantees that the indices $\mathscr{H}_j$ can be estimated accurately from the available simulation data. Several additional points are worth highlighting.

- **Estimator 1.** It is unbiased, which is natural since it is expressed as the average of m U-statistics. Its MSE decays at a rate of $\mathcal{O}(1/m)$. Accordingly, it appears more efficient to prioritize large values of m

(the number of points used to estimate the expectation over $\mathbf{\Lambda}$) rather than n (the number of points used to estimate HSIC indices). To reduce the MSE as much as possible, one could be tempted to take $m \gg n$. However, as will be seen later, this is not what is done in practice, as doing so would compromise the performance of conditional calibration, which remains the ultimate objective. One typically works with $n \gg m$ to ensure that a large n-sample of $\left(\mathbf{\Theta}, G^{(k)}\right)$ is available for each value $\mathbf{\Lambda}^{(k)}$.

- **Estimator 2.** It is biased, as expected due to the use of V-statistics. Unlike Estimator 1, the approximation error cannot be fully eliminated simply by increasing m. The bias introduced by the V-statistics gives rise to the $1/n^2$ term in the MSE, which can only be reduced by increasing n. In light of the previous remark, choosing $n \gg m$ (for conditional calibration purposes) makes the V-statistic bias practically negligible, and Estimators 1 and 2 converge at essentially the same rate.
- **Estimators 3 and 4.** The remarks made for Estimators 1 and 2 regarding the presence or absence of bias also apply to Estimators 3 and 4. Importantly, their convergence rate of $\mathcal{O}\left(1/m + 1/n\right)$ is slower than that of Estimators 1 and 2. This rules out (S2) in favor of (S1), which therefore emerges as the most effective sampling strategy for achieving maximum accuracy under a limited simulation budget.
- **Estimator 4.** As shown in the consistency proof of $\widehat{\widehat{\mathscr{H}}}_j^{\,4}$ (see Appendix E.2), the MSE satisfies $\mathcal{O}\left(1/m + 1/n + 1/n^2\right)$ where:
  - $1/m$ is due to the empirical mean over the m observations of $\mathbf{\Lambda}$.
  - $1/n$ is due to Strategy (S2);
  - $1/n^2$ is due to V-statistics.

Because of their slower convergence rates, Estimators 3 and 4 are excluded from further analysis. They are neither considered in Section 4.2 nor implemented in Section 5. By contrast, Estimators 1 and 2 are both retained, as neither is inherently superior to the other. Indeed, they inherit the respective advantages and drawbacks of the U- and V-statistic estimators of the HSIC (unbiased but sometimes negative for Estimator 1, biased but always non-negative for Estimator 2).

Remark 4.8. The alternative sensitivity index $\mathscr{H}_j^{\text{null}}$ introduced in equation (4.11) can also be estimated from the nested Monte Carlo design. A counterpart to equation (4.17) is given by:

$$\widehat{\widehat{\mathscr{H}}}_j^{\,\text{sup}} := \max_{1 \le k \le m} \widehat{H}_j\left(\mathbf{\Lambda}^{(k)}\right) . \tag{4.20}$$

Once again, four estimators can be constructed by applying the two possible HSIC estimators to the two nested designs. Moreover, it can be easily shown that these estimators are consistent. The proofs are very similar to those presented in Appendix E.2. The keys arguments to demonstrate consistency are the extremal law of large numbers and appropriate concentration inequalities [72]. However, deriving explicit convergence rates may be considerably more challenging.

At this stage, the work is not yet complete, because the estimates of the indices $\widehat{\widehat{\mathscr{H}}}_j$ are not directly usable for parameter screening. The problem is the same as with HSIC indices, if not worse. Recall that the scale of an HSIC index, measured via the MMD, depends on the marginal distributions of the two random objects and on the kernels chosen to handle these objects. Consequently, the magnitude of the MMD scale is difficult to interpret: given an HSIC estimate, it is impossible to determine whether the underlying HSIC value is zero or not. This is precisely why a proper independence test must be employed to make the final decision.

In the case of $\mathscr{H}_j$, the same problem is encountered, as the index is expressed as the average of several HSIC estimates. The difficulty is further compounded by the fact that the related HSIC terms are defined conditional on an event $\mathbf{\Lambda} = \boldsymbol{\lambda}$. For each term, dependence is measured between Θ_j and $G_{\boldsymbol{\lambda}} = \widetilde{f}(\mathbf{\Theta}, \boldsymbol{\lambda})$ where the marginal distribution of $G_{\boldsymbol{\lambda}}$ naturally varies with $\boldsymbol{\lambda}$. For two different values $\boldsymbol{\lambda}_1 \neq \boldsymbol{\lambda}_2$, the MMD scales associated with $H_j(\boldsymbol{\lambda}_1)$ and $H_j(\boldsymbol{\lambda}_2)$ are therefore not comparable. In short, as with HSIC indices, estimation alone is insufficient: the screening problem must be reformulated as a hypothesis test, and a proper test procedure must be developed.

4.2. Testing independence in the bi-level uncertainty framework

4.2.1. Problem formulation

The objective is to construct a statistical procedure based on the data $D_{m,n}^{\mathrm{S}_1}$ that allows, for each parameter Θ_j, discrimination between the following two hypotheses:

$$(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp G \quad \text{vs.} \quad (\mathrm{H}_1) : \Theta_j \not\perp\!\!\!\perp G \quad \text{with} \quad G = \widetilde{f}\,(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\,. \tag{4.21}$$

By virtue of Proposition 4.1, these statements may be rewritten as:

$$(\mathrm{H}_0) : \mathscr{H}_j = 0 \quad \text{vs.} \quad (\mathrm{H}_1) : \mathscr{H}_j > 0\ . \tag{4.22}$$

Naturally, the test statistic will be either Estimator 1 or 2. Given that (H_0) coincides with $\mathscr{H}_j = 0$, the test of independence based on $\widehat{\widehat{\mathscr{H}}}_j$ is one-sided (upper-tailed), meaning that (H_0) is rejected when the test statistic takes a sufficiently large value. More precisely, the rejection threshold is the quantile of order $1-\alpha$ of the null distribution, with the significance level α ensuring strict control of Type-I error. Let $D_{m,n}^{\mathrm{obs}}$ denote the available data. Note that:

- $D_{m,n}^{\mathrm{obs}}$ is the observed realization of the random object $D_{m,n}$ that generates the data.
- $\widehat{\widehat{\mathscr{H}}}_j\left(D_{m,n}^{\mathrm{obs}}\right)$ is the observed realization of the test statistic $\widehat{\widehat{\mathscr{H}}}_j\left(D_{m,n}\right)$.

The p-value of the test is defined as the probability, under (H_0), of observing a value of the test statistic at least as extreme as the one actually observed. In our context, it is given by:

$$p_j := \mathbb{P}_{\mathrm{H}_0}\left(\widehat{\widehat{\mathscr{H}}}_j\left(D_{m,n}\right) > \widehat{\widehat{\mathscr{H}}}_j\left(D_{m,n}^{\mathrm{obs}}\right)\right). \tag{4.23}$$

For the HSIC-based test of independence, three distinct test procedures can be used to estimate the p-value, each suited to a particular range of sample sizes (see Appendix C.3).

- **Large sample sizes.** For $n \geq 500$, the asymptotic test aims to approximate the spectral distribution ruling the asymptotic behavior under (H_0) of the rescaled test statistic. The p-value is estimated at the cost of $\mathcal{O}(n^2)$ operations.
- **Small sample sizes.** For $n \leq 100$, only the permutation-based test is applicable. This procedure involves randomly permuting the output data and recomputing the test statistic for each permuted sample, thereby generating an empirical approximation of the null distribution.The p-value is then estimated as the proportion of simulated values that exceed the original observed value. The computational complexity of this approach is $\mathcal{O}(B\,n^2)$, where B denotes the number of permutations.
- **Intermediate sample sizes.** For $100 \leq n \leq 500$, the non-asymptotic Gamma test enables retrieving the p-value of the permutation-based test, but without performing any permutation. The computational complexity is therefore reduced to $\mathcal{O}(n^2)$.

We now aim to adapt these test procedures to our specific screening problem. From a purely technical standpoint, studying the asymptotic behavior of the (rescaled) test statistic appears to be challenging. Indeed, because of the bi-level sampling scheme, classical asymptotic results for U- and V-statistics cannot be directly applied. In any case, given that conditional calibration requires $N_2 = n_{\mathrm{exp}} \times m \times n$ evaluations of the downstream model, the asymptotic framework is clearly unrealistic. By contrast, the two alternative test procedures are relatively easy to implement. A permutation-based test is developed in the next section, while a non-asymptotic Gamma test is presented in Appendix D.2.

4.2.2. Permutation-based test procedure

At this stage, the objective is to test $(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp G$ using the nested Monte Carlo design $D^{\mathrm{S}_1}_{m,n}$. Let $D^{\mathrm{obs}}_{m,n}$ denote the available data. From a practical viewpoint, $D^{\mathrm{obs}}_{m,n}$ is the generated nested design, which will later be used for conditional Bayesian calibration. From a probabilistic viewpoint, $D^{\mathrm{obs}}_{m,n}$ is the only observed realization of $D^{\mathrm{S}_1}_{m,n}$, which represents the underlying data generation process. As explained in Section 3.4, $D^{\mathrm{obs}}_{m,n}$ is composed of an m-sample of $\mathbf{\Lambda}$:

$$D^{\mathrm{obs}}_{\mathbf{\Lambda},m} := \left\{\boldsymbol{\lambda}^{(k)}\right\}_{1\leq k\leq m} , \tag{4.24}$$

together with an n-sample of each pair $\left(\mathbf{\Theta}^{(k)}, G^{(k)}\right)$:

$$D^{\mathrm{obs}}_k := D^{\mathrm{obs}}_{\mathbf{\Theta}^{(k)}G^{(k)},n} = \left\{\left(\boldsymbol{\theta}^{(kl)}, g^{(kl)}\right)\right\}_{1\leq l\leq n} \quad \text{with} \quad g^{(kl)} = \widetilde{f}\left(\boldsymbol{\theta}^{(kl)}, \boldsymbol{\lambda}^{(k)}\right). \tag{4.25}$$

In these equations, lowercase letters indicate that the data are fixed during the test procedure.

To simulate the distribution of the test statistic under (H_0), we draw inspiration from the permutation-based test (see Appendix C.3), which is arguably the simplest strategy to do so in a finite-sample setting. Recall the equivalence stated in Proposition 4.1:

$$\Theta_j \perp\!\!\!\perp G \iff \exists\, \widetilde{\mathcal{D}}_{\mathbf{\Lambda}} \subseteq \mathcal{D}_{\mathbf{\Lambda}} \text{ such that } \mathbb{P}_{\mathbf{\Lambda}}\left(\widetilde{\mathcal{D}}_{\mathbf{\Lambda}}\right) = 1 \text{ and } \forall\, \boldsymbol{\lambda} \in \widetilde{\mathcal{D}}_{\mathbf{\Lambda}},\ \Theta_j \perp\!\!\!\perp G_{\boldsymbol{\lambda}} . \tag{4.26}$$

While not fully rigorous, it is reasonable to assume that the values $\boldsymbol{\lambda}^{(k)}$ belong to $\widetilde{\mathcal{D}}_{\mathbf{\Lambda}}$. Under this assumption, one has:

$$(\mathrm{H}_0) \implies \forall\, 1 \leq k \leq m,\ \Theta_j^{(k)} \perp\!\!\!\perp G^{(k)} \text{ where } G^{(k)} = G_{\boldsymbol{\lambda}^{(k)}} = \widetilde{f}\left(\mathbf{\Theta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right). \tag{4.27}$$

A necessary condition to simulate (H_0) is therefore to enforce independence within each pair $\left(\mathbf{\Theta}^{(k)}, G^{(k)}\right)$. To achieve this, it suffices to apply random permutations to the output data within each sample D^{obs}_k. Let τ_k denote the permutation applied to the k-th sample. In practice, τ_k is randomly selected from the set $\mathbb{S}_n$ of all permutations of $\{1, \ldots, n\}$. The corresponding permuted sample is obtained as follows:

$$D^{\tau_k}_k := D^{\tau_k}_{\mathbf{\Theta}^{(k)}G^{(k)},n} = \left\{\left(\boldsymbol{\theta}^{(kl)}, g^{(k\,\tau_k(l))}\right)\right\}_{1\leq l\leq n} . \tag{4.28}$$

Let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$ denote the collection of all permutations. Accordingly, $D^{\boldsymbol{\tau}}_{m,n}$ denotes the nested design derived from $D^{\mathrm{obs}}_{m,n}$ after applying the permutations τ_k to the samples D^{obs}_k. Then, the reasoning is straightforward.

- Computing $\widehat{H}_j\left(D^{\tau_k}_k\right)$ produces a realization of $\widehat{H}_j\left(\boldsymbol{\lambda}^{(k)}\right)$ under $\Theta_j \perp\!\!\!\perp \widetilde{f}\left(\mathbf{\Theta}^{(k)}, \boldsymbol{\lambda}^{(k)}\right)$.
- Computing $\widehat{\widehat{\mathscr{H}_j}}\left(D^{\boldsymbol{\tau}}_{m,n}\right)$ produces a realization of $\widehat{\widehat{\mathscr{H}_j}}\left(D_{m,n}\right)$ under (H_0).

The complete procedure used to produce a pseudo-realization of the null distribution is illustrated in Figure 4. Repeating this procedure B times, each time with a new collection of permutations $\boldsymbol{\tau}_b = (\tau_{b1}, \ldots, \tau_{bm}) \in (\mathbb{S}_n)^m$, yields an empirical approximation of the null distribution in the form of a B-sample. The p-value is then estimated by:

$$\widehat{p}_j = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\left\{\widehat{\widehat{\mathscr{H}_j}}\left(D^{\boldsymbol{\tau}_b}_{m,n}\right) > \widehat{\widehat{\mathscr{H}_j}}\left(D^{\mathrm{obs}}_{m,n}\right)\right\}} . \tag{4.29}$$

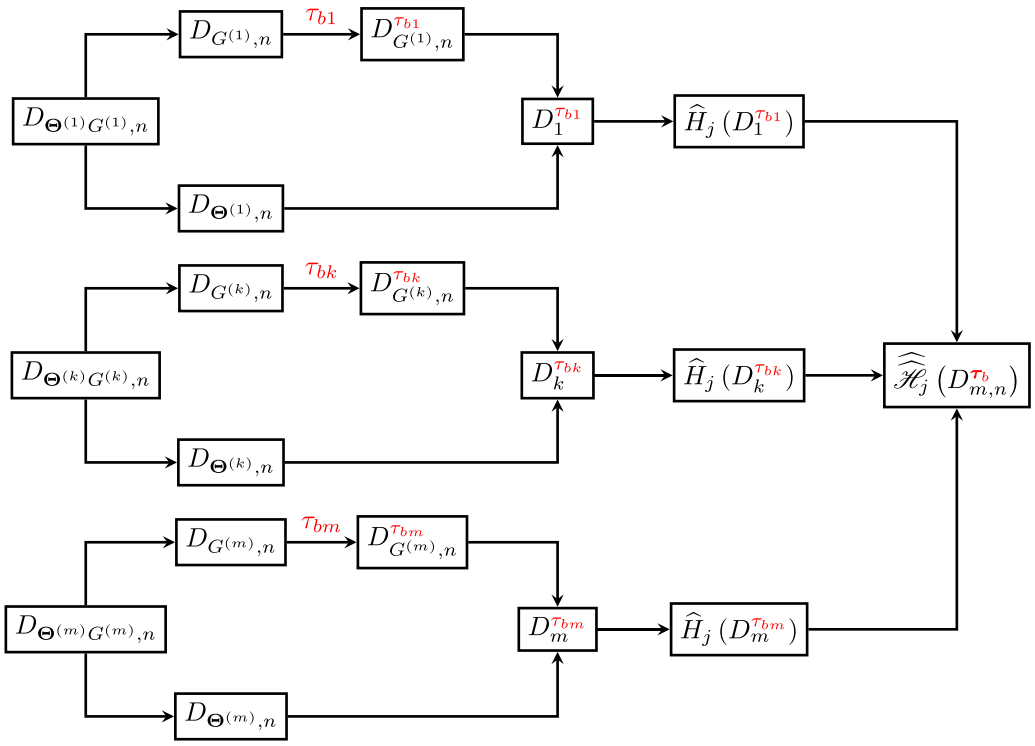

FIGURE 4. Schematic of the procedure used to simulate the b-th realization of the test statistic $\widehat{\widehat{\mathscr{H}}}_j$ under the null hypothesis $(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp \widetilde{f}\,(\boldsymbol{\Theta}, \boldsymbol{\Lambda})$. To this end, a realization of each estimator $\widehat{H}_j\left(\boldsymbol{\lambda}^{(k)}\right)$ must be simulated under independence between Θ_j and $\widetilde{f}\left(\boldsymbol{\Theta}, \boldsymbol{\lambda}^{(k)}\right)$. This is achieved by applying a random permutation τ_{bk} to each sample $D_{G^{(k)},n}$. The permuted output sample $D^{\tau_{bk}}_{G^{(k)},n}$ and the input sample $D_{\boldsymbol{\Theta}^{(k)},n}$ are then reassembled to form a sample $D^{\tau_{bk}}_k$ which is used to estimate $H_j\left(\boldsymbol{\lambda}^{(k)}\right)$. The empirical average of these values yields one realization of $\widehat{\widehat{\mathscr{H}}}_j$ under (H_0).

The overall computational complexity of this test procedure is $\mathcal{O}\left(m\,B\,n^2\right)$. Since the number of permutations is often set to $B = 500$, this can become costly, even for intermediate sample sizes ($100 \leq n \leq 500$). Using the non-asymptotic Gamma test described in Appendix D.2 reduces the complexity to $\mathcal{O}\left(m\,n^2\right)$.

4.3. Summary of contributions

First, we have introduced a sensitivity measure that takes into account the bi-level parametric uncertainty arising from the chaining between Models 1 and 2 (see Fig. 3). This framework leads to consider a set of $\mathbb{P}_{\boldsymbol{\Lambda}}$-informed HSIC indices computed between the parameters Θ_j of Model 2 and the output of interest $G = \widetilde{f}\,(\boldsymbol{\Theta}, \boldsymbol{\Lambda})$. A key result (see Prop. 4.1) is that $\mathscr{H}_j$ provides an exact characterization of the probabilistic independence between Θ_j and G. This makes $\mathscr{H}_j$ a particularly relevant criterion for conditional calibration since the influence exerted by Θ_j is quantified by integrating over the full uncertainty on $\boldsymbol{\Lambda}$.

Second, we have shown that four distinct estimators of $\mathscr{H}_j$ can be constructed from the nested Monte Carlo design intended for conditional calibration. Their statistical properties have been thoroughly investigated (see Prop. 4.5) in order to identify the most reliable estimators. In particular, we have demonstrated that two estimators are unbiased, and that all four estimators are consistent, with convergence rates comparable to those

of classical Monte Carlo estimators. These findings have confirmed that $\mathscr{H}_j$ can be accurately estimated using a reasonable simulation budget. They have also enabled us to discard two estimators based on their MSE decay rates.

Finally, we have highlighted that point estimation alone is insufficient for parameter screening due to the kernel-based metric underlying $\mathscr{H}_j$. Thanks to Proposition 4.1, the screening problem has been reformulated as a hypothesis testing problem, where the null hypothesis corresponds to $\mathscr{H}_j = 0$. Two test procedures, largely inspired by what exists for HSIC indices, have been developed to assess the statistical significance of the estimated indices.

5. Refined analysis of the fission gas behavior model

In this section, we revisit the case study introduced in Section 3.1. The goal remains to screen the parameters $\boldsymbol{\theta}$ of the fission gas behavior model, but this time taking into account the residual uncertainty in the thermal model conductivity λ. To do so, we apply the methodology developed in Section 4. The following procedure needs to be executed.

- Generate the design $D_{\Lambda,m}$ with $m = 20$ and $\mathbb{P}_\Lambda$ estimated in [32].
- Generate all the designs $D_{\boldsymbol{\Theta}^{(k)},n}$ with $n = 200$ and $\mathbb{P}_{\boldsymbol{\Theta}}$ derived from expert judgment.
- Evaluate the chained model y to compute the output data $D_{\boldsymbol{Y}^{(k)},n}$.
- Estimate the sensitivity indices with respect to $G = \widetilde{f}\,(\boldsymbol{\Theta}, \Lambda)$.
- Run the test procedures to estimate the p-values.
- Select the most influential parameters using the p-values ($\alpha = 5\%$) or the detection rates ($\beta = 80\%$).

The definition of the analyzed output G depends on the chosen screening approach (see Rem. 3.3). The final nested design $D^{\mathrm{S}_1}_{m,n}$ is obtained by gathering the input and output data according to equation (3.18). This sampling strategy entails the subsequent use of Estimators 1 and 2. The two test procedures developed in Section 4.2 apply to Estimator 2, whereas only the permutation-based test applies to Estimator 1.

Remark 5.1. Looking at Table 2, one might wonder why m is chosen much smaller than n, whereas the convergence rates of the MSE of Estimators 1 and 2 suggest the opposite. This situation is dictated by the conditional calibration algorithm, which requires large designs in the parameter space $\mathcal{D}_{\boldsymbol{\Theta}}$ to work efficiently. This illustrates a situation where estimator performance is suboptimal because the data were not originally intended for GSA.

The results obtained for Approach A are reported in Figure 5, which is arranged in the same way as Figure 2. It displays the test results for $(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp Y_i$ across all (i, j) pairs. This is done for the two possible estimators and the permutation-based test. For Estimator 1, the detection rates ρ_j summarizing the n_{exp} tests performed for each parameter Θ_j are reported in Column A of Table 3. The detection rates for Estimator 2 are not shown as they exactly coincide with those of Estimator 1.

We observe that parameters Θ_5, Θ_6, Θ_7, and Θ_{11} are influential on the RGF of almost all fuel rods. For these four parameters, the detection rates exceed 90%. Then comes Θ_{10} with $\rho_{10} = 65\%$. If the decision threshold is set to $\beta = 60\%$, only these five parameters are selected. Nevertheless, it is worth noting that the influence of Θ_2 and Θ_3 is not negligible, with detection rates $\rho_2 = 22.5\%$ and $\rho_3 = 47.5\%$, respectively.

