Kernel-based ANOVA decomposition and Shapley effects - Application to global sensitivity analysis

Sébastien da Veiga

Kernel-based ANOVA decomposition and Shapley effects – Application to global sensitivity analysis

Sébastien Da Veiga

Safran Tech, Modeling & Simulation
Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France

Abstract

Global sensitivity analysis is the main quantitative technique for identifying the most influential input variables in a numerical simulation model. In particular when the inputs are independent, Sobol' sensitivity indices attribute a portion of the output of interest variance to each input and all possible interactions in the model, thanks to a functional ANOVA decomposition. On the other hand, moment-independent sensitivity indices focus on the impact of input variables on the whole output distribution instead of the variance only, thus providing complementary insight on the inputs / output relationship. Unfortunately they do not enjoy the nice decomposition property of Sobol' indices and are consequently harder to analyze. In this paper, we introduce two moment-independent indices based on kernel-embeddings of probability distributions and show that the RKHS framework used for their definition makes it possible to exhibit a kernel-based ANOVA decomposition. This is the first time such a desirable property is proved for sensitivity indices apart from Sobol' ones. When the inputs are dependent, we also use these new sensitivity indices as building blocks to design kernel-embedding Shapley effects which generalize the traditional variance-based ones used in sensitivity analysis. Several estimation procedures are discussed and illustrated on test cases with various output types such as categorical variables and probability distributions. All these examples show their potential for enhancing traditional sensitivity analysis with a kernel point of view.

1 Introduction

In the computer experiments community, global sensitivity analysis (GSA) has now emerged as a central tool for exploring the inputs/outputs relationship of a numerical simulation model. Starting from the pioneering work of Sobol (Sobol', 1993) and Saltelli (Saltelli et al., 1999) on the interpretation and estimation of Sobol' indices, the last two decades have been a fertile ground for the development of advanced statistical methodologies and extensions of original Sobol' indices: new estimation procedures (Da Veiga et al. (2009), Da Veiga and Gamboa (2013), Solís (2019), Gamboa et al. (2020)), multivariate outputs with aggregation (Gamboa et al., 2013) and dimensionality reduction (Lamboni et al., 2011), goal-oriented sensitivity analysis (Fort et al., 2016) or moment-independent sensitivity measures (Borgonovo (2007), Da Veiga (2015)), among others. At the heart of the popularity of Sobol' indices is the fundamental functional analysis of variance (ANOVA) decomposition, which opens the path for their interpretation as parts of the output variance and makes it possible to pull apart the input main effects and all their potential interactions, up to their whole influence measured by total Sobol' indices. This decomposition however has two

drawbacks. First, it is only valid when the inputs are independent, although some generalizations were investigated (Chastaing et al., 2012). Secondly, it only concerns the original Sobol' indices, meaning that it is not possible to split the input effects with goal-oriented or moment-independent sensitivity analysis in general.

When the inputs are dependent, total Sobol' indices can still be used to discriminate them when the objective is to build a surrogate model of the system, and other Sobol'-related indices have also been proposed for interpretability (Mara et al., 2015). But the major breakthrough happened when Shapley effects have been defined for GSA by Owen (Owen, 2014). Indeed due to their mathematical foundations from game theory, Shapley effects do not require the independence assumption to enjoy nice properties: each input is assigned a Shapley effect lying between 0 and 1, while the sum of all effects is equal to 1. For a given input all interactions and correlations with other ones are mixed up, but the interpretation as parts of the output variance is kept and input rankings are still sensible. For these reasons Shapley effects are now commonly thought as central importance measures in GSA for dealing with dependence, and their estimation has been thoroughly investigated recently (Song et al., 2016; Iooss and Prieur, 2019; Broto et al., 2020; Plischke et al., 2020).

From an interpretability perspective, other importance measures introduced in the context of goal-oriented and moment-independent sensitivity analysis have proven useful to gain additional insights on a given model. For example quantile-oriented (Fort et al., 2016; Maume-Deschamps and Niang, 2018) or reliability-based measures (Ditlevsen and Madsen, 1996) can help understand which inputs lead to the failure of the system, while optimization-related indices enable dimension reduction for optimization problems (Spagnol et al., 2019). On the other hand, moment-independent sensitivity indices, which quantify the input impact on the whole output distribution instead of the variance only, are powerful complementary tools to grasp further types of input influence. Among them are the f-divergence indices (Da Veiga (2015), Rahman (2016)) with particular cases corresponding to the sensitivity index introduced by Borgonovo (Borgonovo, 2007) and the class of kernel-based sensitivity indices, which rely on the embedding of probability distributions in reproducing kernel Hilbert spaces (RKHS) (Da Veiga, 2015, 2016). Unfortunately an ANOVA-like decomposition is not available yet for any of these indices even in the independent setting: as a consequence this limits the interpretation of their formulation for interactions since without ANOVA it is not possible to remove the main effects, and at the same time the natural normalization constant (equivalent to the total output variance for Sobol' indices) is not known.

In this paper we focus on a general RKHS framework for GSA and prove that an ANOVA decomposition actually exists for two previously introduced kernel-based sensitivity indices in the independent setting. To the best of our knowledge this is the first time such a decomposition is available for other sensitivity indices other than the original Sobol' ones. Not only this makes it possible to properly define higher-order indices, but this further gives access to their natural normalization constant. We also demonstrate that these measures are generalizations of Sobol' indices, in the sense that they are recovered with specific kernels. But the RKHS point of view additionally comes with a large body of work on several kernels specifically designed for particular target applications, such as multivariate, functional, categorical or time-series case studies, thus defining a unified framework for many real GSA test cases. When inputs are not independent, we finally introduce a kernel-based version of Shapley effects similar to the ones proposed by Owen.

The paper is organized as follows. Section 2 first briefly introduces the traditional functional

ANOVA decomposition with Sobol' indices and moment-independent indices. In Section 3 we then discuss the elementary tools from RKHS theory needed to build kernel-based sensitivity indices which are at the core of this work. We further investigate these indices and prove they also arise from an ANOVA decomposition. In addition we define Shapley effects with kernels and the benefits of the RKHS framework for GSA are studied through several examples. Several estimation procedures are then discussed in Section 4, where we generalize some of the recent estimators for Sobol' indices. Finally, Section 5 illustrates the potential of these sensitivity indices with various numerical experiments corresponding to typical GSA applications.

2 Global sensitivity analysis

Let $\eta : \mathcal{X}_1 \times \ldots \times \mathcal{X}_d \to \mathcal{Y}$ denote the numerical simulation model, which is a function of d input variables $X_l \in \mathcal{X}_l$, $l = 1, \ldots, d$, and $Y \in \mathcal{Y}$ the model output given by $Y = \eta(X_1, \ldots, X_d)$. In standard GSA the inputs X_l are further assumed to be independent with known probability distributions P_{X_l}, meaning that the vector of inputs $\mathbf{X} = (X_1, \ldots, X_d)$ is distributed as $\mathrm{P}_{\mathbf{X}} = \mathrm{P}_{X_1} \otimes \ldots \otimes \mathrm{P}_{X_d}$. For any subset $A = \{l_1, \ldots, l_{|A|}\} \in \mathcal{P}_d$ of indices taken from $\{1, \ldots, d\}$ we denote $\mathbf{X}_A = \left(X_{l_1}, \ldots, X_{l_{|A|}}\right) \in \mathcal{X}_A = \mathcal{X}_{l_1} \times \ldots \times \mathcal{X}_{l_{|A|}}$ the vector of inputs with indices in A and $\mathbf{X}_{-A}$ the complementary vector with indices not in A. In this setting, the main objective of global sensitivity analysis is to quantify the impact of any group of input variables $\mathbf{X}_A$ on the model output Y.
In this section we first recall the functional ANOVA decomposition and the definition of Sobol' indices, which fall into the category of *variance-based* indices. Sensitivity indices that account for the whole output distribution, referred to as *moment-independent* indices, are then discussed. Note that in the following, we adopt the notation S for a properly normalized sensitivity index, while $\mathcal{S}$ will stand for an unnormalized index, where normalization is to be understood as an end result from an ANOVA-like decomposition.

2.1 ANOVA decomposition and variance-based sensitivity indices

Here we first assume that $Y \in \mathcal{Y} \subset \mathbb{R}$ is a square integrable scalar output. If the inputs are independent, the function η can then be decomposed according to the ANOVA decomposition:

Theorem 1 (ANOVA decomposition (Hoeffding, 1948; Antoniadis, 1984))**.** *Assume that $\eta : \mathcal{X}_1 \times \ldots \times \mathcal{X}_d \to \mathcal{Y}$ is a square integrable function of d independent random variables $X_1, \ldots, X_d$. Then η admits a decomposition*

$$Y = \eta(X_1, \ldots, X_d) = \sum_{A \subseteq \mathcal{P}_d} \eta_A(\mathbf{X}_A),$$

with η_A depending only on the variables $\mathbf{X}_A$ and satisfying

(a) $\eta_\emptyset = \mathbb{E}(Y)$,

(b) $\mathbb{E}_{X_l}(\eta_A(\mathbf{X}_A)) = 0$ *if* $l \in A$,

(c) $\eta_A(\mathbf{X}_A) = \sum_{B \subset A} (-1)^{|A|-|B|} \mathbb{E}(Y|\mathbf{X}_B)$.

Furthermore, (b) implies that all the terms η_A in the decomposition are mutually orthogonal. As a consequence, the output variance can be decomposed as

$$\operatorname{Var} Y = \sum_{A \subseteq \mathcal{P}_d} \operatorname{Var} \eta_A(\mathbf{X}_A) = \sum_{A \subseteq \mathcal{P}_d} V_A \tag{1}$$

where

$$V_A = \sum_{B \subset A} (-1)^{|A|-|B|} \mathrm{Var}\, \mathbb{E}(Y|\mathbf{X}_B). \tag{2}$$

When this decomposition holds, it is then straightforward to quantify the influence of any subset of inputs $\mathbf{X}_A$ on the output variance by normalizing each term with $\mathrm{Var}\, Y$.