Comparing these results with those from Section 3.3 reveals two important insights. First, the four most influential parameters are identified in the same way. Second, our screening methodology appears to be more conservative: the selected set of parameters is larger, and the detection rates are higher for all parameters. This behavior was expected, as the null hypothesis is much stronger. Indeed, for $(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp \widetilde{f}\,(\boldsymbol{\Theta}, \Lambda)$ to hold,

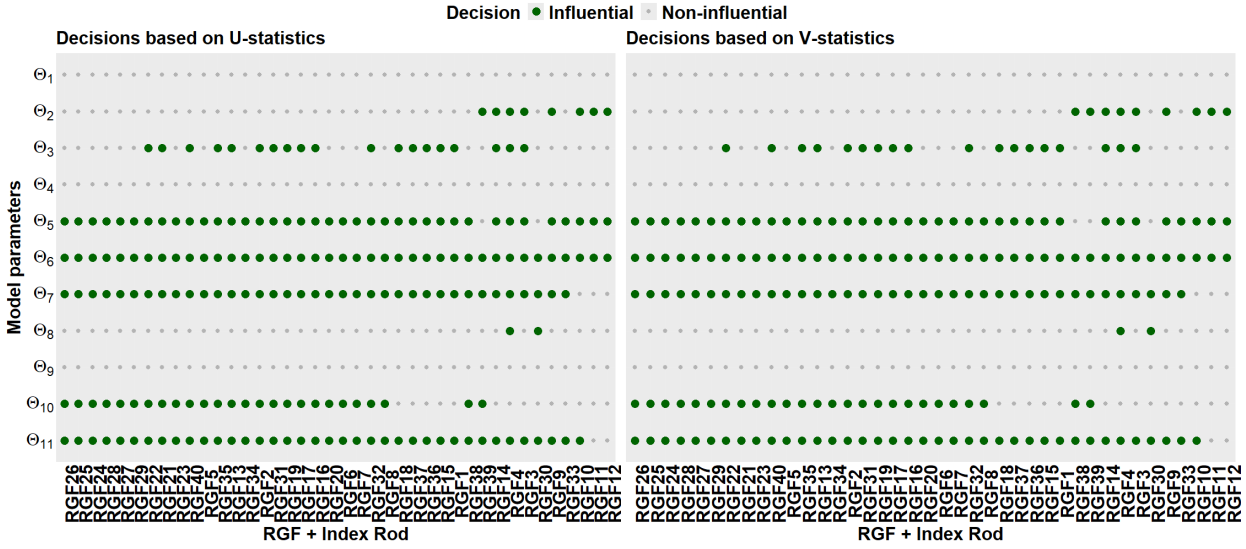

FIGURE 5. Results obtained for Approach A after incorporating the uncertainty on Λ. For each fuel rod, the influence of each parameter Θ_j is assessed using the test procedure developed in Section 4.2.2. The test statistic is either Estimator 1 (left) or Estimator 2 (right). The significance level is set to $\alpha = 5\%$. The associated detection rates are reported in Table 3.

TABLE 3. Comparison of the screening results obtained for the three considered approaches after incorporating the uncertainty on Λ. Only the results obtained with Estimator 1 are shown.

Results with U-statistics	Detection rates	*p*-values	
Parameters \ Approach	A	B	C
Θ_1	0%	1	1
Θ_2	20%	0	0.997
Θ_3	47.5%	0.011	0
Θ_4	0%	1	1
Θ_5	95%	0	0
Θ_6	100%	0	0
Θ_7	92.5%	0	0
Θ_8	5%	0.963	0.903
Θ_9	0%	1	1
Θ_{10}	65%	0.658	0.002
Θ_{11}	95%	0	0

the parameter Θ_j must be independent of $\widetilde{f}(\boldsymbol{\Theta}, \lambda)$ for almost all $\lambda \in \mathcal{D}_\Lambda$, not merely for the nominal value $\lambda = \lambda_{\text{nom}}$. Consequently, (H_0) is rejected more frequently, which increases the power of the test procedure.

The *p*-values obtained for Approaches B and C are also reported in Table 3. The results from the three approaches are highly consistent, each leading to the selection of the same four most influential parameters, namely Θ_5, Θ_6, Θ_7, and Θ_{11}. Despite this overall agreement, a few differences between the three approaches can still be observed. Approach C, which targets the least-squares criterion $L = \widetilde{f}^C(\boldsymbol{\Theta}, \Lambda)$, readily detects Θ_{10} but discards Θ_2. This suggests that Θ_{10} plays an essential role in the conditional calibration procedure, unlike Θ_2.

Interestingly, for Approach B, which operates on the vector $\boldsymbol{Y} = \widetilde{f}^{B}(\boldsymbol{\Theta}, \Lambda)$ of all RGFs, the parameter Θ_{10} is not detected. Recall that in this approach the multivariate output is handled via a multivariate kernel obtained by tensorization of univariate Gaussian kernels. This construction implicitly standardizes the output components, thereby balancing their relative contributions. Such an effect is often desirable for training a metamodel, but in our setting it tends to penalize parameters whose influence is confined to a small subset of outputs. This may partly explain why Θ_{10} is not selected under Approach B. To facilitate the detection of Θ_{10}, it may be worth considering the weighted PCA-kernel (see Prop. 2 in [58]). It is constructed as a weighted linear combination (using the PCA eigenvalues as weights) of univariate kernels applied to the principal components. It has the drawback of not being characteristic (see Appendix C.5.3), but it offers improved performance in terms of detection capabilities. This alternative kernel has been experimented and indeed allows the successful detection of Θ_{10}.

Overall, the three approaches converge towards the selection of six parameters: Θ_3, Θ_5, Θ_6, Θ_7, Θ_{10}, and Θ_{11}. Thus, the proposed screening methodology corroborates the findings of Section 3.2 and allows for a more broader selection of parameters. Additionally, all conclusions are robust with respect to both the choice of estimator and test procedure. In particular, the p-values returned by the non-asymptotic Gamma test closely match those obtained from the permutation-based test. In Figure E.1, the histogram of pseudo-realizations of Estimator 2 under (H_0) is compared with the parametric Gamma approximation, revealing a good fit.

6. Conclusions and perspectives

This paper proposes a screening methodology aimed at reducing the number of parameters involved in the conditional Bayesian calibration of the downstream component of a computational chain. The proposed approach properly accounts for the residual uncertainty arising from the calibration of the upstream component. Besides, it does not induce any additional computational cost, as it can be implemented directly from the nested Monte Carlo design required to carry out conditional calibration.

The proposed sensitivity measure is defined as an integrated version of the HSIC. It explicitly incorporates the upstream uncertainty level and is able to correctly identify the most influential parameters of the downstream model, regardless of the values taken by the upstream model parameters. The proposed estimators are consistent, and their convergence rates are comparable to those of standard Monte Carlo estimators, allowing for accurate inference even with a limited simulation budget. Decision making relies on a test of independence that can be performed using the provided nested design. Two test procedures have been developed, both inspired by existing approaches for HSIC indices. While the permutation-based test entails a high computational complexity, this burden can be alleviated by resorting to the non-asymptotic Gamma test.

The methodology has been successfully applied to an industrial case study from the nuclear industry, involving the chaining between a thermal model and a fission gas behavior model within the ALCYONE code. The results confirm that the proposed approach fulfills its intended purpose: it leads to the selection of a broader set of influential parameters, as a direct consequence of the rigorous treatment of upstream uncertainty.

In its current form, the methodology presents several notable strengths. Since it is built upon the HSIC framework, it inherits most of its advantages, in particular the ability to handle any type of model output. In addition, the approach naturally extends to more complex computational chains involving more than two components. To further align the methodology with operational constraints, future work could focus on extending the theoretical developments to alternative experimental designs, such as Latin hypercube sampling (LHS).

Acknowledgments

The major part of this work was done when Oumar Baldé was a PhD student in CEA, DES, Service de Génie Logiciel pour la Simulation.

References

[1] T.J. Santner, B.J. Williams, W.I. Notz and B.J. Williams, *The Design and Analysis of Computer Experiments*, vol. 1. Springer (2003).

[2] K. Campbell, Statistical calibration of computer simulations. *Reliabil. Eng. Syst. Saf.* **91** (2006) 1358–1363.

[3] T.G. Trucano, L.P. Swiler, T. Igusa, W.L. Oberkampf and M. Pilch, Calibration, validation, and sensitivity analysis: What's what. *Reliabil. Eng. Syst. Saf.* **91** (2006) 1331–1357.

[4] E. de Rocquigny, N. Devictor and S. Tarantola, Uncertainty in Industrial Practice: A Guide to Quantitative Uncertainty Management. John Wiley & Sons (2008).

[5] A. Bouloré, Importance of uncertainty quantification in nuclear fuel behaviour modelling and simulation. *Nucl. Eng. Des.* **355** (2019) 110311.

[6] M. Gu and L. Wang, Scaled Gaussian stochastic process for computer model calibration and prediction. *SIAM/ASA J. Uncertainty Quantif.* **6** (2018) 1555–1583.

[7] R.K. Wong, C.B. Storlie and T.C. Lee, A frequentist approach to computer model calibration. *J. Roy. Statist. Soc. B Statist. Methodol.* **79** (2017) 635–648.

[8] M.C. Kennedy and A. O'Hagan, Bayesian calibration of computer models. *J. Roy. Statist. Soc. B (Statist. Methodol.)* **63** (2001) 425–464.

[9] X. Wu, T. Kozlowski, H. Meidani and K. Shirvan, Inverse uncertainty quantification using the modular Bayesian approach based on Gaussian process. Part 1. Theory. *Nucl. Eng. Des.* **335** (2018) 339–355.

[10] R.E. Kass and L. Wasserman, The selection of prior distributions by formal rules. *J. Am. Statist. Assoc.* **91** (1996) 1343–1370.

[11] S. Chib and E. Greenberg, Understanding the Metropolis–Hastings algorithm. *Am. Statist.* **49** (1995) 327–335.

[12] B. Currin, T. Mitchell, M. Morris and D. Ylvisaker, Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Am. Statist. Assoc.* **86** (1991) 953–963.

[13] J. Sacks, W.J. Welch, T.J. Mitchell and H.P. Wynn, Design and analysis of computer experiments. *Statist. Sci.* **4** (1989) 409–423.

[14] C. Andrieu and J. Thoms, A tutorial on adaptive MCMC. *Statist. Comput.* **18** (2008) 343–373.

[15] V.V. Dighe, M. Becker, T. Göçmen, B. Sanderse and J.-W. van Wingerden, Sensitivity analysis and Bayesian calibration of a dynamic wind farm control model: FLORIDyn, in *Journal of Physics: Conference Series*, vol. 2265. IOP Publishing (2022) 022062.

[16] J. Gou, C. Miao, Q. Duan, Q. Tang, Z. Di, W. Liao, J. Wu and R. Zhou, Sensitivity analysis-based automatic parameter calibration of the VIC model for streamflow simulations over China. *Water Resources Res.* **56** (2020) e2019WR025968.

[17] J.B. Nagel, J. Rieckermann and B. Sudret, Principal component analysis and sparse polynomial chaos expansions for global sensitivity analysis and model calibration: application to urban drainage simulation. *Reliabil. Eng. Syst. Saf.* **195** (2020) 106737.

[18] G. Perret, D. Wicaksono, I.D. Clifford and H. Ferroukhi, Global sensitivity analysis and Bayesian calibration on a series of reflood experiments with varying boundary conditions. *Nucl. Technol.* **208** (2022) 711–722.

[19] D.M. Hamby, A review of techniques for parameter sensitivity analysis of environmental models. *Environ. Monitor. Assessment* **32** (1994) 135–154.

[20] M.D. Morris, Factorial sampling plans for preliminary computational experiments. *Technometrics* **33** (1991) 161–174.

[21] J. Morio, Global and local sensitivity analysis methods for a physical system. *Eur. J. Phys.* **32** (2011) 1577.

[22] A. Saltelli, Making best use of model evaluations to compute sensitivity indices. *Comput. Phys. Commun.* **145** (2002) 280–297.

[23] I.M. Sobol, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* **55** (2001) 271–280.

[24] S. Kucherenko, M. Rodriguez-Fernandez, C. Pantelides and N. Shah, Monte Carlo evaluation of derivative-based global sensitivity measures. *Reliabil. Eng. Syst. Saf.* **94** (2009) 1135–1148.

[25] L. Clouvel, B. Iooss, V. Chabridon, M.I. Idrissi and F. Robin, An overview of variance-based importance measures in the linear regression context: comparative analyses and numerical tests. *Socioenviron. Syst. Model.* **7** (2025) 18681–18681.

[26] B. Iooss and C. Prieur, Shapley effects for sensitivity analysis with correlated inputs: Comparisons with Sobol' indices, numerical estimation and applications. *Int. J. Uncertain. Quantif.* **9** (2019) 493–514.

[27] A.B. Owen and C. Prieur, On Shapley value for measuring importance of dependent inputs. *SIAM/ASA J. Uncertain. Quantif.* **5** (2017) 986–1002.

[28] S. Da Veiga, Global sensitivity analysis with dependence measures. *J. Statist. Computat. Simul.* **85** (2015) 1283–1305.

[29] M. De Lozzo and A. Marrel, New improvements in the use of dependence measures for sensitivity analysis and screening. *J. Statist. Computat. Simul.* **86** (2016) 3038–3058.

[30] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf and A. Smola, A kernel statistical test of independence. *Adv. Neural Inform. Process. Syst.* **20** (2007) 585–592.

[31] C. Introïni, I. Ramière, J. Sercombe, B. Michel, T. Helfer and J. Fauque, ALCYONE: the fuel performance code of the PLEIADES platform dedicated to PWR fuel rods behavior. *Ann. Nucl. Energy* **207** (2024) 110711.

[32] A. Bouloré, C. Struzik, V. Bouineau, F. Gaudier, G. Damblin and S. Bernaud, Modelling of UO2 thermal conductivity: improvement of the irradiation defects contribution and uncertainty quantification. *Nucl. Eng. Des.* **407** (2023) 112304.

[33] G. Damblin, P. Barbillon, M. Keller, A. Pasanisi and É. Parent, Adaptive numerical designs for the calibration of computer codes. *SIAM/ASA J. Uncertain. Quantif.* **6** (2018) 151–179.

[34] C.P. Robert, G. Casella and G. Casella, *Monte Carlo Statistical Methods*, vol. 2. Springer (1999).

[35] G. Damblin and P. Gaillard, Bayesian inference and non-linear extensions of the CIRCE method for quantifying the uncertainty of closure relationships integrated into thermal-hydraulic system codes. *Nucl. Eng. Des.* **359** (2020) 110391.

[36] S. Barde, Bayesian estimation of large-scale simulation models with Gaussian process regression surrogates. *Computat. Statist. Data Anal.* **196** (2024) 107972.

[37] M. Van Oijen, D. Cameron, K. Butterbach-Bahl, N. Farahbakhshazad, P.-E. Jansson, R. Kiese, K.H. Rahn, C. Werner and J. Yeluripati. A Bayesian framework for model calibration, comparison and analysis: application to four models for the biogeochemistry of a Norway spruce forest. *Agric. For. Meteorol.* **151** (2011) 1609–1621.

[38] X. Wu, T. Mui, G. Hu, H. Meidani and T. Kozlowski, Inverse uncertainty quantification of TRACE physical model parameters using sparse gird stochastic collocation surrogate model. *Nucl. Eng. Des.* **319** (2017) 185–200.

[39] A. Marrel, B. Iooss and V. Chabridon, The ICSCREAM methodology: identification of penalizing configurations in computer experiments using screening and metamodel: applications in thermal-hydraulics. *Nucl. Sci. Eng.* **196** (2022) 301–321.

[40] A. Marrel and B. Iooss, Probabilistic surrogate modeling by Gaussian process: a new estimation algorithm for more robust prediction. *Reliabil. Eng. Syst. Saf.* **247** (2024) 110120.

[41] B. Zhou, Z. Shi, S. Kucherenko and H. Zhao, A unified approach for global sensitivity analysis based on active subspace and Kriging. *Reliabil. Eng. Syst. Saf.* **217** (2022) 108080.

[42] M. Binois and N. Wycoff, A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. *ACM Trans. Evol. Learn. Optim.* **2** (2022) 1–26.

[43] H. Haario, E. Saksman and J. Tamminen, An adaptive Metropolis algorithm. *Bernoulli* **7** (2001) 223–242.

[44] F. Liu, M. Bayarri and J. Berger, Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4** (2009) 119–150.

[45] X. Wu, K. Shirvan and T. Kozlowski, Demonstration of the relationship between sensitivity and identifiability for inverse uncertainty quantification. *J. Computat. Phys.* **396** (2019) 2–30.

[46] X. Xu, C. Sun, G. Huang and B.P. Mohanty, Global sensitivity analysis and calibration of parameters for a physically-based agro-hydrological model. *Environ. Model. Softw.* **83** (2016) 88–102.

[47] M. Zambrano-Bigiarini, Z. Zając and S. Tarantola, Global sensitivity analysis for the calibration of a fully-distributed hydrological model, in *7th International Conference on Sensitivity Analysis of Model Output, MASCOT-SAMO* (2013).

[48] A.B. Owen, Sobol'indices and Shapley value. *SIAM/ASA J. Uncertain. Quantif.* **2** (2014) 245–251.

[49] M. Hérin, M. Il Idrissi, V. Chabridon and B. Iooss, Proportional marginal effects for global sensitivity analysis. *SIAM/ASA J. Uncertain. Quantif.* **12** (2024) 667–692.

[50] F. Gamboa, A. Janon, T. Klein and A. Lagnoux, Sensitivity analysis for multidimensional and functional outputs. *Electron. J. Statist.* **8** (2014) 575–603.

[51] F. Gamboa, T. Klein, A. Lagnoux and L. Moreno, Sensitivity analysis in general metric spaces. *Reliabil. Eng. Syst. Saf.* **212** (2021) 107611.

[52] A. Janon, T. Klein, A. Lagnoux, M. Nodet and C. Prieur, Asymptotic normality and efficiency of two Sobol index estimators. *ESAIM Probab. Statist.* **18** (2014) 342–364.

[53] A. Marrel, B. Iooss, B. Laurent and O. Roustant, Calculations of Sobol' indices for the Gaussian process metamodel. *Reliabil. Eng. Syst. Saf.* **94** (2009) 742–751.

[54] S. Da Veiga, F. Gamboa, A. Lagnoux, T. Klein and C. Prieur, New estimation of Sobol' indices using kernels (2023). https://arxiv.org/pdf/2303.17832.

[55] E. Borgonovo, A new uncertainty importance measure. *Reliabil. Eng. Syst. Saf.* **92** (2007) 771–784.

[56] S. Rahman, The f-sensitivity index. *SIAM/ASA J. Uncertain. Quantif.* **4** (2016) 130–162.

[57] M.R. El Amri and A. Marrel, Optimized HSIC-based tests for sensitivity analysis: application to thermalhydraulic simulation of accidental scenario on nuclear reactor. *Qual. Reliabil. Eng. Int.* **38** (2022) 1386–1403.

[58] M.R. El Amri and A. Marrel, More powerful HSIC-based independence tests, extension to space-filling designs and functional data. *Int. J. Uncertain. Quantif.* **14** (2024) 69–98.

[59] G.-K. Delipei, J. Garnier, J. Le Pallec and B. Normand, Uncertainty analysis methodology for multi-physics coupled rod ejection accident, in *International Conference on Mathematics and Computational Methods Applied to Nuclear Science and Engineering (M&C 2019)* (2019).

[60] N. Marie, S. Li, A. Marrel, M. Marquès, S. Bajard, A. Tosello, J. Perez, B. Grosjean, A. Gerschenfeld, M. Anderhuber *et al.*, VVUQ of a thermal-hydraulic multi-scale tool on unprotected loss of flow accident in SFR reactor. *EPJ N Nucl. Sci. Technol.* **7** (2021) 3.

[61] Á. Rollón de Pinedo, M. Couplet, B. Iooss, N. Marie, A. Marrel, E. Merle and R. Sueur, Functional outlier detection by means of h-mode depth and dynamic time warping. *Appl. Sci.* **11** (2021) 11475.

[62] T. Wang, X. Dai and Y. Liu,. Learning with Hilbert–Schmidt independence criterion: A review and new perspectives. *Knowl. Based Syst.* **234** (2021) 107567.

[63] T. Wang, Z. Hu and H. Liu, A unified view of feature selection based on Hilbert–Schmidt independence criterion. *Chemometrics Intell. Lab. Syst.* **236** (2023) 104807.

[64] T. Wang, J. Lu and G. Zhang, Two-stage fuzzy multiple kernel learning based on Hilbert–Schmidt independence criterion. *IEEE Trans. Fuzzy Syst.* **26** (2018) 3703–3714.

[65] A. Gretton, O. Bousquet, A. Smola and B. Schölkopf, Measuring statistical dependence with Hilbert–Schmidt norms, in *International Conference on Algorithmic Learning Theory*. Springer (2005) 63–77.

[66] S. Bernaud, I. Ramiere, G. Latu and B. Michel, PLEIADES: A numerical framework dedicated to the multiphysics and multiscale nuclear fuel behavior simulation. *Ann. Nucl. Energy* **205** (2024) 110577.

[67] B. Iooss and A. Marrel, Advanced methodology for uncertainty propagation in computer experiments with large number of inputs. *Nucl. Technol.* **205** (2019) 1588–1606.

[68] C. Sriperumbudur, K. Fukumizu and G. Lanckriet, On the relation between universality, characteristic kernels and RKHS embedding of measures, in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings (2010) 773–780.

[69] J. Ziegel, D. Ginsbourger and L. Dümbgen, Characteristic kernels on Hilbert spaces, Banach spaces, and on sets of measures. *Bernoulli* **30** (2024) 1441–1457.

[70] O. Baldé, G. Damblin, A. Marrel, A. Bouloré and L. Giraldi, Nonparametric Bayesian approach for quantifying the conditional uncertainty of input parameters in chained numerical models (2023). https://arxiv.org/pdf/2307.01111.

[71] N. Fellmann, C. Blanchet-Scalliet, C. Helbert, A. Spagnol and D. Sinoquet, Kernel-based sensitivity analysis for (excursion) sets. *Technometrics* **66** (2024) 575–587.

[72] L. Song, A. Smola, A. Gretton, J. Bedo and K. Borgwardt, Feature selection via dependence maximization. *J. Mach. Learn. Res.* **13** (2012) 1393–1434.

[73] B.K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf and G.R. Lanckriet, Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.* **11** (2010) 1517–1561.

[74] Z. Szabó and B.K. Sriperumbudur, Characteristic and universal tensor product kernels. *J. Mach. Learn. Res.* **18** (2017) 1–29.

[75] R. Serfling, Approximation Theorems of Mathematical Statistics. John Wiley & Sons (1980).
[76] K. Zhang, J. Peters, D. Janzing and B. Schölkopf, Kernel-based conditional independence test and application in causal Discovery, in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence* (2011) 804–813.
[77] F. Kazi-Aoual, S. Hitier, R. Sabatier and J.-D. Lebreton, Refined approximations to permutation tests for multivariate inference. *Computat. Statist. Data Anal.* **20** (1995) 643–656.
[78] M. Yamada, W. Jitkrittum, L. Sigal, E.P. Xing and M. Sugiyama, High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computat.* **26** (2014) 185–207.
[79] M. Kanagawa, P. Hennig, D. Sejdinovic and B.K. Sriperumbudur, Gaussian processes and kernel methods: a review on connections and Equivalences (2018). https://arxiv.org/pdf/1807.02582.
[80] A. Gretton and L. Györfi, Consistent nonparametric tests of independence. *J. Mach. Learn. Res.* **11** (2010) 1391–1423.

Appendix A. Preliminaries on characteristic kernels

A.1 Characteristic kernels

Throughout this section, $\mathcal{Z}$ is a separable metric space. The set of all Borel probability measures on $\mathcal{Z}$ is denoted by $\mathcal{M}_1^+(\mathcal{Z})$.

Definition A.1 (Kernel mean embedding)**.** Let $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a kernel. For any probability distribution $\nu \in \mathcal{M}_1^+(\mathcal{Z})$ such that $\mathbb{E}_\nu\left[\sqrt{K(Z,Z)}\right] < \infty$, the *kernel mean embedding* of ν is the function defined as:

$$\begin{array}{rccl} \mu_\nu : & \mathcal{Z} & \longrightarrow & \mathbb{R} \\ & z & \longmapsto & \mathbb{E}_\nu[K(z,Z)] = \displaystyle\int_{\mathcal{Z}} K\left(z,z'\right) \mathrm{d}\nu(z') \ . \end{array} \tag{A.1}$$

In particular, μ_ν belongs to the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ induced by K.

Definition A.2 (Characteristic kernel)**.** Let $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a bounded kernel. This assumption ensures that μ_ν is well-defined for any probability measure $\nu \in \mathcal{M}_1^+(\mathcal{Z})$. The kernel K is said to be characteristic on $\mathcal{Z}$ if the mapping $\nu \in \mathcal{M}_1^+(\mathcal{Z}) \mapsto \mu_\nu \in \mathcal{H}$ is injective. This means that:

$$\forall\, \nu_1, \nu_2 \in \mathcal{M}_1^+(\mathcal{Z}), \quad \mu_{\nu_1} = \mu_{\nu_2} \implies \nu_1 = \nu_2 \ . \tag{A.2}$$

Definition A.3 (Maximum mean discrepancy)**.** Let $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a bounded kernel, and let $\mathcal{H}$ denote the associated RKHS. For any probability measures $\nu_1, \nu_2 \in \mathcal{M}_1^+(\mathcal{Z})$, the maximum mean discrepancy (MMD) is defined by:

$$\mathrm{MMD}\left(\nu_1, \nu_2\right) := \|\mu_{\nu_1} - \mu_{\nu_2}\|_{\mathcal{H}} \ . \tag{A.3}$$

If K is characteristic on $\mathcal{Z}$, the MMD induces a proper metric on $\mathcal{M}_1^+(\mathcal{Z})$, that is:

$$\mathrm{MMD}(\nu_1, \nu_2) = 0 \iff \mu_{\nu_1} = \mu_{\nu_2} \iff \nu_1 = \nu_2 \,. \tag{A.4}$$

Definition A.4 (Integrally strictly positive definite kernel)**.** Let $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a bounded kernel. It is said to be integrally strictly positive definite on $\mathcal{Z}$ if:

$$\forall\, \nu \in \mathcal{M}_1^+(\mathcal{Z}), \quad \int_{\mathcal{Z}} \int_{\mathcal{Z}} K(z, z')\, \mathrm{d}\nu(z)\, \mathrm{d}\nu(z') > 0 \,. \tag{A.5}$$

Theorem A.5 (Theorem 7 in [73])**.** *A bounded kernel $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is characteristic on $\mathcal{Z}$ if and only if it is integrally strictly positive definite.*

This result establishes that being integrally strictly positive definite is a sufficient condition for being characteristic. As shown in [69], a specific class of kernels is integrally strictly positive definite, and therefore characteristic.

Proposition A.6 (Proposition 5.2 in [69])**.** *Let $\mathcal{F}$ be a separable Hilbert space. Let $T : \mathcal{Z} \to \mathcal{F}$ and $\varphi : [0, \infty) \to \mathbb{R}$ be two measurable maps. Assume that T is injective and that the function:*

$$\forall\, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^q, \quad k_\varphi(\boldsymbol{x}, \boldsymbol{x}') = \varphi\left(\|\boldsymbol{x} - \boldsymbol{x}'\|_{\mathbb{R}^q}^2\right) \tag{A.6}$$

defines a characteristic kernel on $\mathbb{R}^q$. Then, the function:

$$\forall\, z, z' \in \mathcal{Z}, \quad K(z, z') = \varphi\left(\|T(z) - T(z')\|_{\mathcal{F}}^2\right) \tag{A.7}$$

defines a characteristic kernel on $\mathcal{Z}$.

A.2 Translation-invariant kernels

Definition A.7 (Translation-invariant function)**.** A function $K : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ is said to be translation-invariant if there exists a shift function $\psi : \mathbb{R}^q \to \mathbb{R}$ such that:

$$\forall\, \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^q, \quad K(\boldsymbol{x}, \boldsymbol{x}') = \psi(\boldsymbol{x} - \boldsymbol{x}') \,. \tag{A.8}$$

Assumption A.8. Let $K : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ be a translation-invariant function. It is assumed that its shift function $\psi : \mathbb{R}^q \to \mathbb{R}$ is continuous and integrable on $\mathbb{R}^q$:

$$\psi \in C^0(\mathbb{R}^q) \cap L^1(\mathbb{R}^q) \,. \tag{A.9}$$

Note that Assumption A.8 is not stringent, as it is verified by most kernels used in practice (especially Gaussian and Matérn kernels).

Theorem A.9 (Bochner)**.** *Let $K : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ be a translation-invariant function satisfying Assumption A.8. Then, K is a kernel if and only if the Fourier transform (FT) of ψ, defined by:*

$$\forall\, \boldsymbol{w} \in \mathbb{R}^q, \quad \widehat{\psi}(\boldsymbol{w}) = \frac{1}{\left(\sqrt{2\pi}\right)^q} \int_{\mathbb{R}^q} \mathrm{e}^{-i\boldsymbol{w}^t\boldsymbol{x}}\, \psi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \,, \tag{A.10}$$

is a real-valued non-negative function. Moreover, when K is a kernel, the associated shift function ψ is more commonly called the signature of K.

Proposition A.10 (Proposition 8 in [68])**.** *Let $K : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ be a translation-invariant kernel satisfying Assumption A.8. Then, K is characteristic on $\mathbb{R}^q$ if and only if the FT of ψ is positive on $\mathbb{R}^q$, which can be equivalently*

expressed as:

$$\operatorname{supp}\left(\widehat{\psi}\right) := \overline{\left\{\boldsymbol{w} \in \mathbb{R}^q : \widehat{\psi}(\boldsymbol{w}) > 0\right\}} = \mathbb{R}^q \ . \tag{A.11}$$

A.3 Tensor product of univariate kernels

In this section, the objective is to construct a characteristic kernel on $\mathbb{R}^q$ that could be used to compute the HSIC on a vector output (see Approach B in Sect. 3.2). Let $\{K_i\}_{i=1}^q$ be a set of q univariate kernels. The *tensor-product kernel* is defined by:

$$\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^q, \quad K_{\text{tens}}\left(\boldsymbol{x}, \boldsymbol{x}'\right) := \left[\bigotimes_{i=1}^q K_i\right]\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \prod_{i=1}^q K_i\left(x_i, x_i'\right). \tag{A.12}$$

The question is whether taking characteristic kernels K_i on $\mathbb{R}$ is sufficient to obtain a characteristic kernel K_{tens} on $\mathbb{R}^q$. In general, the answer is no. Indeed, a counterexample for $q = 3$ may be found in [74]. However, under additional assumptions on the kernels K_i, it is possible to show that K_{tens} remains characteristic.

We first establish that the result holds when tensorizing Gaussian kernels (see Prop. A.11). The proof follows almost immediately from Proposition A.6. We then extend the result to the case where all kernels K_i are translation-invariant (see Prop. A.12). In this case, the proof consists in computing the FT of the signature ψ_{tens} and showing that its support is $\mathbb{R}^q$.

Proposition A.11. *Assume that the kernels $\{K_i\}_{i=1}^q$ are all Gaussian kernels:*

$$\forall\, x_i, x_i' \in \mathbb{R}, \quad K_i\left(x_i, x_i'\right) = \exp\left[-\frac{1}{2}\left(\frac{x_i - x_i'}{\gamma_i}\right)^2\right]. \tag{A.13}$$

Then, the kernel K_{tens} is characteristic on $\mathbb{R}^q$.

Proof. If taking Gaussian kernels, equation (A.12) becomes:

$$K_{\text{tens}}\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \prod_{i=1}^q \exp\left[-\frac{1}{2}\left(\frac{x_i - x_i'}{\gamma_i}\right)^2\right] = \exp\left[-\frac{1}{2}\sum_{i=1}^q \left(\frac{x_i - x_i'}{\gamma_i}\right)^2\right] \tag{A.14}$$

$$= \varphi\left(\|T(\boldsymbol{x}) - T(\boldsymbol{x}')\|^2\right) \ , \tag{A.15}$$

where the last equation is obtained by identification with:

- $\varphi : z \mapsto \exp(-|z|)$;
- $T : \boldsymbol{x} = (x_1, \ldots, x_d) \mapsto (x_1/\gamma_1, \ldots, x_d/\gamma_d)/\sqrt{2}$.

On the one hand, T is a homothety and is therefore injective. On the other hand, with this choice of φ, the kernel k_φ from equation (A.6) is a multivariate Laplacian kernel, which is known to be characteristic on $\mathbb{R}^q$. Since all assumptions of Proposition A.6 are satisfied, one can conclude that K_{tens} is integrally strictly positive definite, hence characteristic on $\mathbb{R}^q$. □

Proposition A.12. *Assume that each kernel K_i is translation-invariant, satisfies Assumption A.8, and is characteristic on $\mathbb{R}$. Then, the kernel K_{tens} is characteristic on $\mathbb{R}^q$.*

Proof. First, observe that K_{tens} is a translation-invariant kernel on $\mathbb{R}^q$:

$$K_{\text{tens}}\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \prod_{i=1}^q K_i(x_i, x_i') = \prod_{i=1}^q \psi_i(x_i - x_i') = \psi_{\text{tens}}\left(\boldsymbol{x} - \boldsymbol{x}'\right). \tag{A.16}$$

Its signature ψ_{tens} is given by the tensor product of the signatures ψ_i of the kernels K_i:

$$\forall\, \boldsymbol{r} \in \mathbb{R}^q, \quad \psi_{\text{tens}}(\boldsymbol{r}) := \left[\bigotimes_{i=1}^{q} \psi_i\right](\boldsymbol{r}) = \prod_{i=1}^{q} \psi_i(r_i)\ . \tag{A.17}$$

According to Proposition A.10, in order to prove that K_{tens} is characteristic on $\mathbb{R}^q$, it is sufficient to prove that the FT of ψ_{tens} is positive on $\mathbb{R}^q$. A direct consequence of the Fubini-Tonelli theorem is the separability of the FT:

$$\begin{aligned} \widehat{\psi}_{\text{tens}}(\boldsymbol{w}) &= \frac{1}{\left(\sqrt{2\pi}\right)^q} \int_{\mathbb{R}^q} \mathrm{e}^{-i\boldsymbol{w}^t\boldsymbol{x}}\, \psi_{\text{tens}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ &= \prod_{i=1}^{q} \left(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{-iw_i x_i}\, \psi_i(w_i)\, \mathrm{d}w_i \right) = \prod_{i=1}^{q} \widehat{\psi}_i(w_i)\ . \end{aligned} \tag{A.18}$$

As all kernels K_i are characteristic on $\mathbb{R}$, the FTs $\widehat{\psi}_i$ are positive on $\mathbb{R}$. It follows that $\widehat{\psi}$ is positive on $\mathbb{R}^q$, which completes the proof by Proposition A.10. □

Appendix B. Technical results on U- and V-statistics

For $s \geq 2$, let $Z_1, \ldots, Z_s$ be independent and identically distributed (i.i.d.) random variables with common distribution $\mathbb{P}_Z \in \mathcal{M}_1^+(\mathcal{Z})$. Let $\eta : \mathcal{Z}^s \to \mathbb{R}$ be a measurable function. Throughout Appendix B, the target quantity is the following expectation:

$$\zeta := \mathbb{E}_{\mathbb{P}_Z}\left[\eta\left(Z_1, \ldots, Z_s\right)\right] = \int_{\mathcal{Z}} \ldots \int_{\mathcal{Z}} \eta\left(z_1, \ldots, z_s\right)\, \mathrm{d}\mathbb{P}_Z(z_1) \ldots \mathrm{d}\mathbb{P}_Z(z_s)\ . \tag{B.1}$$

In this multivariate setting, the function η is said to be *symmetric* if it is invariant under any permutation of its arguments:

$$\forall\, \sigma \in \mathbb{S}_s, \quad \forall\, \boldsymbol{z} \in \mathcal{Z}^s, \quad \eta\left(z_{\sigma(1)}, \ldots, z_{\sigma(s)}\right) = \eta\left(z_1, \ldots, z_s\right), \tag{B.2}$$

where $\mathbb{S}_s$ denotes the set of all permutations of $\{1, \ldots, s\}$. If the function η is not symmetric, it must be replaced by its symmetrized version $\widetilde{\eta}$, defined by:

$$\forall\, \boldsymbol{z} \in \mathcal{Z}^s, \quad \widetilde{\eta}\left(z_1, \ldots, z_s\right) := \frac{1}{s!} \sum_{\sigma \in \mathbb{S}_s} \eta\left(z_{\sigma(1)}, \ldots, z_{\sigma(s)}\right). \tag{B.3}$$

By construction, $\widetilde{\eta}$ is symmetric. Note that $\eta = \widetilde{\eta}$ if and only if η is already symmetric. Moreover, replacing η with $\widetilde{\eta}$ leaves the target quantity unchanged:

$$\mathbb{E}_{\mathbb{P}_Z}\left[\widetilde{\eta}\left(Z_1, \ldots, Z_s\right)\right] = \mathbb{E}_{\mathbb{P}_Z}\left[\eta\left(Z_1, \ldots, Z_s\right)\right] = \zeta\ , \tag{B.4}$$

because the vectors $(Z_1, \ldots, Z_s)$ and $\left(Z_{\sigma(1)}, \ldots, Z_{\sigma(s)}\right)$ have the same joint distribution for any permutation $\sigma \in \mathbb{S}_s$.

Definition B.1 (U-statistic)**.** Let $\left(Z^{(1)}, \ldots, Z^{(n)}\right)$ be an i.i.d. sample from the distribution $\mathbb{P}_Z$, with $n \geq s$. The U-statistic estimator of ζ is constructed as follows:

$$\widehat{\zeta}^{\mathrm{U}} = \binom{n}{s}^{-1} \sum_{\boldsymbol{i} \in \mathcal{A}_n^s} \widetilde{\eta}\left(Z^{(i_1)}, \ldots, Z^{(i_s)}\right)\ , \tag{B.5}$$

where the summation set is:

$$\mathcal{A}_n^s = \{\boldsymbol{i} := (i_1, \ldots, i_s) : 1 \leq i_1 < \ldots < i_s \leq n\} \ \text{ with } \ \mathrm{Card}\left(\mathcal{A}_n^s\right) = \binom{n}{s} = \frac{n!}{s!\,(n-s)!}\ . \tag{B.6}$$

Definition B.2 (V-statistic)**.** Let $\left(Z^{(1)}, \ldots, Z^{(n)}\right)$ be an i.i.d. sample from the distribution $\mathbb{P}_Z$, with $n \geq s$. The V-statistic estimator of ζ is constructed as follows:

$$\widehat{\zeta}^{\mathrm{V}} = \frac{1}{n^s} \sum_{\boldsymbol{i} \in \mathcal{B}_n^s} \widetilde{\eta}\left(Z^{(i_1)}, \ldots, Z^{(i_s)}\right) , \tag{B.7}$$

where the summation set is:

$$\mathcal{B}_n^s = \{\boldsymbol{i} := (i_1, \ldots, i_s) : 1 \leq i_1, \ldots, i_s \leq n\} \quad \text{with} \quad \mathrm{Card}\left(\mathcal{B}_n^s\right) = n^s . \tag{B.8}$$

It is straightforward to see that $\widehat{\zeta}^{\mathrm{U}}$ is unbiased. The constraint $\boldsymbol{i} \in \mathcal{A}_n^s$ on the multi-index $\boldsymbol{i} = (i_1, i_2, \ldots, i_s)$ ensures that $i_1 < i_2 < \ldots < i_s$, so that each term $\widetilde{\eta}\left(Z^{(i_1)}, \ldots, Z^{(i_s)}\right)$ has expectation exactly equal to ζ. On the contrary, $\widehat{\zeta}^{\mathrm{V}}$ is biased because its summation set $\mathcal{B}_n^s$ includes some multi-indices $\boldsymbol{i}$ for which $i_a = i_b$ with $(a, b) \in \{1, \ldots, s\}^2$ and $a \neq b$.

$\widehat{\zeta}^{\mathrm{U}}$ is the unbiased estimator of ζ with the smallest variance. The following lemma provides a simple upper bound for its variance.

Lemma B.3 (Lemma A of Sect. 5.2.1 in [75])**.** *Assume that* $\mathbb{E}_{\mathbb{P}_Z}\left[|\widetilde{\eta}\left(Z_1, \ldots, Z_s\right)|^2\right] < \infty$. *Then, the variance of* $\widehat{\zeta}^{\mathrm{U}}$ *satisfies:*

$$0 \leq \mathbb{V}\left(\widehat{\zeta}^{\mathrm{U}}\right) \leq \frac{s}{n} \mathbb{V}_{\mathbb{P}_Z}\left(\widetilde{\eta}\left(Z_1, \ldots, Z_s\right)\right) . \tag{B.9}$$

An immediate consequence of this lemma is that the variance of $\widehat{\zeta}^{\mathrm{U}}$ vanishes as $n \to \infty$.

Asymptotically, $\widehat{\zeta}^{\mathrm{U}}$ and $\widehat{\zeta}^{\mathrm{V}}$ behave similarly. This is formalized by the following lemma, which provides convergence rates for the moments of their difference.

Lemma B.4 (Lemma of Sect. 5.7.3 in [75])**.** *Let* $r \geq 1$ *be fixed. Assume that:*

$$\forall\, 1 \leq i_1, \ldots, i_s \leq s, \quad \mathbb{E}_{\mathbb{P}_Z}\left[\, |\widetilde{\eta}\left(Z_{i_1}, \ldots, Z_{i_s}\right)|^r \,\right] < \infty . \tag{B.10}$$

Then, one has:

$$\mathbb{E}_{\mathbb{P}_Z}\left[\left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{V}}\right|^r\right] = \mathcal{O}\left(\frac{1}{n^r}\right) . \tag{B.11}$$

In this work, this Big-O convergence rate is insufficient. Instead, we require a moment bound with an explicit constant. From this perspective, Lemma B.5 fills the gap left by Lemma B.4.

Lemma B.5. *Assuming that the conditions of Lemma B.4 hold, one has:*

$$\mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{V}}\right|^r\right] \leq C_r \left(\frac{s(s-1)}{n}\right)^r \tag{B.12}$$

with:

$$C_r = C_r\left(\widetilde{\eta}, \mathbb{P}_Z\right) := \max_{1 \leq i_1, \ldots, i_s \leq s} \mathbb{E}\left[\, |\widetilde{\eta}\left(Z_{i_1}, \ldots, Z_{i_s}\right)|^r \,\right] . \tag{B.13}$$

The book by Serfling [75] only provides a sketch of the proof of Lemma B.4. Following the line of argument suggested there, one can recover Lemma B.5. For completeness, we include below a detailed and self-contained proof of Lemma B.5.

Proof. For technical reasons, we introduce the following two index sets:

$$\mathcal{E}_n^s := \{\boldsymbol{i} = (i_1, \ldots, i_s) : 1 \leq i_1 \neq \ldots \neq i_s \leq n\} \tag{B.14}$$

$$\mathcal{F}_n^s := \left\{ \boldsymbol{i} = (i_1, \ldots, i_s) : \exists\ (a,b) \in \{1,\ldots,s\}^2 \text{ such that } i_a = i_b \text{ and } a \neq b \right\}. \tag{B.15}$$

The elements of $\mathcal{E}_n^s$ are obtained by applying all permutations in $\mathbb{S}_s$ to each element of $\mathcal{A}_n^s$. The set $\mathcal{F}_n^s$ is the complement of $\mathcal{E}_n^s$ in $\mathcal{B}_n^s$. Consequently, their cardinalities are given by:

$$\text{Card}\left(\mathcal{E}_n^s\right) = s!\,\text{Card}\left(\mathcal{A}_n^s\right) = s!\binom{n}{s} = \frac{n!}{(n-s)!} = n(n-1)\ldots(n-s+1) =: (n)_s \tag{B.16}$$

$$\text{Card}\left(\mathcal{F}_n^s\right) = \text{Card}\left(\mathcal{B}_n^s\right) - \text{Card}\left(\mathcal{E}_n^s\right) = n^s - (n)_s\ . \tag{B.17}$$

The key element of the proof is the statistic:

$$\widehat{\zeta}^{\mathrm{W}} := \frac{1}{n^s - (n)_s} \sum_{\boldsymbol{i} \in \mathcal{F}_n^s} \widetilde{\eta}\left(Z^{(i_1)}, \ldots, Z^{(i_s)}\right)\ , \tag{B.18}$$

which is defined analogously to $\widehat{\zeta}^{\mathrm{U}}$ and $\widehat{\zeta}^{\mathrm{V}}$, except that the averaging is performed over the index set $\mathcal{F}_n^s$ (instead of $\mathcal{A}_n^s$ and $\mathcal{B}_n^s$). From now, in order to simplify the notation throughout the proof, we write $\widetilde{\eta}_{\boldsymbol{i}}$ for the terms appearing in $\widehat{\zeta}^{\mathrm{U}}$, $\widehat{\zeta}^{\mathrm{V}}$, and $\widehat{\zeta}^{\mathrm{W}}$.