Definition 1 (Sobol' indices (Sobol', 1993))**.** *Under the same assumptions of Theorem 1, the Sobol' sensitivity index associated to a subset A of input variables is defined as*

$$S_A = \frac{V_A}{\mathrm{Var}\, Y}, \tag{3}$$

while the total Sobol' index associated to A is

$$S_A^T = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B. \tag{4}$$

In particular, the first-order Sobol' index of an input X_l writes

$$S_l = \frac{\mathrm{Var}\, \mathbb{E}(Y|X_l)}{\mathrm{Var}\, Y}$$

and its total Sobol' index is given by

$$S_l^T = \sum_{B \subseteq \mathcal{P}_d,\, l \in B} S_B = 1 - \frac{\mathrm{Var}\, \mathbb{E}(Y|\mathbf{X}_{-l})}{\mathrm{Var}\, Y}.$$

Finally, the ANOVA decomposition (1) readily provides an interpretation of Sobol' indices as a percentage of explained output variance, i.e.

$$\sum_{A \subseteq \mathcal{P}_d} S_A = 1. \tag{5}$$

With these definitions, the impact of each input variable can be quantitatively assessed: the first-order Sobol' index measures the main effect of an input, while the total Sobol' index aggregates all its potential interactions with other inputs. As an illustration, an input variable with low total Sobol' index is thus unimportant and one can freeze it at a default value. When for a given input both first-order and total Sobol' indices are close, this means that this input does not have interactions, while a large gap indicates strong interactions in the model. Furthermore, due to the summation property (5), the interpretation of Sobol' indices as shares of the output variance is an efficient tool for practitioners who aim at understanding precisely the impact and interactions of the inputs of a model on the output. For example the interaction of two inputs X_l and $X_{l'}$ writes

$$S_{ll'} = \frac{\mathrm{Var}\, \mathbb{E}(Y|X_l,\, X_{l'}) - \mathrm{Var}\, \mathbb{E}(Y|X_l) - \mathrm{Var}\, \mathbb{E}(Y|X_{l'})}{\mathrm{Var}\, Y} = \frac{\mathrm{Var}\, \mathbb{E}(Y|X_l,\, X_{l'})}{\mathrm{Var}\, Y} - S_l - S_{l'}. \tag{6}$$

Note that to compute this interaction one subtracts the first-order indices S_l and $S_{l'}$ from the sensitivity index of the subset $(X_l, X_{l'})$ in order to remove the main effects and highlight the interaction only.

2.2 Moment-independent sensitivity indices

Despite their appealing properties, Sobol' indices rank the input variables according to their impact on the output variance only. In a parallel line of work, several authors proposed to investigate instead how inputs influence the whole output distribution, thus introducing a different insight on the inputs/outputs relationship. The starting point (Baucells and Borgonovo (2013), Da Veiga (2015)) is to consider that a given input X_l is important in the model if the probability distribution P_Y of the output changes when X_l is fixed, *i.e.* if the conditional probability distribution $\mathrm{P}_{Y|X_l}$ is different from P_Y. More precisely, if $d(\cdot,\cdot)$ denotes a dissimilarity measure between probability distributions, one can define a sensitivity index for variable X_l given by

$$\mathcal{S}_l = \mathbb{E}_{X_l}\left(d(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right). \tag{7}$$

Such a general formulation is flexible, in the sense that many choices for $d(\cdot,\cdot)$ are available. As an illustration, it is straightforward to show that the unnormalized first-order Sobol' index is retrieved with the naive dissimilarity measure $d(\mathrm{P},\mathrm{Q}) = (\mathbb{E}_{\xi\sim\mathrm{P}}(\xi) - \mathbb{E}_{\xi\sim\mathrm{Q}}(\xi))^2$, which compares probability distributions only through their means. A large class of dissimilarity measures is also given by the so-called *f-divergence* family: assuming that (X_l, Y) has an absolute continuous distribution with respect to the Lebesgue measure on $\mathbb{R}^2$, the f-divergence between P_Y and $\mathrm{P}_{Y|X_l}$ is

$$d_f(\mathrm{P}_Y, \mathrm{P}_{Y|X_l}) = \int f\left(\frac{p_Y(y)}{p_{Y|X_l}(y)}\right) p_{Y|X_l}(y)dy$$

where f is a convex function such that $f(1) = 0$ and p_Y and $p_{Y|X_l}$ are the probability distribution functions of Y and $Y|X_l$, respectively. The corresponding sensitivity index is then

$$\mathcal{S}_l^f = \int f\left(\frac{p_Y(y)p_{X_l}(x)}{p_{X_l,Y}(x,y)}\right) p_{X_l,Y}(x,y)dxdy$$

with p_{X_l} and $p_{X_l,Y}$ the probability distribution functions of X_l and (X_l, Y), respectively. This index has been studied for example in Da Veiga (2015) and Rahman (2016). A notable special case is obtained with the total-variation distance corresponding to $f(t) = |t-1|$, leading to the sensitivity index proposed by Borgonovo (Borgonovo, 2007):

$$\mathcal{S}_l^{TV} = \int |p_Y(y)p_{X_l}(x) - p_{X_l,Y}(x,y)|dxdy.$$

Obviously, definition (7) can be easily extended to measure the influence of any subset of inputs $\mathcal{S}_A = \mathbb{E}_{\mathbf{X}_A}\left(d(P_Y, P_{Y|\mathbf{X}_A})\right)$. But in this case, since there is no ANOVA-like decomposition, there is no longer the guarantee that an interaction index defined following (6):

$$\mathcal{S}_{ll'}^{TV} = \int |p_Y(y)p_{X_l}(x)p_{X_{l'}}(x') - p_{X_l,X_{l'},Y}(x,x',y)|dxdx'dy - \mathcal{S}_l^{TV} - \mathcal{S}_{l'}^{TV}$$

really measures the pure interaction between X_l and $X_{l'}$. Therefore the interpretation of higher-order moment-independent sensitivity indices is cumbersome. On the other hand, even if normalization constants have been proposed through general inequalities on f-divergences (Borgonovo (2007), Rahman (2016)), the lack of an ANOVA decomposition once again impedes the definition of a natural normalization constant equivalent to the output variance for Sobol' indices.

Recently, new moment-independent indices built upon the framework of RKHS embedding of probability distributions have also been investigated (Da Veiga, 2015, 2016). Though originally introduced as an alternative to reduce the curse of dimensionality and make the most of the vast kernel literature, we will see in what follows that they actually exhibit an ANOVA-like decomposition and can therefore be seen as a general kernelized version of Sobol' indices.

3 Kernel-based sensitivity analysis

Before introducing the kernel-based sensitivity indices, we first review some elements of the RKHS embedding of probability distributions (Smola et al., 2007), which will serve as a building block for their definition.

3.1 RKHS embedding of distributions

We first introduce a RKHS $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$ with kernel $k_{\mathcal{X}}$ and dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The *kernel mean embedding* $\mu_{\mathrm{P}} \in \mathcal{H}$ of a probability distribution P on $\mathcal{X}$ is defined as

$$\mu_{\mathrm{P}} = \mathbb{E}_{\xi \sim \mathrm{P}} k_{\mathcal{X}}(\xi, \cdot) = \int_{\mathcal{X}} k_{\mathcal{X}}(\xi, \cdot) d\mathrm{P}(\xi)$$

if $\mathbb{E}_{\xi \sim \mathrm{P}} k_{\mathcal{X}}(\xi, \xi) < \infty$, see Smola et al. (2007). The representation μ_{P} is appealing because, if the kernel $k_{\mathcal{X}}$ is characteristic, the map $\mathrm{P} \to \mu_{\mathrm{P}}$ is injective (Sriperumbudur et al., 2009, 2010). Consequently, the kernel mean embedding can be used in lieu of the probability distribution for several comparisons and manipulations of probability measures but using only inner products or distances in the RKHS. For example, a distance between two probability measures P_1 and P_2 on $\mathcal{X}$ can simply be obtained by computing the distance between their representations in $\mathcal{H}$, *i.e.*

$$\mathrm{MMD}(\mathrm{P}_1, \mathrm{P}_2) = \| \mu_{\mathrm{P}_1} - \mu_{\mathrm{P}_2} \|_{\mathcal{H}},$$

which is a distance if the kernel $k_{\mathcal{X}}$ is characteristic (Sriperumbudur et al., 2009, 2010). This distance is called the *maximum mean discrepancy* (MMD) and it has been recently used in many applications (Muandet et al., 2012; Szabó et al., 2016). Indeed, using the reproducing property of a RKHS one may show (Song, 2008) that

$$\mathrm{MMD}^2(\mathrm{P}_1, \mathrm{P}_2) = \mathbb{E}_{\xi, \xi'} k_{\mathcal{X}}(\xi, \xi') - 2\mathbb{E}_{\xi, \zeta} k_{\mathcal{X}}(\xi, \zeta) + \mathbb{E}_{\zeta, \zeta'} k_{\mathcal{X}}(\zeta, \zeta')$$

where $\xi, \xi' \sim \mathrm{P}_1$ and $\zeta, \zeta' \sim \mathrm{P}_2$ with $\xi,\ \xi',\ \zeta,\ \zeta'$ independent, this notation being used throughout the rest of the paper. This means that the MMD can be computed with expectations of kernels only, unlike other distances between probability distributions which will typically require density estimation.

Another significant application of kernel embeddings concerns the problem of measuring the dependence between random variables. Given a pair of random vectors $(\mathbf{U}, \mathbf{V}) \in \mathcal{X} \times \mathcal{Y}$ with probability distribution $\mathrm{P}_{\mathbf{UV}}$, we define the product RKHS $\mathcal{H} = \mathcal{F} \times \mathcal{G}$ with kernel $k_H((\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}')) = k_{\mathcal{X}}(\mathbf{u}, \mathbf{u}') k_{\mathcal{Y}}(\mathbf{v}, \mathbf{v}')$. A measure of the dependence between $\mathbf{U}$ and $\mathbf{V}$ can then be defined as the distance between the mean embedding of $\mathrm{P}_{\mathbf{UV}}$ and $\mathrm{P}_{\mathbf{U}} \otimes \mathrm{P}_{\mathbf{V}}$, the joint distribution with independent marginals $\mathrm{P}_{\mathbf{U}}$ and $\mathrm{P}_{\mathbf{V}}$:

$$\mathrm{MMD}^2(\mathrm{P}_{\mathbf{UV}}, \mathrm{P}_{\mathbf{U}} \otimes \mathrm{P}_{\mathbf{V}}) = \| \mu_{\mathrm{P}_{\mathbf{UV}}} - \mu_{\mathrm{P}_{\mathbf{U}}} \otimes \mu_{\mathrm{P}_{\mathbf{V}}} \|_{\mathcal{H}}.$$