To prove the inequality in equation (B.12), we proceed in several steps, as recommended in [75]. First, we show that:

$$n^s\left(\widehat{\xi}^{\mathrm{U}} - \widehat{\xi}^{\mathrm{V}}\right) = \left(n^s - (n)_s\right)\left(\widehat{\xi}^{\mathrm{U}} - \widehat{\xi}^{\mathrm{W}}\right). \tag{B.19}$$

Next, we establish that:

$$n^s - (n)_s = \mathcal{O}\left(n^{s-1}\right). \tag{B.20}$$

Combining equations (B.19) and (B.20), the desired inequality follows easily.

Let us see how to establish equation (B.19). From the definitions of the index sets, one has:

$$n^s\,\widehat{\zeta}^{\mathrm{V}} = \sum_{\boldsymbol{i} \in \mathcal{B}_n^s} \widetilde{\eta}_{\boldsymbol{i}} = \sum_{\boldsymbol{i} \in \mathcal{E}_n^s} \widetilde{\eta}_{\boldsymbol{i}} + \sum_{\boldsymbol{i} \in \mathcal{F}_n^s} \widetilde{\eta}_{\boldsymbol{i}} \tag{B.21}$$

$$= s! \sum_{\boldsymbol{i} \in \mathcal{A}_n^s} \widetilde{\eta}_{\boldsymbol{i}} + \sum_{\boldsymbol{i} \in \mathcal{F}_n^s} \widetilde{\eta}_{\boldsymbol{i}} \tag{B.22}$$

$$= (n)_s\,\widehat{\zeta}^{\mathrm{U}} + \left(n^s - (n)_s\right)\widehat{\zeta}^{\mathrm{W}}\ . \tag{B.23}$$

Equation (B.21) follows from the disjoint union $\mathcal{B}_n^s = \mathcal{E}_n^s \sqcup \mathcal{F}_n^s$. Equation (B.22) uses the symmetry of the function $\widetilde{\eta}$: under this assumption, summing over $\mathcal{E}_n^s$ is equivalent to summing over $\mathcal{A}_n^s$ and multiplying by $s! = \text{Card}(\mathbb{S}_s)$. Equation (B.23) directly stems from Definition B.1 and equation (B.18). Finally, equation (B.23) yields equation (B.19) after straightforward manipulations.

To demonstrate equation (B.20), we use induction on s. Initialization for $s = 1$ is trivial:

$$n^s - (n)_s = n - n = 0 = \mathcal{O}(1)\ .$$

Then, assume that equation (B.20) holds for some $s \geq 1$. It must be shown that it also holds at order $s+1$. One can write:

$$n^{s+1} - (n)_{s+1} = n^{(s+1)} - n(n-1)(n-2)\ldots(n-s) \tag{B.24}$$

$$= n\left[n^s - (n-1)(n-2)\ldots(n-s)\right] \tag{B.25}$$

$$= n\big[n^s - (n)_s\big] + n\big[(n)_s - (n-1)_s\big] \ . \tag{B.26}$$

On the one hand, by the induction hypothesis, there exists $\gamma_s > 0$ such that:

$$n^s - (n)_s \leq \gamma_s \, n^{s-1} \ . \tag{B.27}$$

On the other hand, straightforward algebra yields:

$$(n)_s - (n-1)_s = n(n-1)(n-2)\dots(n-s+1) - (n-1)(n-2)\dots(n-s) \tag{B.28}$$

$$= \big[(n-1)(n-2)\dots(n-s+1)\big]\big[n-(n-s)\big] \tag{B.29}$$

$$= (n-1)(n-2)\dots(n-s+1)\, s \tag{B.30}$$

$$\leq n^{s-1}\, s \ . \tag{B.31}$$

To obtain the last inequality, note that each of the $s-1$ factors in $(n-1)(n-2)\cdots(n-s+1)$ is bounded above by n. Then, by substituting equations (B.27) and (B.31) in equation (B.26), we obtain:

$$n^{s+1} - (n)_{s+1} \leq (s+\gamma_s)\, n^s = \gamma_{s+1}\, n^s \quad \text{with} \quad \gamma_{s+1} = s + \gamma_s. \tag{B.32}$$

This shows that equation (B.20) holds at order $s+1$, and therefore completes the induction. Moreover, equation (B.32) allows the constants γ_s to be determined explicitly. From the initialization step, one has $\gamma_1 = 0$. The recursion $\gamma_{s+1} = s + \gamma_s$ then gives:

$$\gamma_s = (s-1) + \gamma_{s-1} = (s-1) + (s-2) + \ldots + 2 + 1 + \gamma_1 = \sum_{k=1}^{s-1} k = \frac{s\,(s-1)}{2} \ . \tag{B.33}$$

Consequently, equation (B.20) can be written explicitly as:

$$n^s - (n)_s \leq \left(\frac{s(s-1)}{2}\right) n^{s-1} \ . \tag{B.34}$$

Now that equations (B.19) and (B.20) have been established, we can derive equation (B.12).

In Lemma B.4, for the chosen exponent $r \geq 1$, it was assumed that:

$$\forall\, 1 \leq i_1, \dots, i_s \leq s, \quad \mathbb{E}_{\mathbb{P}_Z}\big[\,|\widetilde{\eta}\,(Z_{i_1}, \dots, Z_{i_s})|^r\,\big] < \infty \ . \tag{B.35}$$

This assumption ensures finiteness of the constant:

$$C_r = \max_{1 \leq i_1, \dots, i_s \leq s} \mathbb{E}_{\mathbb{P}_Z}\big[\,|\widetilde{\eta}\,(Z_{i_1}, \dots, Z_{i_s})|^r\,\big] \ . \tag{B.36}$$

In the context of Lemma B.5, C_r is a constant with respect to the given i.i.d. sample, but it depends on $\widetilde{\eta}$ and $\mathbb{P}_Z$.

Combining equations (B.19) and (B.34) yields the following inequalities:

$$\left(\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{V}}\right) \leq \left(\frac{\gamma_s}{n}\right)\left(\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{W}}\right) \tag{B.37}$$

$$\left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{V}}\right|^r \leq \left(\frac{\gamma_s}{n}\right)^r \left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{W}}\right|^r \tag{B.38}$$

$$\mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{V}}\right|^r\right] \leq \left(\frac{\gamma_s}{n}\right)^r \mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{W}}\right|^r\right] . \tag{B.39}$$

Using the triangle inequality in the space $\mathbb{L}^r(\Omega)$ of all random variables with finite r-th moment, we obtain:

$$\left(\mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{W}}\right|^r\right]\right)^{1/r} = \left\|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{W}}\right\|_{\mathbb{L}^r} \leq \left\|\widehat{\zeta}^{\mathrm{U}}\right\|_{\mathbb{L}^r} + \left\|\widehat{\zeta}^{\mathrm{W}}\right\|_{\mathbb{L}^r} . \tag{B.40}$$

Raising both sides of the inequality to the power r results in:

$$\left\|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{W}}\right\|_{\mathbb{L}^r}^r \leq \left(\left\|\widehat{\zeta}^{\mathrm{U}}\right\|_{\mathbb{L}^r} + \left\|\widehat{\zeta}^{\mathrm{W}}\right\|_{\mathbb{L}^r}\right)^r \leq 2^{r-1}\left(\left\|\widehat{\zeta}^{\mathrm{U}}\right\|_{\mathbb{L}^r}^r + \left\|\widehat{\zeta}^{\mathrm{W}}\right\|_{\mathbb{L}^r}^r\right). \tag{B.41}$$

The two inequalities above follow from the properties of the power function $\varphi_r(t) = t^r$ on $[0, +\infty)$, which is notably increasing and convex for $r \geq 1$. Specifically, the second inequality is an instance of Jensen's inequality applied to a sum of two non-negative terms with equal weights $1/2$. After replacing $\mathbb{L}^r$-norms with expectations, equation (B.41) writes:

$$\mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{W}}\right|^r\right] \leq 2^{r-1}\left(\mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{U}}\right|^r\right] + \mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{W}}\right|^r\right]\right). \tag{B.42}$$

It now remains to bound the two expectations appearing on the right-hand side. With more convenient notations, the U-statistic defined in equation (B.1) can be written as:

$$\widehat{\zeta}^{\mathrm{U}} = \frac{1}{\mathrm{Card}(\mathcal{A}_n^s)} \sum_{\boldsymbol{i} \in \mathcal{A}_n^s} \widetilde{\eta}_{\boldsymbol{i}} . \tag{B.43}$$

In the same spirit as equations (B.40) and (B.41), by successively applying the triangle inequality, the monotonicity of φ_r, and Jensen's inequality, we obtain:

$$\left|\widehat{\zeta}^{\mathrm{U}}\right| \leq \frac{1}{\mathrm{Card}(\mathcal{A}_n^s)} \sum_{\boldsymbol{i} \in \mathcal{A}_n^s} |\widetilde{\eta}_{\boldsymbol{i}}| \tag{B.44}$$

$$\left|\widehat{\zeta}^{\mathrm{U}}\right|^r \leq \left(\frac{1}{\mathrm{Card}(\mathcal{A}_n^s)} \sum_{\boldsymbol{i} \in \mathcal{A}_n^s} |\widetilde{\eta}_{\boldsymbol{i}}|\right)^r \leq \frac{1}{\mathrm{Card}(\mathcal{A}_n^s)} \sum_{\boldsymbol{i} \in \mathcal{A}_n^s} |\widetilde{\eta}_{\boldsymbol{i}}|^r . \tag{B.45}$$

Taking the expectation on both sides leads to:

$$\mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{U}}\right|^r\right] \leq \frac{1}{\mathrm{Card}(\mathcal{A}_n^s)} \sum_{\boldsymbol{i} \in \mathcal{A}_n^s} \mathbb{E}\left[\,|\widetilde{\eta}_{\boldsymbol{i}}|^r\,\right] \leq C_r \times \frac{1}{\mathrm{Card}(\mathcal{A}_n^s)} \sum_{\boldsymbol{i} \in \mathcal{A}_n^s} 1 = C_r . \tag{B.46}$$

The same reasoning applies to the other expectation. Injecting the bound C_r in equation (B.42) yields:

$$\mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{W}}\right|^r\right] \leq 2^r\, C_r . \tag{B.47}$$

Finally, bringing together equations (B.39) and (B.47) gives:

$$\mathbb{E}\left[\left|\widehat{\zeta}^{\mathrm{U}} - \widehat{\zeta}^{\mathrm{V}}\right|^r\right] \leq \left(\frac{\gamma_s}{n}\right)^r 2^r\, C_r \leq \left(\frac{s(s-1)}{n}\right)^r C_r . \tag{B.48}$$

□

Appendix C. A closer look at the HSIC

This section focuses on the Hilbert-Schmidt independence criterion (HSIC). First, Appendix C.1 introduces the HSIC between two random objects and presents several equivalent formulations. Next, Appendix C.2 describes how to estimate the HSIC from a sample of joint observations of the two objects. Then, Appendix C.3 explains how the HSIC can be used to test independence between two objects. Following this, Appendix C.4 provides several statistical examples illustrating the use of HSIC tests. Finally, Appendix C.5 addresses the particular case of parameter screening for model calibration.

Throughout this section, X and Y denote the two random objects under consideration. It is only assumed that they take values in separable spaces $\mathcal{X}$ and $\mathcal{Y}$. Their marginal distributions are denoted by $\mathbb{P}_X$ and $\mathbb{P}_Y$, respectively. The pair $Z := (X, Y)$ takes values in the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and follows the joint distribution $\mathbb{P}_Z = \mathbb{P}_{XY}$.

C.1 Definition of the HSIC

Any dependence measure aims to quantify the discrepancy between the joint distribution $\mathbb{P}_{XY}$, representing the actual dependence within the pair $Z = (X, Y)$ and the product of the marginal distributions $\mathbb{P}_X \otimes \mathbb{P}_Y$, representing a hypothetical situation of independence between the two objects. The HSIC is one such dependence measure, which captures statistical dependence using covariance kernels.

The idea behind the HSIC is to transport the distributions $\mathbb{P}_{XY}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$ into an RKHS of functions $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where their discrepancy can be more easily quantified. To this end, a kernel is assigned to both objects. Let $K_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ denote the kernels associated with X and Y, respectively. The corresponding RKHSs are denoted by $\mathcal{H}_x$ and $\mathcal{H}_y$. The HSIC is then defined as the square of the MMD between $\mathbb{P}_{XY}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$ in the RKHS $\mathcal{H}_x \otimes \mathcal{H}_y$ induced by the tensor-product kernel $K_x \otimes K_y$ [65]. This can be written as:

$$\mathrm{HSIC}(X, Y) := \mathrm{MMD}^2\left(\mathbb{P}_{XY}, \mathbb{P}_X \otimes \mathbb{P}_Y\right) = \|\mu_{\mathbb{P}_{XY}} - \mu_{\mathbb{P}_X \otimes \mathbb{P}_Y}\|^2_{\mathcal{H}_x \otimes \mathcal{H}_y} \,. \tag{C.1}$$

Let us assume that both kernels are characteristic. Under this assumption, the HSIC equals zero if and only if the two random objects are independent:

$$\mathrm{HSIC}(X, Y) = 0 \quad \Longleftrightarrow \quad X \perp\!\!\!\perp Y \,, \tag{C.2}$$

where the symbol $\perp\!\!\!\perp$ indicates an independence relationship. As will be seen later in Appendix C.3, this property is essential for independence testing [30].

By using the integral expression of the kernel mean embedding given in equation (A.1), the HSIC can be rewritten entirely in terms of kernel-based moments:

$$\begin{aligned} \mathrm{HSIC}(X, Y) = {} & \mathbb{E}\big[K_x(X_1, X_2)\, K_y(Y_1, Y_2)\big] \\ & + \mathbb{E}\big[K_x(X_1, X_2)\big]\, \mathbb{E}\big[K_y(Y_1, Y_2)\big] \\ & - 2\, \mathbb{E}\big[K_x(X_1, X_2)\, K_y(Y_1, Y_3)\big]. \end{aligned} \tag{C.3}$$

Here, $Z_1 = (X_1, Y_1)$, $Z_2 = (X_2, Y_2)$ and $Z_3 = (X_3, Y_3)$ are three independent pairs, with common distribution $\mathbb{P}_Z = \mathbb{P}_{XY}$. This reformulation makes the HSIC amenable to inference. The next section explains how to construct an estimator of the HSIC from a given sample of observations.

C.2 Estimation of the HSIC

Throughout Appendix C, the data consist of an n-sample of Z. It is composed of n independent and identically distributed (i.i.d.) observations of Z:

$$D_n := \left\{\left(Z^{(i)}\right)\right\}_{1 \le i \le n} = \left\{\left(X^{(i)}, Y^{(i)}\right)\right\}_{1 \le i \le n} \quad \text{with} \quad Z^{(i)} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_Z \,. \tag{C.4}$$

The idea is to express the HSIC in the canonical form given in equation (B.1), enabling the use of a U- or V-statistic estimator. By merging all expectations, the HSIC expression established in equation (C.3) can be compacted into a single

expectation:

$$\mathrm{HSIC}(X,Y) = \mathbb{E}_{\mathbb{P}_Z}\left[\eta\left(Z_1,Z_2,Z_3,Z_4\right)\right] , \tag{C.5}$$

where the variables Z_i are independent copies of $Z \sim \mathbb{P}_Z$, and η is the four-argument function defined by:

$$\eta(z_1,z_2,z_3,z_4) = K_x(x_1,x_2)\,K_y(y_1,y_2) + K_x(x_1,x_2)\,K_y(y_3,y_4) - 2\,K_x(x_1,x_2)\,K_y(y_1,y_3) . \tag{C.6}$$

While the kernels K_x and K_y are symmetric, the function η itself is not. It is therefore necessary to introduce its symmetrized version. Specializing equation (B.3) yields:

$$\widetilde{\eta}(z_1,z_2,z_3,z_4) = \frac{1}{4!}\sum_{\sigma\in\mathbb{S}_4}\eta\left(z_{\sigma(1)},z_{\sigma(2)},z_{\sigma(3)},z_{\sigma(4)}\right) , \tag{C.7}$$

where $\mathbb{S}_4$ denotes the set of all permutations of $\{1,\ldots,4\}$. Since any permutation of the vector (Z_1,Z_2,Z_3,Z_4) has the same distribution, one still has:

$$\mathrm{HSIC}(X,Y) = \mathbb{E}_{\mathbb{P}_Z}\left[\widetilde{\eta}\left(Z_1,Z_2,Z_3,Z_4\right)\right] , \tag{C.8}$$

but the four-argument fonction is now symmetric. Then, the HSIC can be estimated by using either a U- or V-statistic:

- $$\widehat{H}_n^{\mathrm{U}} := \binom{n}{4}^{-1}\sum_{1\le i_1<\ldots<i_4\le n}\widetilde{\eta}\left(Z^{(i_1)},\ldots,Z^{(i_4)}\right) , \tag{C.9}$$
- $$\widehat{H}_n^{\mathrm{V}} := \frac{1}{n^4}\sum_{1\le i_1,\ldots,i_4\le n}\widetilde{\eta}\left(Z^{(i_1)},\ldots,Z^{(i_4)}\right) . \tag{C.10}$$

These two estimators can then be instantiated by substituting $\widetilde{\eta}$ by its expression from equation (C.7). In both cases, the resulting quadruple sum can be decomposed into three terms, each of which can be further simplified by accounting for repeated terms through combinatorial counting. The final expression of the U-statistic is given by:

$$\begin{aligned}\widehat{H}_n^{\mathrm{U}} = {} & \frac{1}{(n)_2}\sum_{(i_1,i_2)\in\mathcal{I}_n^2} K_x\left(X^{(i_1)},X^{(i_2)}\right)\,K_y\left(Y^{(i_1)},Y^{(i_2)}\right) \\ & + \frac{1}{(n)_4}\sum_{(i_1,i_2,i_3,i_4)\in\mathcal{I}_n^4} K_x\left(X^{(i_1)},X^{(i_2)}\right)\,K_y\left(Y^{(i_3)},Y^{(i_4)}\right) \\ & - \frac{2}{(n)_3}\sum_{(i_1,i_2,i_3)\in\mathcal{I}_n^3} K_x\left(X^{(i_1)},X^{(i_2)}\right)\,K_y\left(Y^{(i_1)},Y^{(i_3)}\right) .\end{aligned} \tag{C.11}$$

Regarding the notations above, $(n)_k := n!/(n-k)!$ and the summation set $\mathcal{I}_n^k$ is defined by:

$$\mathcal{I}_n^k = \{\boldsymbol{i} = (i_1,i_2,\ldots,i_k) : 1\le i_1\ne i_2\ne\ldots\ne i_k\le n\} \quad \text{with} \quad \mathrm{Card}\left(\mathcal{I}_n^k\right) = (n)_k . \tag{C.12}$$

Likewise, the final expression of the V-statistic is given by:

$$\begin{aligned}\widehat{H}_n^{\mathrm{V}} = {} & \frac{1}{n^2}\sum_{(i_1,i_2)\in\mathcal{J}_n^2} K_x\left(X^{(i_1)},X^{(i_2)}\right)\,K_y\left(Y^{(i_1)},Y^{(i_2)}\right) \\ & + \frac{1}{n^4}\sum_{(i_1,i_2,i_3,i_4)\in\mathcal{J}_n^4} K_x\left(X^{(i_1)},X^{(i_2)}\right)\,K_y\left(Y^{(i_3)},Y^{(i_4)}\right)\end{aligned} \tag{C.13}$$

$$-\frac{2}{n^3}\sum_{(i_1,i_2,i_3)\in\mathcal{J}_n^3} K_x\left(X^{(i_1)},X^{(i_2)}\right)K_y\left(Y^{(i_1)},Y^{(i_3)}\right),$$

where the summation set $\mathcal{J}_n^k$ is defined by:

$$\mathcal{J}_n^k=\{\boldsymbol{i}=(i_1,i_2,\ldots,i_k):1\leq i_1,i_2,\ldots,i_k\leq n\}\quad\text{with}\quad \mathrm{Card}\left(\mathcal{J}_n^k\right)=n^k\ . \tag{C.14}$$

The V-statistic estimator admits a simple matrix representation:

$$\widehat{H}_n^{\mathrm{V}}=\frac{1}{n^2}\,\mathrm{Tr}\left(\boldsymbol{H}\boldsymbol{L}_x\boldsymbol{H}\boldsymbol{L}_y\right)=\frac{1}{n^2}\,\mathrm{Tr}\left(\widetilde{\boldsymbol{L}}_x\widetilde{\boldsymbol{L}}_y\right)=\frac{1}{n^2}\,\mathbf{1}_n^t\left(\widetilde{\boldsymbol{L}}_x\odot\widetilde{\boldsymbol{L}}_y\right)\mathbf{1}_n\ , \tag{C.15}$$

where:

- $\mathrm{Tr}(\cdot)$ denotes the matrix trace operator.
- $\odot$ denotes the elementwise matrix product.
- $\boldsymbol{L}_x=\left[K_x\left(X^{(i)},X^{(j)}\right)\right]_{1\leq i,j\leq n}$ and $\boldsymbol{L}_y=\left[K_y\left(Y^{(i)},Y^{(j)}\right)\right]_{1\leq i,j\leq n}$ are the Gram matrices.
- $\boldsymbol{H}=\boldsymbol{I}_n-(1/n)\mathbf{1}_n\mathbf{1}_n^t$ is the centering matrix, with $\boldsymbol{I}_n\in\mathbb{R}^{n\times n}$ the identity matrix, and $\mathbf{1}_n\in\mathbb{R}^n$ a vector of ones.
- $\widetilde{\boldsymbol{L}}_x:=\boldsymbol{H}\boldsymbol{L}_x\boldsymbol{H}$ and $\widetilde{\boldsymbol{L}}_y:=\boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H}$ are the double-centered Gram matrices.

The second equality in equation (C.15) allows rewriting the HSIC in terms of two double-centered symmetric matrices, as required by the non-asymptotic Gamma test (see Appendix D). The third equality highlights that $\mathcal{O}(n^2)$ operations are sufficient to compute the V-statistic.

As a U-statistic, $\widehat{H}_n^{\mathrm{U}}$ is the unbiased estimator of $\mathrm{HSIC}(X,Y)$ with the smallest variance [75]. However, this estimator may take negative values, which is somehow unsettling (because the HSIC is a non-negative quantity). In contrast, the V-statistic $\widehat{H}_n^{\mathrm{V}}$ has complementary properties: it is biased, asymptotically unbiased and always non-negative.

C.3 Independence testing with the HSIC

The HSIC can be used to test independence between the random objects X and Y. The idea is to make a decision between the two following hypotheses:

$$(\mathrm{H}_0):X\perp\!\!\!\perp Y\quad\text{vs.}\quad(\mathrm{H}_1):\ X\not\perp\!\!\!\perp Y, \tag{C.16}$$

on the only basis of the observed data:

$$D_n^{\mathrm{obs}}:=\left\{z^{(i)}\right\}_{1\leq i\leq n}=\left\{\left(x^{(i)},y^{(i)}\right)\right\}_{1\leq i\leq n}\ . \tag{C.17}$$

The data in D_n^{obs} are written in lowercase to emphasize that D_n^{obs} is a fixed dataset, as opposed to D_n which models the underlying random mechanism generating the data. D_n^{obs} is used to perform the test, whereas D_n is only needed for theoretical considerations.

Assuming that K_x and K_y are characteristic kernels, recall the fundamental property stated in equation (C.2). The HSIC vanishes if and only if X and Y are independent objects. Consequently, the two hypotheses in equation (C.16) can be reformulated as:

$$\mathrm{HSIC}(X,Y)=0\quad\text{vs.}\quad\mathrm{HSIC}(X,Y)>0\ . \tag{C.18}$$

Two important points follow from this reformulation of the problem. First, a natural choice for the test statistic is to take either the U- or V-statistic estimator of the HSIC. In the following, it is denoted by $\widehat{H}^{\mathrm{stat}}(D_n)$, allowing both cases to be handled simultaneously. Second, the test is one-sided (upper-tailed), meaning that (H_0) is rejected when the observed value $\widehat{H}^{\mathrm{stat}}\left(D_n^{\mathrm{obs}}\right)$ of the test statistic $\widehat{H}^{\mathrm{stat}}(D_n)$ exceeds the quantile of order $(1-\alpha)$ of the null distribution.

The significance level, often set to $\alpha = 5\%$, ensures strict control of Type-I error. Alternatively, the p-value can be used to quantify how extreme the observed value is relative to the null distribution:

$$p\text{-val} := \mathbb{P}_{\mathrm{H}_0}\left(\widehat{H}^{\mathrm{stat}}(D_n) \geq \widehat{H}^{\mathrm{stat}}(D_n^{\mathrm{obs}})\right). \tag{C.19}$$

The null hypothesis (H_0) is then rejected when the p-value is smaller than α. Generally speaking, estimating a test p-value requires knowledge of the null distribution. It may be known exactly, approximated by a parametric model, or simulated using a routine applied to the data. Depending on the sample size, three procedures are available for the HSIC-based independence test.

- **The asymptotic test.** When n is large ($n \geq 500$), asymptotic theory can be used. The rescaled HSIC estimator $n\,\widehat{H}^{\mathrm{stat}}(D_n)$ converges to a degenerate limiting distribution. More precisely, this limit is a spectral distribution characterized by the eigenvalues of the centered kernels $\widetilde{K}_x$ and $\widetilde{K}_y$ (obtained by centering K_x and K_y with respect to $\mathbb{P}_X$ and $\mathbb{P}_Y$). Several techniques have been proposed to approximate the spectral distribution, including asymptotic moment formulas to fit a Gamma distribution [30] or kernel PCA to estimate the eigenvalues [76]. Regardless of the approach, the p-value is obtained at the cost of $\mathcal{O}(n^2)$ operations.
- **The permutation-based test.** When n is small ($n \leq 100$), the asymptotic test cannot be applied. The null distribution of the HSIC estimator must be simulated with a permutation mechanism. Random permutations are sequentially applied to the output data and the test statistic is recomputed on each permuted sample [29, 57]. For any given permutation $\tau \in \mathbb{S}_n$, the permuted sample D_n^τ is derived from D_n^{obs} by permuting the outputs $y^{(i)}$ according to τ:

$$D_n^\tau = \left\{\left(x^{(i)}, y^{(\tau(i))}\right)\right\}_{1\leq i\leq n}. \tag{C.20}$$

  The resulting HSIC estimate $\widehat{H}^{\mathrm{stat}}(D_n^\tau)$ can be regarded as a realization of the test statistic under (H_0). Repeating this procedure for B permutations yields B realizations from the null distribution. Using an empirical estimator of the probability in equation (C.19), the p-value is approximated as:

$$\widehat{p}\text{-val} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\left\{\widehat{H}^{\mathrm{stat}}\left(D_n^{\tau_b}\right) > \widehat{H}^{\mathrm{stat}}\left(D_n^{\mathrm{obs}}\right)\right\}}. \tag{C.21}$$

  Overall, the computational complexity of this procedure amounts to $\mathcal{O}(B\,n^2)$.
- **The non-asymptotic Gamma test.** When n is intermediate ($100 \leq n \leq 500$), the permutation-based test becomes computationally expensive. To alleviate this burden, a parametric estimate of the null distribution can be constructed. A computational trick allows the first two moments of the test statistic under the permutation distribution to be derived analytically, without explicitly generating permutations [58, 77]. These moments are used to fit a Gamma distribution, from which the p-value is obtained. With an algorithmic complexity of only $\mathcal{O}(n^2)$, this procedure can be viewed as a low-complexity version of the permutation-based test. Appendix D.1 provides all technical details needed to implement this procedure.

C.4 Statistical applications of the HSIC

As highlighted in Appendix C.3, the HSIC provides a nonparametric independence test between two random objects. Importantly, it only requires a sample of n joint observations.

In the seminal work by Gretton et al. [30], the HSIC test was first used as a validation tool for *independent component analysis* (ICA). In that setting, the data consist of signals obtained from mixtures of mutually independent sources. The goal of ICA is to estimate the mixing matrix and to reconstruct the source signals. At the end of the procedure, independence between the reconstructed sources can be assessed through HSIC tests.

Subsequently, the HSIC was promoted as a powerful tool for *feature selection* in supervised learning [72]. A natural idea in this context is to prioritize explanatory variables that exhibit strong statistical dependence with the target variable. By relying on HSIC tests, one can easily detect variables that have a strongly nonlinear effect on the target. However, this strategy may lead to select variables with similar explanatory power, thereby unnecessarily complicating

the predictive model. To mitigate this redundancy issue, the HSIC-Lasso leverages an ℓ^1-penalty to reduce intra-feature dependencies [78].

In most safety assessment studies, the output variable $Y = f(\boldsymbol{X})$ is computed from a set of input variables $\boldsymbol{X} = (X_1, \ldots, X_d)$ by evaluating a deterministic simulation code $f : \mathcal{X} \to \mathcal{Y}$. In high-dimensional settings, where f takes as input several dozens of parameters, many numerical methods in learning and optimization suffer from the so-called *curse of dimensionality*. To remedy this problem, a preliminary step of *screening* is often performed in order to identify the inputs X_i with negligible influence on Y, so that they can be fixed at their nominal values. As explained by Da Veiga [28], this can be achieved using HSIC tests, since they allow identifying the inputs for which it is reasonable to accept the null hypothesis $(\mathrm{H}_0) : X_i \perp\!\!\!\perp Y$.

Remark C.1. Let us assume that X_i is not a dummy input, meaning that $x_i \mapsto f(\boldsymbol{x})$ is not a constant function. In most cases, this functional relationship induces a probabilistic dependence between X_i and Y. Hence, from a theoretical standpoint, the alternative hypothesis $(\mathrm{H}_1) : X_i \not\perp\!\!\!\perp Y$ holds, even if the influence of X_i on Y is negligible. Moreover, when characteristic kernels are used, the HSIC test is asymptotically consistent [30]. This property ensures that the probability of detecting an input variable X_i for which (H_1) is true converges to one as the sample size n increases. Therefore, the HSIC test is expected to detect all input variables. In practice, this is not what happens because n is limited (typically $10^2 \leq n \leq 10^3$ for most industrial numerical simulators). In this regime, it is difficult to distinguish weak dependence from true independence using the available data. As a result, the HSIC test leads to reject the null hypothesis (H_0) for weakly influential input variables, making it a powerful screening tool.

C.5 Using the HSIC for parameter screening

Parameter screening prior to Bayesian calibration is a specific instance of the general screening problem, with a few particularities. Using the notations introduced in Section 3.2, the HSIC test can be employed to select the most influential calibration parameters Θ_j with respect to the output of interest $G_{\mathrm{nom}} = f(\boldsymbol{\Theta})$. The function f represents the deterministic relationship between the calibration parameters $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_p)$ and the output under study G_{nom}. It encapsulates the full simulation chain y and depends on the chosen screening strategy. Recall the three possible definitions of G_{nom}:

- **Approach A:** $G_{\mathrm{nom}} = f_i^A(\boldsymbol{\Theta}) = Y_i = y(\boldsymbol{x}_i, \lambda_{\mathrm{nom}}, \boldsymbol{\Theta})$;
- **Approach B:** $G_{\mathrm{nom}} = f^B(\boldsymbol{\Theta}) = \boldsymbol{Y} = [Y_i]_{1 \leq i \leq n_{\mathrm{exp}}}$;
- **Approach C:** $G_{\mathrm{nom}} = f^C(\boldsymbol{\Theta}) = L = \left(\boldsymbol{z} - \boldsymbol{Y}\right)^t \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1} \left(\boldsymbol{z} - \boldsymbol{Y}\right)$.

The estimation of HSIC indices (see Appendix C.2) and the execution of independence tests (see Appendix C.3) is straightforward, provided that all variables are equipped with appropriate kernels. All calibration parameters Θ_j and the outputs Y_i and L are scalar and continuous. Consequently, these variables can be equipped with standard univariate kernels (such as Gaussian or Matérn kernels). In our numerical experiments (see Sects. 3.3 and 5), Gaussian kernels are employed, and their bandwidth parameters are chosen using the median heuristic. The only critical point is the choice of a kernel for the multivariate output $\boldsymbol{Y} \in \mathbb{R}^q$ (with $q = n_{\mathrm{exp}}$). Several options can be considered.

C.5.1 RBF kernels

The simplest choice is to consider multivariate kernels with radial isotropic structure, commonly referred to as *radial basis function* (RBF) kernels. For example, the multivariate Gaussian kernel is defined as:

$$\forall\, \boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^q, \quad K_{\mathcal{N}}\left(\boldsymbol{y}, \boldsymbol{y}'\right) = \exp\left(-\frac{\|\boldsymbol{y} - \boldsymbol{y}'\|^2}{2\,\gamma}\right). \tag{C.22}$$

According to Proposition A.10, $K_{\mathcal{N}}$ is characteristic on $\mathbb{R}^q$. The FT of its signature ψ, required to establish equation (A.11), may be found in [79] (see Ex. 2.7). Note that $K_{\mathcal{N}}$ is *radial* (as it only depends on the Euclidean distance between $\boldsymbol{y}$ and $\boldsymbol{y}'$) and *isotropic* (as it involves a unique bandwidth parameter γ). It is therefore not recommended when the components of $\boldsymbol{Y}$ have very different marginal distributions.

C.5.2 Tensor-product kernels

An alternative kernel is the tensor-product kernel K_{tens} defined in equation (A.12). The idea is to associate a univariate kernel K_i with each component Y_i, and take their tensor product. In the case of Gaussian kernels, this yields:

$$\forall\, \boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^q, \quad K_{\text{tens}}\left(\boldsymbol{y}, \boldsymbol{y}'\right) = \prod_{i=1}^{q} K_i\left(y_i, y_i'\right) = \prod_{i=1}^{q} \exp\left[\frac{1}{2}\left(\frac{y_i - y_i'}{\gamma_i}\right)^2\right]. \tag{C.23}$$

If all univariate kernels K_i are translation-invariant and characteristic on $\mathbb{R}$, then K_{tens} is also translation-invariant and characteristic on $\mathbb{R}^q$ (see Prop. A.12). K_{tens} is considerably more flexible than $K_{\mathcal{N}}$, as each bandwidth parameter γ_i can be adapted to the scale of Y_i. Empirically, however, K_{tens} tends to disadvantage parameters whose influence is confined to a small number of output components.

C.5.3 Weighted PCA kernel

To address this limitation, we consider the weighted PCA kernel proposed by El Amri and Marrel [58]. For simplicity, the random vector $\boldsymbol{Y}$ is assumed to be centered and scaled. Let $\boldsymbol{\Sigma}$ denote its covariance matrix:

$$\boldsymbol{\Sigma} = \left[\sigma_{kl}\right]_{1 \leq k,l \leq q} \quad \text{with} \quad \sigma_{kl} = \operatorname{Cov}\left(Y_k, Y_l\right). \tag{C.24}$$

In practice, this matrix is estimated empirically from centered and scaled observations of $\boldsymbol{Y}$. Let $\{\lambda_i\}_{i=1}^q$ and $\{\boldsymbol{v}_i\}_{i=1}^q$ denote the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$. Using first the canonical basis $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_q)$ and then the eigenbasis $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q)$, the vector $\boldsymbol{Y}$ can be decomposed as follows:

$$\boldsymbol{Y} = \sum_{i=1}^{q} Y_i\, \boldsymbol{e}_i = \sum_{i=1}^{q} C_i\, \boldsymbol{v}_i \quad \text{with} \quad C_i = \langle \boldsymbol{Y}, \boldsymbol{v}_i \rangle_{\mathbb{R}^q} = \boldsymbol{v}_i^t\, \boldsymbol{Y}\ . \tag{C.25}$$

The total variance of $\boldsymbol{Y}$ is defined as the sum of the variances of its individual components:

$$\operatorname{Tr}\left(\boldsymbol{\Sigma}\right) = \sum_{i=1}^{q} \mathbb{V}\left(Y_i\right) = \sum_{i=1}^{q} \lambda_i\ . \tag{C.26}$$

We retain the smallest number r of principal components that ensures a loss of information smaller than $\epsilon = 5\%$ of the total variability:

$$r := \min\left\{1 \leq k \leq q : \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{q} \lambda_i \geq 1 - \epsilon\right\}. \tag{C.27}$$

The projection $\mathcal{P} : \mathbb{R}^q \to \mathbb{R}^r$ from the physical space to the latent space of principal components is given by:

$$\mathcal{P}\left(\boldsymbol{y}\right) := \left[c_i\right]_{1 \leq i \leq r} = \boldsymbol{V}_r^t\, \boldsymbol{y} \quad \text{with} \quad \boldsymbol{V}_r = \left[\ \boldsymbol{v}_1 \ \middle|\ \boldsymbol{v}_2 \ \middle|\ \ldots \ \middle|\ \boldsymbol{v}_r\ \right] \in \mathbb{R}^{q \times r} \tag{C.28}$$

The weighted PCA kernel is then defined as:

$$\forall\, \boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^q, \quad K_{\text{PCA}}^r\left(\boldsymbol{y}, \boldsymbol{y}'\right) = \sum_{i=1}^{r} \lambda_i\, K_i\left(c_i, c_i'\right), \tag{C.29}$$

where the functions K_i are univariate kernels (for instance Gaussian kernels) assigned to the first r principal components. Whenever $r < q$, K_{PCA}^r ceases to be characteristic on $\mathbb{R}^q$. Indeed, one can find two distributions ($\mathbb{P}_1 \neq \mathbb{P}_2$) that share identical marginal distributions for the first r principal components but differ on the remaining components. As the resulting kernel mean embeddings are identical ($\mu_{\mathbb{P}_1} = \mu_{\mathbb{P}_2}$), K_{PCA}^r cannot be characteristic. Nevertheless, empirical

evidence shows that this kernel provides increased detection power for inputs whose influence is localized on a small number of output components.

Appendix D. Non-asymptotic Gamma test procedure

D.1 Standard procedure

Let us consider the classical setting, in which the HSIC is used to make a decision between the two following hypotheses:

$$(\mathrm{H}_0) : X \perp\!\!\!\perp Y \quad \text{vs.} \quad (\mathrm{H}_1) : \; X \not\perp\!\!\!\perp Y \; . \tag{D.1}$$

An overview of existing test procedures is provided in Appendix C.3. In particular, it is reported that the non-asymptotic Gamma test is well suited to intermediate sample sizes ($100 \leq n \leq 500$). In this regime, both the asymptotic test and the permutation-based test suffer from limitations. On the one hand, the sample size n is not sufficiently large for the asymptotic test to be used reliably. On the other hand, the computational complexity of the permutation-based test, namely $\mathcal{O}(B\,n^2)$, can already be substantial, especially if the number of permutations B is set very high. In this context, the non-asymptotic Gamma test lies midway, thanks to a Gamma approximation of the null distribution obtained without relying on any asymptotic result. It combines two advantages: it can be used in the non-asymptotic regime (contrary to the asymptotic test) and it has a low computational complexity (compared to that of the permutation-based test).

This test procedure applies only to the V-statistic estimator of the HSIC because its key element is the associated matrix representation:

$$\widehat{H}^{\mathrm{V}}\left(D_n^{\mathrm{obs}}\right) = \frac{1}{n^2}\,\mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{W}\right) \quad \text{with} \quad \begin{cases} \boldsymbol{A} = \widetilde{\boldsymbol{L}}_x = \boldsymbol{H}\boldsymbol{L}_x\boldsymbol{H} \\ \boldsymbol{W} = \widetilde{\boldsymbol{L}}_y = \boldsymbol{H}\boldsymbol{L}_y\boldsymbol{H} \end{cases} . \tag{D.2}$$

All details regarding the construction of this estimator are provided in Appendix C.2. The available data D_n^{obs} is a Monte Carlo sample of n joint observations, as in equation (C.17). To simulate (H_0), the permutation-based test randomly permutes the output data and recomputes the test statistic for each permuted sample. Using a sequence of B permutations, denoted by $\{\tau_b\}_{b=1}^{B}$, the simulated values of the V-statistic are:

$$\left\{\widehat{H}^{\mathrm{V}}\left(D_n^{\tau_b}\right)\right\}_{1\leq b\leq B} \quad \text{with} \quad \forall\, 1 \leq b \leq B, \quad D_n^{\tau_b} = \left\{\left(x^{(i)}, y^{(\tau_b(i))}\right)\right\}_{1\leq i\leq n} . \tag{D.3}$$

The starting point is to notice that the histogram of these values appears to be consistent with a Gamma distribution. This is unsurprising, as n is already moderately large, and it is well established that the spectral distribution governing the asymptotic behavior of the V-statistic can be closely approximated by a Gamma distribution [30, 80].

A Gamma distribution could be fitted to the sample of HSIC values provided in equation (D.3), but this would not help reduce the computational complexity of the permutation-based test. Conveniently, as explained in [58], a Gamma approximation can be constructed directly, without relying on the values simulated by the permutation-based test.

Before continuing, we clarify some elements of the probabilistic modeling. Let τ be a random variable uniformly distributed over the set $\mathbb{S}_n$ of all possible permutations. In the following, the probability measure $\mathbb{P}_\tau$ will be referred to as the *permutation distribution*. Since the sample D_n^{obs} is fixed, the random variable $\widehat{H}^{V}\left(D_n^{\tau}\right)$ is the image of τ under the operation of permuting the output data and computing the V-statistic.

When a permutation $\tau \in \mathbb{S}_n$ is applied to the output data, only the matrix $\boldsymbol{W}$ is modified: its rows and columns are permuted according to τ. The new value of the V-statistic admits the following matrix representation:

$$\mathcal{S}(\tau) := n^2\,\widehat{H}^{\mathrm{V}}\left(D_n^{\tau}\right) = \mathrm{Tr}\big(\boldsymbol{A}\boldsymbol{W}^{\tau\tau}\big) \quad \text{with} \quad \boldsymbol{W}^{\tau\tau} := \left[W_{\tau(i)\tau(j)}\right]_{1\leq i,j\leq n} . \tag{D.4}$$

The first two moments of $\mathcal{S}(\tau)$ under $\mathbb{P}_\tau$ can be computed exactly. This result can be stated in a general setting. For any square matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, we first introduce the following three functionals:

$$D_1^{M} := \mathrm{Tr}(\boldsymbol{M}) = \sum_{i=1}^{n} M_{ii} \tag{D.5}$$

$$D_2^{\boldsymbol{M}} := \mathrm{Tr}\left(\boldsymbol{M}^{\odot 2}\right) = \sum_{i=1}^{n} M_{ii}^2 \tag{D.6}$$

$$S_2^{\boldsymbol{M}} := \mathrm{Tr}\left(\boldsymbol{M}^2\right) = \mathrm{Sum}\left(\boldsymbol{M}^{\odot 2}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} M_{ij}^2 \ . \tag{D.7}$$

For a given matrix $\boldsymbol{M}$, computing these three quantities requires n, $2n$ and $2n^2$ operations, respectively. To simplify the formulas to come, we also introduce some auxiliary notations:

$$f_1(\boldsymbol{M}) := (n-1)\, S_2^{\boldsymbol{M}} - \left(D_1^{\boldsymbol{M}}\right)^2 \tag{D.8}$$

$$f_2(\boldsymbol{M}) := n\,(n+1)\, D_2^{\boldsymbol{M}} - (n-1)\left[\left(D_1^{\boldsymbol{M}}\right)^2 + 2\, S_2^{\boldsymbol{M}}\right] \tag{D.9}$$

$$p_n := (n-1)^2\,(n+1)\,(n-2) \tag{D.10}$$

$$q_n := (n+1)\,n\,(n-1)\,(n-2)\,(n-3)\ . \tag{D.11}$$

The main result, stated in the following lemma, can now be written in a compact form.