This measure is the so-called *Hilbert-Schmidt independence criterion* (HSIC, see Gretton et al. (2005a,b)) and can be expanded as

$$\begin{aligned}\mathrm{HSIC}(\mathbf{U},\mathbf{V}) &= \mathrm{MMD}^2(\mathrm{P}_{\mathbf{UV}},\mathrm{P}_{\mathbf{U}}\otimes\mathrm{P}_{\mathbf{V}}) \\ &= \mathbb{E}_{\mathbf{U},\mathbf{U}',\mathbf{V},\mathbf{V}'}k_{\mathcal{X}}(\mathbf{U},\mathbf{U}')k_{\mathcal{Y}}(\mathbf{V},\mathbf{V}') \\ &+ \mathbb{E}_{\mathbf{U},\mathbf{U}'}k_{\mathcal{X}}(\mathbf{U},\mathbf{U}')\mathbb{E}_{\mathbf{V},\mathbf{V}'}k_{\mathcal{Y}}(\mathbf{V},\mathbf{V}') \\ &- 2\mathbb{E}_{\mathbf{U},\mathbf{V}}\left[\mathbb{E}_{\mathbf{U}'}k_{\mathcal{X}}(\mathbf{U},\mathbf{U}')\mathbb{E}_{\mathbf{V}'}k_{\mathcal{Y}}(\mathbf{V},\mathbf{V}')\right] \qquad (8)\end{aligned}$$

where $(\mathbf{U}',\mathbf{V}')$ is an independent copy of $(\mathbf{U},\mathbf{V})$. Once again, the reproducing property implies that HSIC can be expressed as expectations of kernels, which facilitates its estimation when compared to other dependence measures such as the mutual information.

3.2 Kernel-based ANOVA decomposition

The RKHS framework introduced above can readily be used to define kernel-based sensitivity indices. The first approach relies on the MMD, while the second one builds upon HSIC. We discuss them below and show that in particular both of them admit an ANOVA-like decomposition.

3.2.1 MMD-based sensitivity index

The first natural idea is to come back to the general formulation for moment-independent indices (7) and use the MMD as the dissimilarity measure to compare P_Y and $\mathrm{P}_{Y|X_l}$ as proposed in Da Veiga (2016):

$$\begin{aligned}\mathcal{S}_l^{\mathrm{MMD}} &= \mathbb{E}_{X_l}\mathrm{MMD}^2(\mathrm{P}_Y,\mathrm{P}_{Y|X_l}) \\ &= \mathbb{E}_{X_l}\mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y}k_{\mathcal{Y}}(\xi,\xi') - 2\mathbb{E}_{X_l}\mathbb{E}_{\xi\sim\mathrm{P}_Y,\zeta\sim\mathrm{P}_{Y|X_l}}k_{\mathcal{Y}}(\xi,\zeta) + \mathbb{E}_{X_l}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|X_l}}k_{\mathcal{Y}}(\zeta,\zeta') \\ &= \mathbb{E}_{X_l}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|X_l}}k_{\mathcal{Y}}(\zeta,\zeta') - \mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y}k_{\mathcal{Y}}(\xi,\xi')\end{aligned}$$

where we have defined a RKHS $\mathcal{G}$ of functions $\mathcal{Y}\to\mathbb{R}$ with kernel $k_{\mathcal{Y}}$. More generally, we can also consider the unnormalized MMD-based sensitivity index for a group of variables $\mathbf{X}_A$ given by $\mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y,\mathrm{P}_{Y|\mathbf{X}_A})\right) = \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}_{Y|\mathbf{X}_A}}k_{\mathcal{Y}}(\zeta,\zeta') - \mathbb{E}_{\xi,\xi'\sim\mathrm{P}_Y}k_{\mathcal{Y}}(\xi,\xi')$, provided the following assumption holds:

Assumption 1. *$\forall A\subseteq\mathcal{P}_d$ and $\forall\mathbf{x}_A\in\mathcal{X}_A$, $\mathbb{E}_{\xi\sim\mathrm{P}_{Y|\mathbf{X}_A=\mathbf{x}_A}}k_{\mathcal{Y}}(\xi,\xi)<\infty$ with the convention $\mathrm{P}_{Y|\mathbf{X}_A}=\mathrm{P}_Y$ if $A=\emptyset$.*

First note that if we focus on the scalar output case $\mathcal{Y}\subset\mathbb{R}$ with the linear kernel $k_{\mathcal{Y}}(y,y')=yy'$, we have

$$\begin{aligned}\mathcal{S}_A^{\mathrm{MMD}} &= \mathbb{E}_{\mathbf{X}_A}\left(\mathbb{E}_{\xi\sim\mathrm{P}_Y}(\xi) - \mathbb{E}_{\zeta\sim\mathrm{P}_{Y|\mathbf{X}_A}}(\zeta)\right)^2 \\ &= \mathbb{E}_{\mathbf{X}_A}\left(\mathbb{E}Y - \mathbb{E}(Y|\mathbf{X}_A)\right)^2 \\ &= \mathrm{Var}\,\mathbb{E}(Y|\mathbf{X}_A),\end{aligned}$$

that is, we recover the unnormalized Sobol' index for $\mathbf{X}_A$. $\mathcal{S}_A^{\mathrm{MMD}}$ can thus be seen as a kernelized version of Sobol' indices since the latter can be retrieved with a specific kernel. However it is obvious that since the linear kernel is not characteristic, the MMD in this case is not a distance,

which means that $\mathcal{S}_A^{\mathrm{MMD}}$ is no longer a moment-independent index.
To make another connection with Sobol' indices, we now recall Mercer's theorem, a notable representation theorem for kernels.