Lemma D.1. *Assume that $\boldsymbol{A}$ and $\boldsymbol{W}$ are two double-centered symmetric matrices of size n. Let $\mathcal{S}(\tau)$ be the random variable defined in equation (D.4). Its first two moments under the permutation distribution $\mathbb{P}_\tau$ are given by:*

$$\mu_{\mathrm{perm}} := \mathbb{E}_{\mathbb{P}_\tau}\left[\mathrm{Tr}\big(\boldsymbol{A}\boldsymbol{W}^{\tau\tau}\big)\right] = \frac{1}{n!}\sum_{\tau\in\mathbb{S}_n} \mathrm{Tr}\big(\boldsymbol{A}\boldsymbol{W}^{\tau\tau}\big) \tag{D.12}$$

$$= \frac{1}{n-1}\, D_1^{\boldsymbol{A}}\, D_1^{\boldsymbol{W}} =: E_0\left(\boldsymbol{A},\boldsymbol{W}\right) \tag{D.13}$$

$$\sigma^2_{\mathrm{perm}} := \mathbb{V}_{\mathbb{P}_\tau}\left(\mathrm{Tr}\big(\boldsymbol{A}\boldsymbol{W}^{\tau\tau}\big)\right) = \mathbb{E}_{\mathbb{P}_\tau}\left[\mathrm{Tr}\big(\boldsymbol{A}\boldsymbol{W}^{\tau\tau}\big)^2\right] - (\mu_{\mathrm{perm}})^2 \tag{D.14}$$

$$= \frac{2}{p_n}\, f_1(\boldsymbol{A})\, f_1(\boldsymbol{W}) + \frac{1}{q_n}\, f_2(\boldsymbol{A})\, f_2(\boldsymbol{W}) =: V_0\left(\boldsymbol{A},\boldsymbol{W}\right). \tag{D.15}$$

The calculations are detailed in [77] (see Sect. 3). In particular, the formulas in Lemma D.1 show that the exact values of μ_{perm} and σ^2_{perm} can be obtained at the cost of $\mathcal{O}(n^2)$ operations.

Back to the HSIC framework, applying Lemma D.1 to equation (D.2) yields:

$$\mu_H := \mathbb{E}_\tau\left[\widehat{H}^{\mathrm{V}}\left(D_n^\tau\right)\right] = \frac{1}{n^2}\, E_0\left(\boldsymbol{A},\boldsymbol{W}\right) \tag{D.16}$$

$$\sigma_H^2 := \mathbb{V}_\tau\left(\widehat{H}^{\mathrm{V}}\left(D_n^\tau\right)\right) = \frac{1}{n^4}\, V_0\left(\boldsymbol{A},\boldsymbol{W}\right). \tag{D.17}$$

Then, we resort to the method of moments to construct a Gamma approximation of the V-statistic $\widehat{H}^{\mathrm{V}}\left(D_n^\tau\right)$ under the permutation distribution $\mathbb{P}_\tau$. Recall that, for a random variable $S \sim \mathrm{Ga}(a,b)$ following a Gamma distribution with shape parameter a and scale parameter b, the first two moments satisfy:

$$\begin{cases} \mu_S := \mathbb{E}[S] = a\,b \\ \sigma_S^2 := \mathbb{V}(S) = a\,b^2 \end{cases} \quad \text{and thus} \quad \begin{cases} a = \mu_S^2/\sigma_S^2 \\ b = \sigma_S^2/\mu_S \end{cases} \ . \tag{D.18}$$

Consequently, the parameters of the Gamma distribution used to approximate the null distribution can be estimated by:

$$\widehat{a}_H = \mu_H^2/\sigma_H^2 \quad \text{and} \quad \widehat{b}_H = \sigma_H^2/\mu_H. \tag{D.19}$$

Finally, the test p-value can be computed directly from the survival function of the estimated Gamma distribution.

Throughout the test procedure, it is never necessary to permute the output data and recompute the V-statistic. In terms of algorithm complexity, the construction of the Gamma approximation entails $\mathcal{O}(n^2)$ operations, which is much cheaper than the permutation-based test. However, as with the asymptotic test, using a Gamma approximation remains heuristic [80]. Even though this parametric choice was validated empirically in the numerical study proposed in [58], there is no theoretical guarantee that it strictly controls Type-I and Type-II errors.

D.2 New procedure in the bi-level uncertainty framework

The permutation-based test procedure developed in Section 4.2 allows one to decide between the two following hypotheses:

$$(\mathrm{H}_0) : \Theta_j \perp\!\!\!\perp G \quad \text{vs.} \quad (\mathrm{H}_1) : \Theta_j \not\perp\!\!\!\perp G \quad \text{with} \quad G = \widetilde{f}\,(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\,. \tag{D.20}$$

In this context, any of the four estimators of $\mathscr{H}_j$ proposed in equation (4.18) can be used as test statistic. Interestingly, the permutation-based test applies to all estimators. However, only Estimators 1 and 2 are considered in Section 4.2, since they lead to faster MSE convergence rates, as reported in Table 2.

The computational complexity of the permutation-based test is $\mathcal{O}(B\,m\,n^2)$. Once again, this raises the classical issue of the high computational complexity associated with permutation-based test procedures. To reduce this complexity, our idea is to leverage the computational trick of the non-asymptotic Gamma test. Since this procedure is specific to the V-statistic estimator of the HSIC, we will focus exclusively on Estimator 2:

$$\widehat{\widehat{\mathscr{H}}}_j^{\,2}\left(D_{m,n}^{\mathrm{obs}}\right) = \frac{1}{m}\sum_{k=1}^{m} \widehat{H}_j^{\mathrm{V}}\left(D_k^{\mathrm{obs}}\right). \tag{D.21}$$

As detailed in Section 3.4, $D_{m,n}^{\mathrm{obs}}$ is a nested Monte Carlo design generated according to (S1), and each design D_k^{obs} is an n-sample of the pair $\left(\boldsymbol{\Theta}^{(k)}, G^{(k)}\right)$. Replacing the V-statistics by their matrix representations leads to:

$$\widehat{\widehat{\mathscr{H}}}_j^{\,2}\left(D_{m,n}^{\mathrm{obs}}\right) = \frac{1}{m\,n^2}\sum_{k=1}^{m} \mathrm{Tr}\left(\boldsymbol{A}_{jk}\boldsymbol{W}_k\right) \quad \text{with} \quad \begin{cases} \boldsymbol{A}_{jk} := \widetilde{\boldsymbol{L}}_{\theta_j}^{(k)} = \boldsymbol{H}\boldsymbol{L}_{\theta_j}^{(k)}\boldsymbol{H} \\ \boldsymbol{W}_k := \widetilde{\boldsymbol{L}}_{g}^{(k)} = \boldsymbol{H}\boldsymbol{L}_{g}^{(k)}\boldsymbol{H} \end{cases}, \tag{D.22}$$

where $\boldsymbol{L}_{\theta_j}^{(k)}$ and $\boldsymbol{L}_g^{(k)}$ denote the Gram matrices built from the observations of $\Theta_j^{(k)}$ and $G^{(k)}$ found in the sample D_k^{obs}. To simulate the null distribution, a permutation τ_k is applied to the output data of each design D_k^{obs}. A realization of the null distribution is thus obtained by randomly selecting a set of m permutations $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$, applying them to the samples D_k^{obs}, and recomputing the test statistic:

$$\mathcal{S}_j(\boldsymbol{\tau}) := \widehat{\widehat{\mathscr{H}}}_j^{\,2}\left(D_{m,n}^{\boldsymbol{\tau}}\right) = \frac{1}{m}\sum_{k=1}^{m} \widehat{H}_j^{\mathrm{V}}\left(D_k^{\tau_k}\right) = \frac{1}{m\,n^2}\sum_{k=1}^{m} \mathrm{Tr}\left(\boldsymbol{A}_{jk}\boldsymbol{W}_k^{\tau_k\tau_k}\right). \tag{D.23}$$

If the process is repeated B times, a B-sample of the null distribution is produced. When plotting the associated histogram (see Fig. E.1), it appears fully consistent with a Gamma distribution. This suggests that a Gamma approximation could be constructed.

As for the non-asymptotic Gamma test, the idea is to calculate the exact values of the first two moments, from which a Gamma distribution is fitted. The only difference is that the permutation scheme here involves a set of permutations $\boldsymbol{\tau} \in (\mathbb{S}_n)^m$, instead of a single permutation $\tau \in \mathbb{S}_n$. Fortunately, Lemma D.1 allows for a straightforward derivation of the first two moments of $\mathcal{S}_j(\boldsymbol{\tau})$ under $\mathbb{P}_{\boldsymbol{\tau}}$:

$$\mu_j = \mathbb{E}_{\boldsymbol{\tau}}\left[\mathcal{S}_j(\boldsymbol{\tau})\right] = \frac{1}{m\,n^2}\sum_{k=1}^{m} \mathbb{E}_{\tau_k}\left[\mathrm{Tr}\left(\boldsymbol{A}_{jk}\boldsymbol{W}_k^{\tau_k\tau_k}\right)\right] = \frac{1}{m\,n^2}\sum_{k=1}^{m} E_0\left(\boldsymbol{A}_{jk}, \boldsymbol{W}_k\right) \tag{D.24}$$

$$\sigma_j^2 = \mathbb{V}_{\boldsymbol{\tau}}\big(\mathcal{S}_j(\boldsymbol{\tau})\big) = \frac{1}{m^2\, n^4} \sum_{k=1}^{m} \mathbb{V}_{\tau_k}\big(\mathrm{Tr}\left(\boldsymbol{A}_{jk} \boldsymbol{W}_k^{\tau_k \tau_k}\right)\big) = \frac{1}{m^2\, n^4} \sum_{k=1}^{m} V_0\left(\boldsymbol{A}_{jk}, \boldsymbol{W}_k\right). \tag{D.25}$$

The variance formula holds because of the mutual independence of the random permutations in the set $\boldsymbol{\tau}$. Once μ_j and σ_j^2 are computed, the parameters $(\widehat{a}_j, \widehat{b}_j)$ of the Gamma approximation and the final p-value follow directly.

Figure E.1 reveals that the Gamma approximation closely matches the histogram of the null distribution obtained from the permutation-based test, which provides empirical support for this test procedure.

With an overall computational complexity of $\mathcal{O}(m\, n^2)$, instead of $\mathcal{O}(B\, m\, n^2)$, this non-asymptotic Gamma test stands as a low-complexity alternative to the permutation-based test.

Appendix E. Technical proofs

We begin with a simple lemma that will be used in a proof of this section.

Lemma E.1. *Let $(Z_1, \ldots, Z_m)$ be an i.i.d. sample from the distribution $\mathbb{P}_Z$. The second moment of the empirical mean $\bar{Z}_n := \frac{1}{m}\sum_{k=1}^{m} Z_k$ is given by:*

$$\mathbb{E}\left[\bar{Z}_m^2\right] = \frac{1}{m}\,\mathbb{E}\left[Z_1^2\right] + \frac{m-1}{m}\,\left(\mathbb{E}\left[Z_1\right]\right)^2 \leq \mathbb{E}\left[Z_1^2\right] \tag{E.1}$$

Proof. Let μ_1, μ_2 and σ^2 be the first moment, the second moment, and the variance of $\mathbb{P}_Z$, respectively. Then, one has $\sigma^2 = \mu_2 - \mu_1^2$ (by definition) and $\mu_1^2 \leq \mu_2$ (by convexity). For the empirical mean $\bar{Z}_m$, it is well-known that:

$$\mathbb{E}\left[\bar{Z}_m\right] = \mu_1 \quad \text{and} \quad \mathbb{V}\left(\bar{Z}_m^2\right) = \frac{\sigma^2}{m}\,. \tag{E.2}$$

Therefore, the second moment satisfies:

$$\mathbb{E}\left[\bar{Z}_n^2\right] = \mathbb{V}\left(\bar{Z}_n\right) + \mathbb{E}\left[\bar{Z}_n\right]^2 = \frac{\sigma^2}{m} + \mu_1^2 = \frac{\mu_2 - \mu_1^2}{m} + \mu_1^2 = \frac{1}{m}\,\mu_2 + \frac{m-1}{m}\,\mu_1^2 \leq \mu_2\,. \tag{E.3}$$

□

E.1 Proof of Proposition 4.1

Proof. Since the kernels K_{θ_j} and K_g are characteristic, the fundamental property of the HSIC stated in equation (C.2) holds. In the present setting, it reads:

$$\forall\, \boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\Lambda}}, \quad H_j(\boldsymbol{\lambda}) = 0 \iff \mathrm{HSIC}(\Theta_j, G_{\boldsymbol{\lambda}}) = 0 \iff \Theta_j \perp\!\!\!\perp G_{\boldsymbol{\lambda}}\,. \tag{E.4}$$

As a consequence, the following equivalences hold:

$$\mathscr{H}_j = 0 \iff \mathbb{E}_{\boldsymbol{\Lambda}}\left[H_j(\boldsymbol{\Lambda})\right] = 0 \tag{E.5}$$

$$\iff H_j(\boldsymbol{\Lambda}) = 0 \quad \mathbb{P}_{\boldsymbol{\Lambda}}\text{-a.s.} \tag{E.6}$$

$$\iff \exists\, \widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}} \subseteq \mathcal{D}_{\boldsymbol{\Lambda}} \text{ such that } \mathbb{P}_{\boldsymbol{\Lambda}}(\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}) = 1 \text{ and } \forall\, \boldsymbol{\lambda} \in \widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}, H_j(\boldsymbol{\lambda}) = 0 \tag{E.7}$$

$$\iff \exists\, \widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}} \subseteq \mathcal{D}_{\boldsymbol{\Lambda}} \text{ such that } \mathbb{P}_{\boldsymbol{\Lambda}}(\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}) = 1 \text{ and } \forall\, \boldsymbol{\lambda} \in \widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}, \Theta_j \perp\!\!\!\perp G_{\boldsymbol{\lambda}}\,. \tag{E.8}$$

Equation (E.6) is immediate since $H_j(\boldsymbol{\lambda}) = \mathrm{HSIC}(\Theta_j, G_{\boldsymbol{\lambda}}) \geq 0$ for all $\boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\Lambda}}$. Equation (E.7) is obtained by expliciting the almost sure equality with respect to $\mathbb{P}_{\boldsymbol{\Lambda}}$. Finally, using equation (E.4) for all $\boldsymbol{\lambda} \in \widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}} \subseteq \mathcal{D}_{\boldsymbol{\Lambda}}$ yields equation (E.8). This last equation indicates that Θ_j and $G_{\boldsymbol{\lambda}} = \widetilde{f}\,(\boldsymbol{\Theta}, \boldsymbol{\lambda})$ are independent for almost all values $\boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\Lambda}}$. This is in fact sufficient

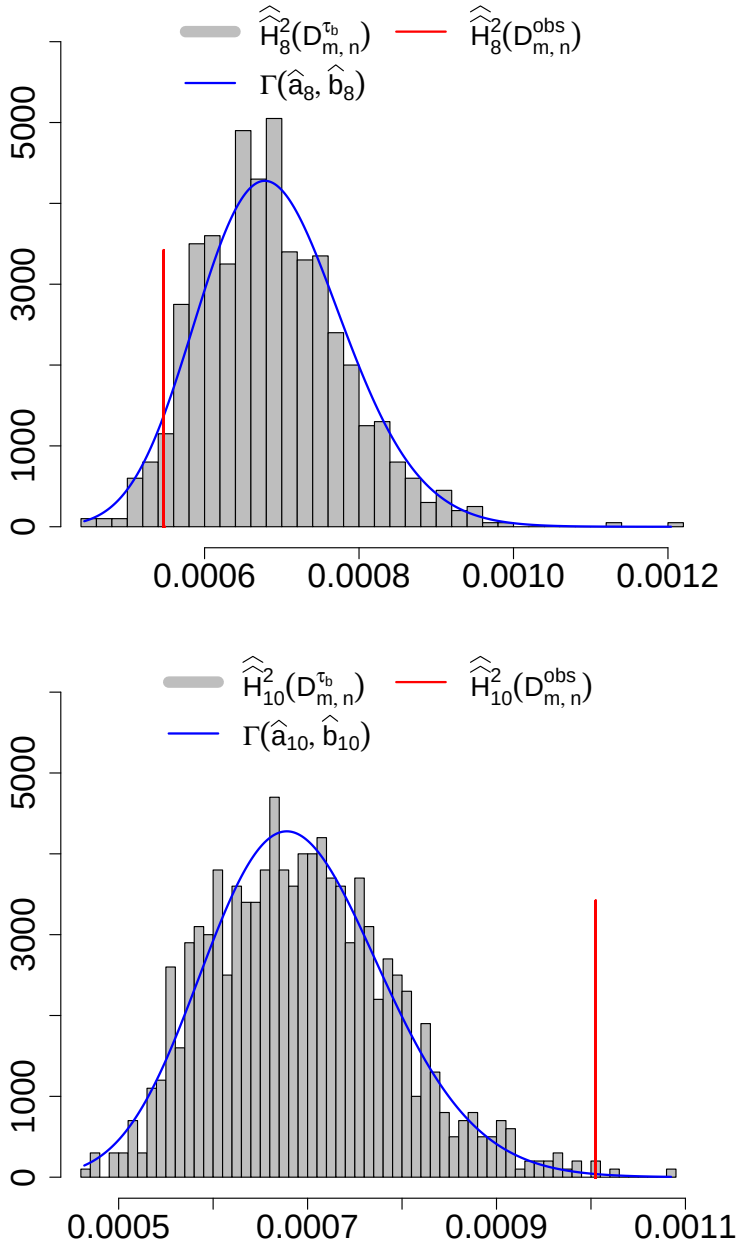

FIGURE E.1. Approximation of the null distribution of Estimator 2 by the permutation-based test and the non-asymptotic Gamma test. The red vertical line indicates the observed value of the test statistic $\widehat{\widehat{\mathscr{H}}}_j^2\left(D_{m,n}^{\text{obs}}\right)$. The histogram shows the values $\left\{\widehat{\widehat{\mathscr{H}}}_j^2\left(D_{m,n}^{\tau_b}\right)\right\}_{1\leq b\leq B}$ simulated by the permutation-based test (with $B = 10^3$). The blue curve represents the parametric estimate constructed by the non-asymptotic Gamma test. Results of Approach C are shown for Θ_8 (top) and Θ_{10} (bottom).

to conclude that $\Theta_j \perp\!\!\!\perp G = \widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})$. This can be easily verified using the factorization property of expectations. For any pair (φ, ψ) of bounded Borelian functions, one must check that:

$$\mathbb{E}\left[\varphi(\Theta_j)\,\psi(G)\right] = \mathbb{E}\left[\varphi(\Theta_j)\right]\,\mathbb{E}\left[\psi(G)\right]. \tag{E.9}$$

Let us successively rewrite the mixed moment:

$$\begin{aligned}
\mathbb{E}\left[\varphi(\Theta_j)\,\psi(G)\right] &= \mathbb{E}\left[\varphi(\Theta_j)\,\psi\left(\widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\right)\right] && \text{(E.10)}\\
&= \mathbb{E}\left[\varphi(\Theta_j)\,\psi\left(\widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\right)\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\Lambda})\right] && \text{(E.11)}\\
&= \mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathbb{E}\left[\varphi(\Theta_j)\,\psi\left(\widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\right)\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\Lambda})\middle|\boldsymbol{\Lambda}\right]\right] && \text{(E.12)}\\
&= \mathbb{E}_{\boldsymbol{\Lambda}}\left[g(\boldsymbol{\Lambda})\right]. && \text{(E.13)}
\end{aligned}$$

Equation (E.11) holds because $\mathbb{P}_{\boldsymbol{\Lambda}}(\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}) = 1$. Equation (E.12) follows directly from the tower property of conditional expectation. Furthermore, since $\boldsymbol{\Theta} \perp\!\!\!\perp \boldsymbol{\Lambda}$, the function $g : \mathcal{D}_{\boldsymbol{\lambda}} \to \mathbb{R}$ introduced in equation (E.13) can be expressed as:

$$\begin{aligned}
g(\boldsymbol{\lambda}) &= \mathbb{E}\left[\varphi(\Theta_j)\,\psi\left(\widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\right)\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\Lambda})\middle|\boldsymbol{\Lambda} = \boldsymbol{\lambda}\right] && \text{(E.14)}\\
&= \mathbb{E}\left[\varphi(\Theta_j)\,\psi\left(\widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\lambda})\right)\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\lambda})\right] = \mathbb{E}\left[\varphi(\Theta_j)\,\psi(G_{\boldsymbol{\lambda}})\,\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\lambda})\right] && \text{(E.15)}
\end{aligned}$$

To factorize the last expectation, two cases must be distinguished. On the one hand, if $\boldsymbol{\lambda} \in \widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}$, one has:

$$\begin{aligned}
\mathbb{E}\left[\varphi(\Theta_j)\,\psi(G_{\boldsymbol{\lambda}})\,\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\lambda})\right] &= \mathbb{E}\left[\varphi(\Theta_j)\,\psi(G_{\boldsymbol{\lambda}})\right] && \text{since } \mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\lambda}) = 1\ ;\\
&= \mathbb{E}\left[\varphi(\Theta_j)\right]\,\mathbb{E}\left[\psi(G_{\boldsymbol{\lambda}})\right] && \text{since } \Theta_j \perp\!\!\!\perp G_{\boldsymbol{\lambda}}\\
&= \mathbb{E}\left[\varphi(\Theta_j)\right]\,\mathbb{E}\left[\psi(G_{\boldsymbol{\lambda}})\,\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\lambda})\right] && \text{since } \mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\lambda}) = 1\ .
\end{aligned} \tag{E.16}$$

On the other hand, if $\boldsymbol{\lambda} \notin \widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}$, one has:

$$\underbrace{\mathbb{E}\left[\varphi(\Theta_j)\,\psi(G_{\boldsymbol{\lambda}})\,\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\lambda})\right]}_{=0} = 0 = \mathbb{E}\left[\varphi(\Theta_j)\right]\,\underbrace{\mathbb{E}\left[\psi(G_{\boldsymbol{\lambda}})\,\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\lambda})\right]}_{=0}\ . \tag{E.17}$$

This shows that the factorization holds for all $\boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\Lambda}}$. Hence, the function g can now be expressed as:

$$\begin{aligned}
g(\boldsymbol{\lambda}) &= \mathbb{E}\left[\varphi(\Theta_j)\right]\,\mathbb{E}\left[\psi(G_{\boldsymbol{\lambda}})\,\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\lambda})\right] && \text{(E.18)}\\
&= \mathbb{E}\left[\varphi(\Theta_j)\right]\,\mathbb{E}\left[\psi\left(\widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\right)\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\Lambda})\middle|\boldsymbol{\Lambda} = \boldsymbol{\lambda}\right]. && \text{(E.19)}
\end{aligned}$$

Injecting this last expression in equation (E.13) leads to:

$$\begin{aligned}
\mathbb{E}\left[\varphi(\Theta_j)\,\psi(G)\right] &= \mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathbb{E}\left[\varphi(\Theta_j)\right]\,\mathbb{E}\left[\psi\left(\widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\right)\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\Lambda})\middle|\boldsymbol{\Lambda}\right]\right] && \text{(E.20)}\\
&= \mathbb{E}\left[\varphi(\Theta_j)\right]\,\mathbb{E}_{\boldsymbol{\Lambda}}\left[\mathbb{E}\left[\psi\left(\widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\right)\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\Lambda})\middle|\boldsymbol{\Lambda}\right]\right] && \text{(E.21)}\\
&= \mathbb{E}\left[\varphi(\Theta_j)\right]\,\mathbb{E}\left[\psi\left(\widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\Lambda})\right)\mathbf{1}_{\widetilde{\mathcal{D}}_{\boldsymbol{\Lambda}}}(\boldsymbol{\Lambda})\right] && \text{(E.22)}\\
&= \mathbb{E}\left[\varphi(\Theta_j)\right]\,\mathbb{E}\left[\psi(G)\right]. && \text{(E.23)}
\end{aligned}$$

To move from equation (E.20) to equation (E.21), the constant $\mathbb{E}[\varphi(\Theta_j)]$ is taken outside the expectation operator. The remaining steps follow the same reasoning as above, namely the tower property and the fact that $\mathbb{P}_{\mathbf{\Lambda}}(\widetilde{\mathcal{D}}_{\mathbf{\Lambda}}) = 1$. Since equation (E.9) holds for all pairs (φ, ψ) of bounded Borelian functions, this proves that $\Theta_j \perp\!\!\!\perp G$.

Now that $\mathscr{H}_j = 0 \Rightarrow \Theta_j \perp\!\!\!\perp G$ has been proven, we turn to the other implication. It is assumed that $\Theta_j \perp\!\!\!\perp G$. Recall how $\mathscr{H}_j$ is expressed in terms of kernel-based moments:

$$\begin{aligned}\mathscr{H}_j = \mathbb{E}_{\mathbf{\Lambda}}[H_j(\mathbf{\Lambda})] = \mathbb{E}\left[K_{\theta_j}(\Theta_j, \Theta_j') K_g\left(\widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}), \widetilde{f}(\mathbf{\Theta}', \mathbf{\Lambda})\right)\right] \\ + \mathbb{E}\left[K_{\theta_j}(\Theta_j, \Theta_j')\right] \mathbb{E}\left[K_g\left(\widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}), \widetilde{f}(\mathbf{\Theta}', \mathbf{\Lambda})\right)\right] \\ - 2\,\mathbb{E}\left[K_{\theta_j}(\Theta_j, \Theta_j') K_g\left(\widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}), \widetilde{f}(\mathbf{\Theta}'', \mathbf{\Lambda})\right)\right],\end{aligned} \tag{E.24}$$

where $\mathbf{\Theta} \perp\!\!\!\perp \mathbf{\Theta}' \perp\!\!\!\perp \mathbf{\Theta}'' \sim \mathbb{P}_{\mathbf{\Theta}}$. The expectations appearing in the first and third lines can be factorized and coincide exactly with the second line:

$$\begin{aligned}&\mathbb{E}\left[K_{\theta_j}(\Theta_j, \Theta_j') K_g\left(\widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}), \widetilde{f}(\mathbf{\Theta}', \mathbf{\Lambda})\right)\right] \\ &= \mathbb{E}\left[K_{\theta_j}(\Theta_j, \Theta_j') K_g\left(\widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}), \widetilde{f}(\mathbf{\Theta}'', \mathbf{\Lambda})\right)\right] \\ &= \mathbb{E}\left[K_{\theta_j}(\Theta_j, \Theta_j')\right] \mathbb{E}\left[K_g\left(\widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}), \widetilde{f}(\mathbf{\Theta}'', \mathbf{\Lambda})\right)\right].\end{aligned} \tag{E.25}$$

Let us justify how to factorize the first expectation. On the one hand, by hypothesis, one has:

$$\Theta_j \perp\!\!\!\perp \widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}) \quad \text{and} \quad \Theta_j' \perp\!\!\!\perp \widetilde{f}(\mathbf{\Theta}', \mathbf{\Lambda}). \tag{E.26}$$

On the other hand, Remark 3.4 provides:

$$\mathbf{\Theta} \perp\!\!\!\perp \mathbf{\Theta}' \perp\!\!\!\perp \mathbf{\Theta}'' \perp\!\!\!\perp \mathbf{\Lambda}\,, \tag{E.27}$$

which in turn gives:

$$\Theta_j \perp\!\!\!\perp \widetilde{f}(\mathbf{\Theta}', \mathbf{\Lambda}) \quad \text{and} \quad \Theta_j' \perp\!\!\!\perp \widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}) \tag{E.28}$$

Equations (E.26) and (E.28) can be summarized as:

$$\begin{bmatrix}\Theta_j \\ \Theta_j'\end{bmatrix} \perp\!\!\!\perp \begin{bmatrix}\widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}) \\ \widetilde{f}(\mathbf{\Theta}', \mathbf{\Lambda})\end{bmatrix}, \tag{E.29}$$

which eventually yields:

$$K_{\theta_j}(\Theta_j, \Theta_j') \perp\!\!\!\perp K_g\left(\widetilde{f}(\mathbf{\Theta}, \mathbf{\Lambda}), \widetilde{f}(\mathbf{\Theta}', \mathbf{\Lambda})\right). \tag{E.30}$$

This justifies the factorization for the first expectation in equation (E.25). The same reasoning holds for the other expectation. Injecting the two factorized expressions in equation (E.24) leads to $\mathscr{H}_j = 0$. For this implication, note that characteristic kernels are not required. □

E.2 Proof of Proposition 4.5

E.2.0 Preparatory results

For a fixed value of $\boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\Lambda}}$, the HSIC index between Θ_j and $G_{\boldsymbol{\lambda}} = \widetilde{f}(\boldsymbol{\Theta}_j, \boldsymbol{\lambda})$ is given by:

$$\begin{aligned} H_j(\boldsymbol{\lambda}) = \mathbb{E}\left[K_{\theta_j}\left(\Theta_j^1, \Theta_j^2\right) K_g\left(\widetilde{f}\left(\boldsymbol{\Theta}^1, \boldsymbol{\lambda}\right), \widetilde{f}\left(\boldsymbol{\Theta}^2, \boldsymbol{\lambda}\right)\right)\right] \\ + \mathbb{E}\left[K_{\theta_j}\left(\Theta_j^1, \Theta_j^2\right)\right] \mathbb{E}\left[K_g\left(\widetilde{f}\left(\boldsymbol{\Theta}^3, \boldsymbol{\lambda}\right), \widetilde{f}\left(\boldsymbol{\Theta}^4, \boldsymbol{\lambda}\right)\right)\right] \\ - 2\,\mathbb{E}\left[K_{\theta_j}\left(\Theta_j^1, \Theta_j^2\right) K_g\left(\widetilde{f}\left(\boldsymbol{\Theta}^1, \boldsymbol{\lambda}\right), \widetilde{f}\left(\boldsymbol{\Theta}^3, \boldsymbol{\lambda}\right)\right)\right], \end{aligned} \tag{E.31}$$

where $\boldsymbol{\Theta}^1, \boldsymbol{\Theta}^2, \boldsymbol{\Theta}^3$ and $\boldsymbol{\Theta}^4$ are four independent and identically distributed random vectors, sharing the same joint distribution $\mathbb{P}_{\boldsymbol{\Theta}}$. Following the same approach as in Appendix C.2, $H_j(\boldsymbol{\lambda})$ can be written as a single expectation:

$$H_j(\boldsymbol{\lambda}) = \mathbb{E}_{\mathbb{P}_{\boldsymbol{\Theta}}}\left[\eta_{\boldsymbol{\lambda}}\left(\boldsymbol{\Theta}^1, \boldsymbol{\Theta}^2, \boldsymbol{\Theta}^3, \boldsymbol{\Theta}^4\right)\right], \tag{E.32}$$

where the function $\eta_{\boldsymbol{\lambda}}$ is defined as:

$$\begin{aligned} \eta_{\boldsymbol{\lambda}}: \quad \mathcal{D}_{\boldsymbol{\Theta}} \times \mathcal{D}_{\boldsymbol{\Theta}} \times \mathcal{D}_{\boldsymbol{\Theta}} \times \mathcal{D}_{\boldsymbol{\Theta}} &\longrightarrow \mathbb{R} \\ \left(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3, \boldsymbol{\theta}^4\right) &\longmapsto K_{\theta_j}\left(\theta_j^1, \theta_j^2\right) K_g\left(\widetilde{f}\left(\boldsymbol{\theta}^1, \boldsymbol{\lambda}\right), \widetilde{f}\left(\boldsymbol{\theta}^2, \boldsymbol{\lambda}\right)\right) \\ &\quad + K_{\theta_j}\left(\theta_j^1, \theta_j^2\right) K_g\left(\widetilde{f}\left(\boldsymbol{\theta}^3, \boldsymbol{\lambda}\right), \widetilde{f}\left(\boldsymbol{\theta}^4, \boldsymbol{\lambda}\right)\right) \\ &\quad - 2\, K_{\theta_j}\left(\theta_j^1, \theta_j^2\right) K_g\left(\widetilde{f}\left(\boldsymbol{\theta}^1, \boldsymbol{\lambda}\right), \widetilde{f}\left(\boldsymbol{\theta}^3, \boldsymbol{\lambda}\right)\right). \end{aligned} \tag{E.33}$$

As $\eta_{\boldsymbol{\lambda}}$ is not symmetric, it must be symmetrized:

$$\widetilde{\eta}_{\boldsymbol{\lambda}}\left(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3, \boldsymbol{\theta}^4\right) = \frac{1}{4!} \sum_{\sigma \in \mathbb{S}_4} \eta_{\boldsymbol{\lambda}}\left(\boldsymbol{\theta}^{\sigma(1)}, \boldsymbol{\theta}^{\sigma(2)}, \boldsymbol{\theta}^{\sigma(3)}, \boldsymbol{\theta}^{\sigma(4)}\right). \tag{E.34}$$

This symmetrization does not modify the target expectation:

$$H_j(\boldsymbol{\lambda}) = \mathbb{E}_{\mathbb{P}_{\boldsymbol{\Theta}}}\left[\widetilde{\eta}_{\boldsymbol{\lambda}}\left(\boldsymbol{\Theta}^1, \boldsymbol{\Theta}^2, \boldsymbol{\Theta}^3, \boldsymbol{\Theta}^4\right)\right], \tag{E.35}$$

but is required by the theory of U- and V-statistics.

The boundedness assumption on the kernels implies that:

$$\begin{aligned} &\exists\, M_{\theta_j} > 0: \quad \forall \left(\theta_j, \theta_j'\right) \in \left(\mathcal{D}_{\Theta_j}\right)^2, \quad \left|K_{\theta_j}\left(\theta_j, \theta_j'\right)\right| \leq M_{\theta_j} < \infty \\ &\exists\, M_g > 0: \quad \forall \left(g, g'\right) \in \left(\mathcal{D}_G\right)^2, \quad \left|K_g\left(g, g'\right)\right| \leq M_g < \infty\,. \end{aligned} \tag{E.36}$$

As a consequence, both $\eta_{\boldsymbol{\lambda}}$ and its symmetrized version $\widetilde{\eta}_{\boldsymbol{\lambda}}$ are bounded. The corresponding bounds admit explicit expressions in terms of M_{θ_j} and M_g:

$$\forall\, \boldsymbol{\lambda} \in \mathcal{D}_{\boldsymbol{\Lambda}}, \quad \forall \left(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3, \boldsymbol{\theta}^4\right) \in \left(\mathcal{D}_{\boldsymbol{\Theta}}\right)^4, \quad \left|\eta_{\boldsymbol{\lambda}}\left(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3, \boldsymbol{\theta}^4\right)\right| \leq 4\, M_{\theta_j}\, M_g \tag{E.37}$$

$$\left|\widetilde{\eta}_{\boldsymbol{\lambda}}\left(\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3, \boldsymbol{\theta}^4\right)\right| \leq 4\, M_{\theta_j}\, M_g\,. \tag{E.38}$$

This ensures that the function $\widetilde{\eta}_{\boldsymbol{\lambda}}$ satisfies the assumptions required by the lemmas stated in Appendix B. Specifically, using equation (E.38), one can easily show that:

$$H_j(\boldsymbol{\lambda}) = \mathbb{E}\left[\widetilde{\eta}_{\boldsymbol{\lambda}}\left(\boldsymbol{\Theta}^1, \boldsymbol{\Theta}^2, \boldsymbol{\Theta}^3, \boldsymbol{\Theta}^4\right)\right] \leq 4\, M_{\theta_j}\, M_g \tag{E.39}$$

$$\mathbb{V}_{\mathbb{P}_{\boldsymbol{\Theta}}}\left(\widetilde{\eta}_{\boldsymbol{\lambda}}\left(\boldsymbol{\Theta}^1,\boldsymbol{\Theta}^2,\boldsymbol{\Theta}^3,\boldsymbol{\Theta}^4\right)\right) \leq \left(4\,M_{\theta_j}\,M_g\right)^2 < \infty \tag{E.40}$$

$$C_r\left(\widetilde{\eta}_{\boldsymbol{\lambda}},\mathbb{P}_{\boldsymbol{\Theta}}\right) := \max_{1\leq i_1,\ldots,i_4\leq 4} \mathbb{E}\Big[\left|\widetilde{\eta}_{\boldsymbol{\lambda}}\left(\boldsymbol{\Theta}^{i_1},\ldots,\boldsymbol{\Theta}^{i_4}\right)\right|^r\Big] \leq \left(4\,M_{\theta_j}\,M_g\right)^r < \infty\,. \tag{E.41}$$

Let us now consider a Monte Carlo sample composed of n joint observations of $(\boldsymbol{\Theta}, G_{\boldsymbol{\lambda}})$ where $G_{\boldsymbol{\lambda}} = \widetilde{f}\left(\boldsymbol{\Theta},\boldsymbol{\lambda}\right)$. The dataset can be written as:

$$D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n} = \left\{\left(\boldsymbol{\Theta}^{(l)},G_{\boldsymbol{\lambda}}^{(l)}\right)\right\}_{1\leq l\leq n} \quad \text{with} \quad G_{\boldsymbol{\lambda}}^{(l)} = \widetilde{f}\left(\boldsymbol{\Theta}^{(l)},\boldsymbol{\lambda}\right). \tag{E.42}$$

Let $\widehat{H}_j^{\mathrm{U}}(\boldsymbol{\lambda})$ and $\widehat{H}_j^{\mathrm{V}}(\boldsymbol{\lambda})$ denote the U- and V-statistic estimators of $H_j(\boldsymbol{\lambda})$ based on the data from $D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}$. Using equations (E.40) and (E.41), Lemma B.3 and Lemma B.5 can be directly specialized, which yields the following corollary.

Corollary E.2. *Under the boundedness assumption on kernels, one has:*

$$\mathbb{V}\left(\widehat{H}_j^{\mathrm{U}}(\boldsymbol{\lambda})\right) \leq \frac{\alpha}{n} = \mathcal{O}\left(\frac{1}{n}\right) \qquad \textit{with} \quad \alpha = 64\,M_{\theta_j}^2\,M_g^2 \tag{E.43}$$

$$\mathbb{E}\left[\left|\widehat{H}_j^{\mathrm{U}}(\boldsymbol{\lambda}) - \widehat{H}_j^{\mathrm{V}}(\boldsymbol{\lambda})\right|^r\right] \leq \frac{\beta_r}{n^r} = \mathcal{O}\left(\frac{1}{n^r}\right) \quad \textit{with} \quad \beta_r = \left(48\,M_{\theta_j}^2\,M_g^2\right)^r. \tag{E.44}$$

Note that the constants α and β_r depend only on the kernel bounds M_{θ_j} and M_g, and are therefore independent of $\boldsymbol{\lambda}$ and $\mathbb{P}_{\boldsymbol{\Theta}}$.

E.2.1 Consistency of Estimator 1

Estimator 1 is given by:

$$\widehat{\widehat{\mathscr{H}}}_j^{\,1}\left(D_{m,n}^{\mathrm{S}_1}\right) = \frac{1}{m}\sum_{k=1}^{m}\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(k)}G^{(k)},n}\right), \tag{E.45}$$

where $D_{m,n}^{\mathrm{S}_1}$ is a nested Monte Carlo design generated according to Strategy (S1). To generate $D_{m,n}^{\mathrm{S}_1}$, proceed as follows.

1. Generate an m-sample $D_{\boldsymbol{\Lambda},m} = \left\{\boldsymbol{\Lambda}^{(k)}\right\}_{1\leq k\leq m}$ from $\mathbb{P}_{\boldsymbol{\Lambda}}$.
2. For each $\boldsymbol{\Lambda}^{(k)}$ in $D_{\boldsymbol{\Lambda},m}$:
   - Generate an n-sample $D_{\boldsymbol{\Theta}^{(k)},n} = \left\{\boldsymbol{\Theta}^{(kl)}\right\}_{1\leq l\leq n}$ from $\mathbb{P}_{\boldsymbol{\Theta}}$.
   - Compute the output values $G^{(kl)} = \widetilde{f}\left(\boldsymbol{\Theta}^{(kl)},\boldsymbol{\Lambda}^{(k)}\right)$ and store them in $D_{G^{(k)},n}$.
   - Merge $D_{\boldsymbol{\Theta}^{(k)},n}$ and $D_{G^{(k)},n}$ to form $D_{\boldsymbol{\Theta}^{(k)}G^{(k)},n}$.

This sampling process guarantees the following probabilistic relationships between the data:

$$D_{\Theta^{(1)},n} \perp\!\!\!\perp \ldots \perp\!\!\!\perp D_{\Theta^{(m)},n} \sim \left(\mathbb{P}_{\boldsymbol{\Theta}}\right)^{\otimes n} \tag{E.46}$$

$$D_{G^{(1)},n} \perp\!\!\!\perp \ldots \perp\!\!\!\perp D_{G^{(m)},n} \sim \mathbb{Q}_A \tag{E.47}$$

$$D_{\boldsymbol{\Theta}^{(1)}G^{(1)},n} \perp\!\!\!\perp \ldots \perp\!\!\!\perp D_{\boldsymbol{\Theta}^{(m)}G^{(m)},n} \sim \mathbb{Q}_B\,. \tag{E.48}$$

Above, $\mathbb{Q}_A$ denotes the joint distribution of the data contained in each sample $D_{G^{(k)},n}$. The corresponding cumulative distribution function (CDF) is given by:

$$\forall\,(g_1,\ldots,g_n)\in(\mathcal{D}_G)^n\,,\quad F_A(g_1,\ldots,g_n) = \int_{\mathcal{D}_{\boldsymbol{\Lambda}}} F_{G_{\boldsymbol{\lambda}}}(g_1)\ldots F_{G_{\boldsymbol{\lambda}}}(g_n)\,\mathrm{d}\mathbb{P}_{\boldsymbol{\Lambda}}(\boldsymbol{\lambda})\,, \tag{E.49}$$

where $F_{G_{\boldsymbol{\lambda}}}$ is the CDF of $G_{\boldsymbol{\lambda}} = \widetilde{f}(\boldsymbol{\Theta}, \boldsymbol{\lambda})$. Similarly, $\mathbb{Q}_B$ is the joint distribution of the data contained in each sample $D_{\boldsymbol{\Theta}^{(k)} G^{(k)},n}$. Its CDF is:

$$\forall \left((\boldsymbol{\theta}_1, g_1), \ldots, (\boldsymbol{\theta}_n, g_n) \right) \in (\mathcal{D}_{\boldsymbol{\Theta}} \times \mathcal{D}_G)^n ,$$
$$F_B\big((\boldsymbol{\theta}_1, g_1), \ldots, (\boldsymbol{\theta}_m, g_m) \big) = \int_{\mathcal{D}_{\boldsymbol{\Lambda}}} F_{\boldsymbol{\Theta}\, G_{\boldsymbol{\lambda}}}(\boldsymbol{\theta}_1, g_1) \ldots F_{\boldsymbol{\Theta}\, G_{\boldsymbol{\lambda}}}(\boldsymbol{\theta}_n, g_n)\, \mathrm{d}\mathbb{P}_{\boldsymbol{\Lambda}}(\boldsymbol{\lambda}) , \quad \text{(E.50)}$$

where $F_{\boldsymbol{\Theta}\, G_{\boldsymbol{\lambda}}}$ is the CDF of $(\boldsymbol{\Theta}, G_{\boldsymbol{\lambda}})$.

Proof. We first prove that $\widehat{\widehat{\mathscr{H}}}_j^{\,1}$ is unbiased.

$$\mathbb{E}\left[\widehat{\widehat{\mathscr{H}}}_j^{\,1}\right] = \frac{1}{m}\sum_{k=1}^{m} \mathbb{E}\left[\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(k)} G^{(k)},n}\right)\right] \quad \text{(E.51)}$$
$$= \mathbb{E}\left[\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G^{(1)},n}\right)\right] \text{ with equation (E.48)} \quad \text{(E.52)}$$
$$= \mathbb{E}_{\boldsymbol{\Lambda}^{(1)}}\left[\mathbb{E}\left[\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G^{(1)},n}\right)\middle|\boldsymbol{\Lambda}^{(1)}\right]\right] \text{ with the tower property.} \quad \text{(E.53)}$$

In particular, the conditional expectation can be computed as follows:

$$\mathbb{E}\left[\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G^{(1)},n}\right)\middle|\boldsymbol{\Lambda}^{(1)} = \boldsymbol{\lambda}\right] = \mathbb{E}\left[\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G_{\boldsymbol{\lambda}},n}\right)\right] = H_j(\boldsymbol{\lambda}) . \quad \text{(E.54)}$$

The first equality holds because $\boldsymbol{\Lambda}^{(1)} \perp\!\!\!\perp \boldsymbol{\Theta}^{(1)}$. The second equality is obtained by recalling that a U-statistic is unbiased. Then, injecting equation (E.54) in equation (E.53) gives:

$$\mathbb{E}\left[\widehat{\widehat{\mathscr{H}}}_j^{\,1}\right] = \mathbb{E}_{\boldsymbol{\Lambda}}\left[H_j(\boldsymbol{\Lambda}^{(1)})\right] = \mathscr{H}_j , \quad \text{(E.55)}$$

which confirms that Estimator 1 is unbiased.