Theorem 2 (Mercer, see Aubin (2000))**.** *Suppose $k_{\mathcal{Y}}$ is a continuous symmetric positive definite kernel on a compact set $\mathcal{Y}$ and consider the integral operator $T_{k_{\mathcal{Y}}} : \mathbb{L}^2(\mathcal{Y}) \to \mathbb{L}^2(\mathcal{Y})$ defined by*

$$\left(T_{k_{\mathcal{Y}}} f\right)(x) = \int_{\mathcal{Y}} k_{\mathcal{Y}}(y,u) f(u) du.$$

Then there is an orthonormal basis $\{e_r\}$ of $\mathbb{L}^2(\mathcal{Y})$ consisting of eigenfunctions of $T_{k_{\mathcal{Y}}}$ such that the corresponding sequence of eigenvalues $\{\lambda_r\}$ are non-negative. The eigenfunctions corresponding to non-zero eigenvalues are continuous on $\mathcal{Y}$ and $k_{\mathcal{Y}}$ has the following representation

$$k_{\mathcal{Y}}(y,y') = \sum_{r=1}^{\infty} \lambda_r e_r(y) e_r(y')$$

where the convergence is absolute and uniform.

Assume now that the output $Y \in \mathcal{Y}$ with $\mathcal{Y}$ a compact set, meaning that Mercer's theorem holds. Then $k_{\mathcal{Y}}$ admits a representation

$$k_{\mathcal{Y}}(y,y') = \sum_{r=1}^{\infty} \phi_r(y) \phi_r(y')$$

where $\phi_r(y) = \sqrt{\lambda_r} e_r(y)$ are orthogonal functions in $\mathbb{L}^2(\mathcal{Y})$. In this setting we can write

$$\begin{aligned}
\mathcal{S}_A^{\mathrm{MMD}} = \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right) &= \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi') - \mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}} k_{\mathcal{Y}}(\zeta,\zeta') \\
&= \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} \left(\sum_{r=1}^{\infty} \phi_r(\xi)\phi_r(\xi')\right) \\
&\quad -\mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}} \left(\sum_{r=1}^{\infty} \phi_r(\zeta)\phi_r(\zeta')\right).
\end{aligned}$$

Now, since the convergence of the series is absolute, we can interchange the expectations and the summations to get

$$\begin{aligned}
\mathcal{S}_A^{\mathrm{MMD}} &= \sum_{r=1}^{\infty} \left\{ \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} \left(\phi_r(\xi)\phi_r(\xi')\right) - \mathbb{E}_{\zeta,\zeta'\sim\mathrm{P}} \left(\phi_r(\zeta)\phi_r(\zeta')\right) \right\} \\
&= \sum_{r=1}^{\infty} \left\{ \mathbb{E}_{\mathbf{X}_A}\mathbb{E}\left(\phi_r(Y)|\mathbf{X}_A\right)^2 - \mathbb{E}\left(\phi_r(Y)\right)^2 \right\} \\
&= \sum_{r=1}^{\infty} \mathrm{Var}\,\mathbb{E}\left(\phi_r(Y)|\mathbf{X}_A\right). \qquad (9)
\end{aligned}$$

In other words, the MMD-based sensitivity index $\mathcal{S}_A^{\mathrm{MMD}}$ generalizes the Sobol' one in the sense that it measures the impact of the inputs not only on the conditional expectation of the output,

but on a possibly infinite number of transformations ϕ_r of the output, given by the eigenfunctions of the kernel.

We can now state the main theorem of this section on the ANOVA-like decomposition for $\mathcal{S}_A^{\mathrm{MMD}}$. Recall that the variance decomposition (1) states that the variance of the output can be decomposed as $\operatorname{Var} Y = \sum_{A \subseteq \mathcal{P}_d} V_A$ where each term is given by

$$V_A = \sum_{B \subset A} (-1)^{|A|-|B|} \operatorname{Var} \mathbb{E}(Y|\mathbf{X}_B).$$

The MMD-based equivalent is obtained with the following theorem.

Theorem 3 (ANOVA decomposition for MMD)**.** *Under the same assumptions of Theorem 1 (in particular, the random vector* $\mathbf{X}$ *has independent components) and with Assumption 1, denote* $\mathrm{MMD}^2_{\mathrm{tot}} = \mathbb{E}k_{\mathcal{Y}}(Y,Y) - \mathbb{E}k_{\mathcal{Y}}(Y,Y')$ *where* Y' *is an independent copy of* Y*. Then the total MMD can be decomposed as*

$$\mathrm{MMD}^2_{\mathrm{tot}} = \sum_{A \subseteq \mathcal{P}_d} \mathrm{MMD}^2_A$$

where each term is given by

$$\mathrm{MMD}^2_A = \sum_{B \subset A} (-1)^{|A|-|B|} \mathbb{E}_{\mathbf{X}_B} \left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_B}) \right).$$

The proof is given in Appendix A.1. Theorem 3 is very similar to the ANOVA one given in (1): one can note that the total variance of the output is replaced by a generalized variance $\mathrm{MMD}^2_{\mathrm{tot}}$ defined by the kernel, and that each subset effect is obtained by removing lesser order ones in the MMD distance of the conditional distributions (instead of the variance of the conditional expectations in the ANOVA). The following corollary states that these two decompositions coincide when the kernel is chosen as the linear one.