Now, let us prove that this estimator is consistent. To this end, we prove that Estimator 1 converges in mean square to $\mathscr{H}_j$. In fact, the derivations are simpler and provide a convergence rate in the process. The mean square error (MSE) can be written as:

$$\mathrm{MSE}\left(\widehat{\widehat{\mathscr{H}}}_j^{\,1}\right) = \mathbb{E}\left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,1} - \mathscr{H}_j\right)^2\right] = \mathbb{V}\left(\widehat{\widehat{\mathscr{H}}}_j^{\,1}\right) \text{ as } \widehat{\widehat{\mathscr{H}}}_j^{\,1} \text{ is unbiased} \quad \text{(E.56)}$$
$$= \frac{1}{m}\mathbb{V}\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G^{(1)},n}\right)\right) \text{ with equation (E.48)} \quad \text{(E.57)}$$

The total variance formula yields:

$$\mathbb{V}\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G^{(1)},n}\right)\right) = \mathbb{V}_{\boldsymbol{\Lambda}^{(1)}}\left(\mathbb{E}\left[\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G^{(1)},n}\right)\middle|\boldsymbol{\Lambda}^{(1)}\right]\right) + \mathbb{E}_{\boldsymbol{\Lambda}^{(1)}}\left[\mathbb{V}\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G^{(1)},n}\right)\middle|\boldsymbol{\Lambda}^{(1)}\right)\right] . \quad \text{(E.58)}$$

On the one hand, the bound provided in equation (E.39) can be used in equation (E.54) to obtain:

$$\mathbb{V}_{\boldsymbol{\Lambda}^{(1)}}\left(\mathbb{E}\left[\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G^{(1)},n}\right)\middle|\boldsymbol{\Lambda}^{(1)}\right]\right) = \mathbb{V}_{\boldsymbol{\Lambda}^{(1)}}\left(H_j(\boldsymbol{\Lambda}^{(1)})\right) \leq \left(4\, M_{\theta_j}\, M_g\right)^2 . \quad \text{(E.59)}$$

On the other hand, the conditional variance can be bounded using Corollary E.2:

$$\mathbb{V}\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G^{(1)},n}\right)\middle|\boldsymbol{\Lambda}^{(1)} = \boldsymbol{\lambda}\right) = \mathbb{V}\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)} G_{\boldsymbol{\lambda}},n}\right)\right) \leq \frac{\alpha}{n} , \quad \text{(E.60)}$$

which then yields:

$$\mathbb{E}_{\mathbf{\Lambda}^{(1)}}\left[\mathbb{V}\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\mathbf{\Theta}^{(1)}G^{(1)},n}\right)\middle|\mathbf{\Lambda}^{(1)}\right)\right] \leq \frac{\alpha}{n}. \tag{E.61}$$

Combining equations (E.59) and (E.61) leads to:

$$\mathrm{MSE}\left(\widehat{\widehat{\mathscr{H}}}_j^{\,1}\right) \leq \frac{1}{m}\left(\left(4\,M_x\,M_y\right)^2 + \frac{\alpha}{n}\right) = \mathcal{O}\left(\frac{1}{m}\right). \tag{E.62}$$

Therefore, Estimator 1 converges in mean square to $\mathscr{H}_j$, with the MSE decaying at a rate of $1/m$. □

E.2.2 Consistency of Estimator 2

Estimator 2 is given by:

$$\widehat{\widehat{\mathscr{H}}}_j^{\,2}\left(D_{m,n}^{\mathrm{S}_1}\right) = \frac{1}{m}\sum_{k=1}^{m}\widehat{H}_j^{\mathrm{V}}\left(D_{\mathbf{\Theta}^{(k)}G^{(k)},n}\right), \tag{E.63}$$

where $D_{m,n}^{\mathrm{S}_1}$ is a nested Monte Carlo design generated according to Strategy (S1). Compared to Estimator 1, the only difference lies in the use of V-statistics instead of U-statistics to estimate HSIC indices. Since V-statistics are not unbiased in general, Estimator 2 is typically biased. However, because the experimental design remains $D_{m,n}^{\mathrm{S}_1}$, the probabilistic relationships among the data highlighted in equations (E.46)–(E.48) still hold.

Proof. For Estimator 2, the proof strategy relies on a decomposition of the MSE obtained by introducing Estimator 1 as an intermediate quantity. Using the classical inequality $(a+b)^2 \leq 2\left(a^2+b^2\right)$, the MSE of Estimator 2 can be bounded as follows:

$$\frac{1}{2}\,\mathrm{MSE}\left(\widehat{\widehat{\mathscr{H}}}_j^{\,2}\right) = \frac{1}{2}\,\mathbb{E}\left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,2} - \mathscr{H}_j\right)^2\right] = \frac{1}{2}\,\mathbb{E}\left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,2} - \widehat{\widehat{\mathscr{H}}}_j^{\,1} + \widehat{\widehat{\mathscr{H}}}_j^{\,1} - \mathscr{H}_j\right)^2\right] \tag{E.64}$$

$$\leq \mathbb{E}\left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,2} - \widehat{\widehat{\mathscr{H}}}_j^{\,1}\right)^2\right] + \mathbb{E}\left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,1} - \mathscr{H}_j\right)^2\right]. \tag{E.65}$$

The second term is the MSE of Estimator 1, and it was established in equation (E.62) that:

$$\mathbb{E}\left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,1} - \mathscr{H}_j\right)^2\right] = \mathcal{O}\left(\frac{1}{m}\right). \tag{E.66}$$

Regarding the first term, it is the mean squared difference between Estimators 1 and 2. Note that the difference may be written as:

$$\widehat{\widehat{\mathscr{H}}}_j^{\,2} - \widehat{\widehat{\mathscr{H}}}_j^{\,1} = \frac{1}{m}\sum_{k=1}^{m}\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\mathbf{\Theta}^{(k)}G^{(k)},n}\right) - \widehat{H}_j^{\mathrm{U}}\left(D_{\mathbf{\Theta}^{(k)}G^{(k)},n}\right)\right) = \frac{1}{m}\sum_{k=1}^{m}\Delta_{jk}\,. \tag{E.67}$$

Equation (E.48) indicates that the datasets $\left\{D_{\mathbf{\Theta}^{(k)}G^{(k)},n}\right\}_{1\leq k\leq m}$ form an i.i.d. sample from $\mathbb{Q}_B$. As a consequence, the variables $\{\Delta_{jk}\}_{1\leq k\leq m}$ also form an i.i.d. sample as well. Applying Lemma E.1 to this sample, we obtain:

$$\mathbb{E}\left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,2} - \widehat{\widehat{\mathscr{H}}}_j^{\,1}\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{m}\sum_{k=1}^{m}\Delta_{jk}\right)^2\right] \leq \mathbb{E}\left[\Delta_{1j}^2\right]. \tag{E.68}$$

The expectation on the right-hand side can reformulated as:

$$\mathbb{E}\left[\Delta_{1j}^2\right] = \mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}^{(1)}G^{(1)},n}\right) - \widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)}G^{(1)},n}\right)\right)^2\right] \tag{E.69}$$

$$= \mathbb{E}_{\boldsymbol{\Lambda}^{(1)}}\left[\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}^{(1)}G^{(1)},n}\right) - \widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)}G^{(1)},n}\right)\right)^2 \middle| \boldsymbol{\Lambda}^{(1)}\right]\right], \tag{E.70}$$

and Corollary E.2 (with $r = 2$) can be used to bound the conditional moment:

$$\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}^{(1)}G^{(1)},n}\right) - \widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)}G^{(1)},n}\right)\right)^2 \middle| \boldsymbol{\Lambda}^{(1)} = \boldsymbol{\lambda}\right]$$

$$= \mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}^{(1)}G_{\boldsymbol{\lambda}},n}\right) - \widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}^{(1)}G_{\boldsymbol{\lambda}},n}\right)\right)^2\right] \leq \frac{\beta_2}{n^2} = \mathcal{O}\left(\frac{1}{n^2}\right). \tag{E.71}$$

Since the constant β_2 does not depend on $\boldsymbol{\lambda}$, the converge rate remains unchanged after taking the expectation with respect to $\boldsymbol{\Lambda}^{(1)}$, which gives $\mathbb{E}\left[\Delta_{j1}^2\right] = \mathcal{O}(1/n^2)$. Finally, putting together the elements of equations (E.65), (E.66), (E.68) and (E.71) leads to:

$$\mathrm{MSE}\left(\widehat{\widehat{\mathscr{H}}}_j^{\,2}\right) = \mathcal{O}\left(\frac{1}{m} + \frac{1}{n^2}\right). \tag{E.72}$$

□

E.2.3 Consistency of Estimator 3

The consistency proofs for Estimators 3 and 4 follow the same line of reasoning as those for Estimators 1 and 2. The MSE is split into two well-chosen expectation terms. The tower property of conditional expectation is then used to handle the two levels of uncertainty separately.

Estimator 3 is given by:

$$\widehat{\widehat{\mathscr{H}}}_j^{\,3}\left(D_{m,n}^{\mathrm{S}_2}\right) = \frac{1}{m}\sum_{k=1}^{m}\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G^{(k)},n}\right), \tag{E.73}$$

where $D_{m,n}^{\mathrm{S}_2}$ is a nested Monte Carlo design generated according to Strategy (S2). To generate $D_{m,n}^{\mathrm{S}_2}$, proceed as follows.

1. Generate an m-sample $D_{\boldsymbol{\Lambda},m} = \left\{\boldsymbol{\Lambda}^{(k)}\right\}_{1\leq k\leq m}$ from $\mathbb{P}_{\boldsymbol{\Lambda}}$.
2. Generate an n-sample $D_{\boldsymbol{\Theta},n} = \left\{\boldsymbol{\Theta}^{(l)}\right\}_{1\leq l\leq n}$ from $\mathbb{P}_{\boldsymbol{\Theta}}$.
3. For each $\boldsymbol{\Lambda}^{(k)}$ in $D_{\boldsymbol{\Lambda},m}$:
   - Compute the output values $G^{(kl)} = \widetilde{f}\left(\boldsymbol{\Theta}^{(l)}, \boldsymbol{\Lambda}^{(k)}\right)$ and store them in $D_{G^{(k)},n}$.
   - Merge $D_{\boldsymbol{\Theta},n}$ and $D_{G^{(k)},n}$ to form $D_{\boldsymbol{\Theta}G^{(k)},n}$.

The same design $D_{\boldsymbol{\Theta},n}$ provides the input observations of $\boldsymbol{\Theta}$ used to construct all output designs $D_{G^{(k)},n}$. Therefore, these designs are no longer independent, which constitutes a major difference from the data generated under (S1). However, they still share the same distribution. More precisely, using the notations introduced in equations (E.47) and (E.48), one can write:

$$\forall\, 1 \leq k \leq m, \quad D_{G^{(k)},n} \sim \mathbb{Q}_A \quad \text{and} \quad D_{\boldsymbol{\Theta}G^{(k)},n} \sim \mathbb{Q}_B\ . \tag{E.74}$$

Similarly, all pairs $\left(\boldsymbol{\Lambda}^{(k)}, D_{\boldsymbol{\Theta}G^{(k)},n}\right)$ share the same distribution $\mathbb{Q}_C$, yet are not independent for the same reason as explained above. The distribution $\mathbb{Q}_C$ can be defined as the push-forward measure of $\mathbb{P}_{\boldsymbol{\Lambda}} \otimes (\mathbb{P}_{\boldsymbol{\Theta}})^{\otimes n}$ through the measurable

mapping:

$$\begin{array}{rccccccc} \psi: & \mathcal{D}_{\boldsymbol{\Lambda}} & \times & (\mathcal{D}_{\boldsymbol{\Theta}})^{\otimes n} & \longrightarrow & \mathcal{D}_{\boldsymbol{\Lambda}} & \times & (\mathcal{D}_{\boldsymbol{\Theta}})^{\otimes n} \quad \times \quad (\mathcal{D}_G)^{\otimes n} \\ & \left(\boldsymbol{\lambda} \right. & , & \left. (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) \right) & \longmapsto & \left(\boldsymbol{\lambda} \right. & , & \left. (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) \right. \quad , \quad \left. \left(\widetilde{f}(\boldsymbol{\lambda}, \boldsymbol{\theta}_1), \dots, \widetilde{f}(\boldsymbol{\lambda}, \boldsymbol{\theta}_1) \right) \right) . \end{array} \tag{E.75}$$

Therefore, in the same spirit as equation (E.74), one can write:

$$\forall 1 \leq k \leq m, \quad \left(\boldsymbol{\Lambda}^{(k)}, D_{\boldsymbol{\Theta} G^{(k)}, n} \right) \sim \mathbb{Q}_C \ . \tag{E.76}$$

Proof. To establish the unbiasedness of Estimator 3, we proceed exactly as for Estimator 1. The only difference is that equation (E.48) can no longer be invoked because it is not satisfied by the data generated under (S2). However, equation (E.74) suffices. In fact, independence between the output samples $D_{G^{(k)},n}$ is not required to prove unbiasedness.

It remains to show consistency. For Estimator 3, the quantity used to split the MSE is the pseudo-estimator $\widehat{\mathscr{H}}_j$ from equation (4.16):

$$\widehat{\mathscr{H}}_j := \widehat{\mathscr{H}}_j \left(D_{\boldsymbol{\Lambda},m} \right) = \frac{1}{m} \sum_{k=1}^{m} H_j \left(\boldsymbol{\Lambda}^{(k)} \right) \quad \text{with} \quad H_j \left(\boldsymbol{\Lambda}^{(k)} \right) = \text{HSIC} \left(\Theta_j, G^{(k)} \mid \boldsymbol{\Lambda}^{(k)} \right) . \tag{E.77}$$

It is useful from a theoretical perspective, although it can be computed in practice since the theoretical values of HSIC indices are not tractable. The MSE can therefore be split as follows:

$$\frac{1}{2} \text{MSE} \left(\widehat{\widehat{\mathscr{H}}}_j^{\,3} \right) = \frac{1}{2} \mathbb{E} \left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,3} - \mathscr{H}_j \right)^2 \right] = \frac{1}{2} \mathbb{E} \left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,3} - \widehat{\mathscr{H}}_j + \widehat{\mathscr{H}}_j - \mathscr{H}_j \right)^2 \right] \tag{E.78}$$

$$\leq \mathbb{E} \left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,3} - \widehat{\mathscr{H}}_j \right)^2 \right] + \mathbb{E} \left[\left(\widehat{\mathscr{H}}_j - \mathscr{H}_j \right)^2 \right] . \tag{E.79}$$

By convexity of the square function, the quantity inside the first expectation satisfies:

$$\left(\widehat{\widehat{\mathscr{H}}}_j^{\,3} - \widehat{H}_j \right)^2 = \left(\frac{1}{m} \sum_{k=1}^{m} \left(\widehat{H}_j^{\mathrm{U}} \left(D_{\boldsymbol{\Theta} G^{(k)}, n} \right) - H_j \left(\boldsymbol{\Lambda}^{(k)} \right) \right) \right)^2 \tag{E.80}$$

$$\leq \frac{1}{m} \sum_{k=1}^{m} \left(\widehat{H}_j^{\mathrm{U}} \left(D_{\boldsymbol{\Theta} G^{(k)}, n} \right) - H_j \left(\boldsymbol{\Lambda}^{(k)} \right) \right)^2 . \tag{E.81}$$

Then, taking expectations gives:

$$\mathbb{E} \left[\left(\widehat{\widehat{\mathscr{H}}}_j^{\,3} - \widehat{H}_j \right)^2 \right] \leq \frac{1}{m} \sum_{k=1}^{m} \mathbb{E} \left[\left(\widehat{H}_j^{\mathrm{U}} \left(D_{\boldsymbol{\Theta} G^{(k)}, n} \right) - H_j \left(\boldsymbol{\Lambda}^{(k)} \right) \right)^2 \right] \tag{E.82}$$

$$= \mathbb{E} \left[\left(\widehat{H}_j^{\mathrm{U}} \left(D_{\boldsymbol{\Theta} G^{(1)}, n} \right) - H_j \left(\boldsymbol{\Lambda}^{(1)} \right) \right)^2 \right] \tag{E.83}$$

$$= \mathbb{E}_{\boldsymbol{\Lambda}^{(1)}} \left[\mathbb{E} \left[\left(\widehat{H}_j^{\mathrm{U}} \left(D_{\boldsymbol{\Theta} G^{(1)}, n} \right) - H_j \left(\boldsymbol{\Lambda}^{(1)} \right) \right)^2 \middle| \boldsymbol{\Lambda}^{(1)} \right] \right] . \tag{E.84}$$

Equation (E.83) is a consequence of all terms in the sum sharing the same distribution $\mathbb{Q}_C$, as stated in equation (E.76). The conditional expectation in equation (E.84) can be bounded using the simple arguments introduced earlier ($\boldsymbol{\Lambda}^{(1)} \perp\!\!\!\perp \boldsymbol{\Theta}$,

the unbiasedness of U-statistics, and Corollary E.2):

$$\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G^{(1)},n}\right)-H_j\left(\boldsymbol{\Lambda}^{(1)}\right)\right)^2\middle|\boldsymbol{\Lambda}^{(1)}=\boldsymbol{\lambda}\right]=\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)-H_j(\boldsymbol{\lambda})\right)^2\right] \tag{E.85}$$

$$=\mathbb{V}\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)\right) \tag{E.86}$$

$$\leq\frac{\alpha}{n}\,. \tag{E.87}$$

Since the constant α does not depend on $\boldsymbol{\lambda}$, taking expectations with respect to $\boldsymbol{\Lambda}^{(1)}$ provides:

$$\mathbb{E}\left[\left(\widehat{\widehat{\mathscr{H}_j}}^{3}-\widehat{\mathscr{H}_j}\right)^2\right]=\mathcal{O}\left(\frac{1}{n}\right). \tag{E.88}$$

The Big-O rate associated with the second expectation in equation (E.78) can easily be derived:

$$\mathbb{E}\left[\left(\widehat{\mathscr{H}_j}-\mathscr{H}_j\right)^2\right]=\mathbb{V}\left(\widehat{\mathscr{H}_j}\right)=\frac{1}{m}\mathbb{V}_{\boldsymbol{\Lambda}^{(1)}}\left(H_j\left(\boldsymbol{\Lambda}^{(1)}\right)\right)\leq\frac{1}{m}\left(4\,M_x\,M_y\right)^2=\mathcal{O}\left(\frac{1}{m}\right). \tag{E.89}$$

The first equality holds because $\mathbb{E}\left[\widehat{\mathscr{H}_j}\right]=\mathscr{H}_j$. Moreover, the bound on the variance of $H_j\left(\boldsymbol{\Lambda}^{(1)}\right)$ follows from equation (E.39). Finally, combining equations (E.88) and (E.89) yields:

$$\mathrm{MSE}\left(\widehat{\widehat{\mathscr{H}_j}}^{3}\right)=\mathcal{O}\left(\frac{1}{m}+\frac{1}{n}\right). \tag{E.90}$$

□

E.2.4 Consistency of Estimator 4

Estimator 4 is given by:

$$\widehat{\widehat{\mathscr{H}_j}}^{4}\left(D_{m,n}^{\mathrm{S}_2}\right)=\frac{1}{m}\sum_{k=1}^{m}\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}^{(k)}G^{(k)},n}\right), \tag{E.91}$$

where $D_{m,n}^{\mathrm{S}_2}$ is a nested Monte Carlo design generated according to Strategy (S2). Compared to Estimator 3, the only difference lies in the use of V-statistics, which typically introduces bias. The probabilistic relationships among the data highlighted in equations (E.74) and (E.75) remain valid.

Proof. Up to equation (E.85), we proceed exactly as for Estimator 3. The analogous expression is:

$$\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}G^{(1)},n}\right)-H_j\left(\boldsymbol{\Lambda}^{(1)}\right)\right)^2\middle|\boldsymbol{\Lambda}^{(1)}=\boldsymbol{\lambda}\right]=\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)-H_j(\boldsymbol{\lambda})\right)^2\right]. \tag{E.92}$$

The expectation on the right-hand side can be split as follows:

$$\frac{1}{2}\,\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)-H_j(\boldsymbol{\lambda})\right)^2\right]$$

$$\leq\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)-H_j(\boldsymbol{\lambda})\right)^2\right]+\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)-\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)\right)^2\right]. \tag{E.93}$$

Bounds for these two terms are provided directly by Corollary E.2:

$$\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)-H_j(\boldsymbol{\lambda})\right)^2\right]=\mathbb{V}\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)\right)\leq\frac{\alpha}{n} \tag{E.94}$$

$$\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)-\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G_{\boldsymbol{\lambda}},n}\right)\right)^2\right]\leq\frac{\beta_2}{n^2}\,. \tag{E.95}$$

Reassembling the full line of reasoning, we arrive at:

$$\begin{aligned}\frac{1}{4}\,\mathrm{MSE}\left(\widehat{\widehat{\mathscr{H}}_j}^4\right)\leq{}&\mathbb{E}_{\boldsymbol{\Lambda}^{(1)}}\left[\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G^{(1)},n}\right)-H_j\left(\boldsymbol{\Lambda}^{(1)}\right)\right)^2\middle|\boldsymbol{\Lambda}^{(1)}\right]\right]\\&+\mathbb{E}_{\boldsymbol{\Lambda}^{(1)}}\left[\mathbb{E}\left[\left(\widehat{H}_j^{\mathrm{V}}\left(D_{\boldsymbol{\Theta}G^{(1)},n}\right)-\widehat{H}_j^{\mathrm{U}}\left(D_{\boldsymbol{\Theta}G^{(1)},n}\right)\right)^2\middle|\boldsymbol{\Lambda}^{(1)}\right]\right]\\&+\mathbb{E}\left[\left(\widehat{\mathscr{H}_j}-\mathscr{H}_j\right)^2\right].\end{aligned} \tag{E.96}$$

Finally, combining the Big-O rates obtained for all three expectations yields:

$$\mathrm{MSE}\left(\widehat{\widehat{\mathscr{H}}_j}^4\right)=\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)+\mathcal{O}\left(\frac{1}{m}\right)=\mathcal{O}\left(\frac{1}{m}+\frac{1}{n}\right). \tag{E.97}$$

□