Corollary 1. *When* $Y \in \mathcal{Y} \subset \mathbb{R}$ *and* $k_{\mathcal{Y}}(y,y') = yy'$ *in Theorem 3, the decomposition is identical to the decomposition (1), which means that*

$$\mathrm{MMD}^2_{\mathrm{tot}} = \operatorname{Var} Y \text{ and } \forall B \in \mathcal{P}_d,\ \mathbb{E}_{\mathbf{X}_B} \left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_B}) \right) = \operatorname{Var} \mathbb{E}(Y|\mathbf{X}_B).$$

It further implies $\forall A \subseteq \mathcal{P}_d,\ \mathrm{MMD}^2_A = V_A$.

Thanks to Theorem 3 we can now define properly normalized MMD-based indices.

Definition 2 (MMD-based sensitivity indices)**.** *In the frame of Theorem 3, let* $A \subseteq \mathcal{P}_d$*. The normalized MMD-based sensitivity index associated to a subset* A *of input variables is defined as*

$$S_A^{\mathrm{MMD}} = \frac{\mathrm{MMD}^2_A}{\mathrm{MMD}^2_{\mathrm{tot}}},$$

while the total MMD-based index associated to A *is*

$$S_A^{T,\mathrm{MMD}} = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B^{\mathrm{MMD}} = 1 - \frac{\mathbb{E}_{\mathbf{X}_{-A}} \left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_{-A}}) \right)}{\mathrm{MMD}^2_{\mathrm{tot}}}.$$

From Theorem 3, we have the fundamental identity providing the interpretation of MMD-based indices as percentage of the explained generalized variance $\mathrm{MMD}^2_{\mathrm{tot}}$*:*

$$\sum_{A \subseteq \mathcal{P}_d} S_A^{\mathrm{MMD}} = 1.$$

Finally, we exhibit a generalized law of total variance for $\mathrm{MMD}^2_{\mathrm{tot}}$ which will yield another formulation for the total MMD-based index.

Proposition 1 (Generalized law of total variance)**.** *Assuming Assumption 1 holds, we have*

$$\mathrm{MMD}^2_{\mathrm{tot}} = \mathbb{E}_{\mathbf{X}_A}\left[\mathbb{E}_{\xi \sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi' \sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi')\right] + \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right).$$

The proof is to be found in Appendix A.2. This is a generalization in the sense that the total variance is replaced by $\mathrm{MMD}^2_{\mathrm{tot}}$, the conditional variance by $\mathbb{E}_{\xi \sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi' \sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi')$ and the variance of the conditional expectation by $\mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)$. In particular, all these terms reduce to the ones in the classical law of total variance if one uses the linear kernel $k_{\mathcal{Y}}(y,y') = yy'$ in the scalar case. This gives the following corollary.

Corollary 2 (Other formulation of total MMD-based index)**.** *In the frame of Theorem 3, we have for all* $A \subseteq \mathcal{P}_d$

$$S_A^{T,\mathrm{MMD}} = \frac{\mathbb{E}_{\mathbf{X}_{-A}}\left[\mathbb{E}_{\xi \sim \mathrm{P}_{Y|\mathbf{X}_{-A}}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi' \sim \mathrm{P}_{Y|\mathbf{X}_{-A}}} k_{\mathcal{Y}}(\xi,\xi')\right]}{\mathrm{MMD}^2_{\mathrm{tot}}}.$$

3.2.2 HSIC-based sensitivity indices

Another approach for combining kernel embeddings with sensitivity analysis consists in directly using HSIC as a sensitivity index. For example Da Veiga (2015) considers the unnormalized index

$$\mathcal{S}_A^{HS} = \mathrm{HSIC}(\mathbf{X}_A, Y)$$

relying on a product RKHS $\mathcal{H}_A = \mathcal{F}_A \times \mathcal{G}$ with kernel $k_{\mathcal{H}_A}((\mathbf{x},y),(\mathbf{x}',y')) = k_{\mathcal{X}_A}(\mathbf{x}_A,\mathbf{x}'_A)k_{\mathcal{Y}}(y,y')$ and provided the following assumption holds:

Assumption 2. $\forall A \subseteq \mathcal{P}_d$, $\mathbb{E}_{\xi \sim \mathrm{P}_{X_A}} k_{\mathcal{X}_A}(\xi,\xi) < \infty$ *and* $\mathbb{E}_{\xi \sim \mathrm{P}_Y} k_{\mathcal{Y}}(\xi,\xi) < \infty$.

In Da Veiga (2015) an empirical normalization inspired by the definition of the distance correlation criterion (Székely et al., 2007) was also proposed. But similarly to the MMD decomposition above, it is actually possible to exhibit an ANOVA-like decomposition for HSIC, thus providing a natural normalization constant. The main ingredient is an assumption on the kernel $k_{\mathcal{X}}$ associated to the input variables.

Assumption 3. *The reproducing kernel* $k_{\mathcal{X}}$ *of* $\mathcal{F}$ *is of the form*

$$k_{\mathcal{X}}(\mathbf{x},\mathbf{x}') = \prod_{l=1}^{p}\left(1 + k_l(x_l,x'_l)\right) \tag{10}$$

where for each $l = 1, \dots, d$, $k_l(\cdot, \cdot)$ is the reproducing kernel of a RKHS $\mathcal{F}_l$ of real functions depending only on variable x_l and such that $1 \notin \mathcal{F}_l$.
In addition, for all $l = 1, \dots, d$ and $\forall x_l \in \mathcal{X}_l$, we have

$$\int_{\mathcal{X}_l} k_l(x_l, x_l') d\mathrm{P}_{X_l}(x_l') = 0. \tag{11}$$

The first part (10) of Assumption 3 may seem stringent, however it can be easily fulfilled by using univariate Gaussian kernels since they define a RKHS which does not contain constant functions (Steinwart et al., 2006).

On the contrary, the second assumption (11) is more subtle. It requires using kernels defining a so-called RKHS of *zero-mean functions* (Wahba et al., 1995). A prominent example of such RKHS is obtained if (a) all input variables are uniformly distributed on $[0, 1]$ and (b) the univariate kernels are chosen among the Sobolev kernels with smoothness parameter $r \geq 1$:

$$k_l(x_l, x_l') = \frac{B_{2r}(|x_l - x_l'|)}{(-1)^{r+1}(2r)!} + \sum_{j=1}^{r} \frac{B_j(x_l) B_j(x_l')}{(j!)^2} \tag{12}$$

where B_j is the Bernoulli polynomial of degree j. Even though applying a preliminary transformation on the inputs in order to get uniform variables is conceivable (with *e.g.* the probability integral transform), a more general and elegant procedure has been proposed by Durrande et al. (2012). Starting from an arbitrary univariate $k(\cdot, \cdot)$, they build a zero-mean kernel $k_0^D(\cdot, \cdot)$ given by

$$k_0^D(x, x') = k(x, x') - \frac{\int k(x, t) d\mathrm{P}(t) \int k(x', t) d\mathrm{P}(t)}{\iint k(s, t) d\mathrm{P}(s) d\mathrm{P}(t)}$$

where $k_0^D(\cdot, \cdot)$ satisfies $\forall x,\ \int k_0^D(x, t) d\mathrm{P}(t) = 0$. Interestingly, they also show that the RKHS $\mathcal{H}_0$ associated to $k_0(\cdot, \cdot)$ is orthogonal to the constant functions, thus satisfying directly the requirements for the product kernel (10).
More recently, several works made use of the Stein operator (Stein et al., 1972) to define the Stein discrepancy in a RKHS (Chwialkowski et al., 2016) which showed great potential for Monte-Carlo integration (Oates et al., 2017) or goodness-of-fit tests (Gorham and Mackey, 2015; Chwialkowski et al., 2016; Jitkrittum et al., 2017) when the target distribution is either impossible to sample or is known up to a normalization constant. More precisely, given a RKHS $\mathcal{H}$ with kernel $k(\cdot, \cdot)$ of functions in $\mathbb{R}^d$ and a (target) probability distribution with density $p(\cdot)$, they define a new RKHS $\mathcal{H}_0$ with kernel $k_0^S(\cdot, \cdot)$ which writes

$$k_0^S(\mathbf{x}, \mathbf{x}') = \nabla_{\mathbf{x}} \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + \frac{\nabla_{\mathbf{x}} p(\mathbf{x})}{p(\mathbf{x})} \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + \frac{\nabla_{\mathbf{x}'} p(\mathbf{x}')}{p(\mathbf{x}')} \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}') + \frac{\nabla_{\mathbf{x}} p(\mathbf{x})}{p(\mathbf{x})} \frac{\nabla_{\mathbf{x}'} p(\mathbf{x}')}{p(\mathbf{x}')} k(\mathbf{x}, \mathbf{x}')$$

and it can be proved that $\forall \mathbf{x} \in \mathbb{R}^d,\ \int k_0^S(\mathbf{x}, \mathbf{x}') p(\mathbf{x}') d\mathbf{x}' = 0$. Unlike $k_0^D(\cdot, \cdot)$, kernel $k_0^S(\cdot, \cdot)$ can still be defined when $p(\cdot)$ is known up to a constant: this property may find interesting applications in GSA when the input distributions are obtained via a preliminary Bayesian data calibration, since it would no longer be required to perform a costly sampling step of their posterior distribution and one could easily use the unnormalized posterior distribution instead.

With Assumption 3, we can now state a decomposition for HSIC-based sensitivity indices.

Theorem 4 (ANOVA decomposition for HSIC)**.** *Under the same assumptions of Theorem 1 (in particular, the random vector* $\mathbf{X}$ *has independent components) and with Assumptions 2 and 3, the HSIC dependence measure between* $\mathbf{X} = (X_1, \ldots, X_d)$ *and* Y *can be decomposed as*

$$\mathrm{HSIC}\left(\mathbf{X}, Y\right) = \sum_{A \subseteq \mathcal{P}_d} \mathrm{HSIC}_A$$

where each term is given by

$$\mathrm{HSIC}_A = \sum_{B \subset A} (-1)^{|A|-|B|} \mathrm{HSIC}\left(\mathbf{X}_B, Y\right)$$

and $\mathrm{HSIC}\left(\mathbf{X}_B, Y\right)$ *is defined with a product RKHS* $\mathcal{H}_B = \mathcal{F}_B \times \mathcal{G}$ *with kernel* $k_B(\mathbf{x}_B, \mathbf{x}'_B) k_{\mathcal{Y}}(y, y') = \prod_{l \in B} \left(1 + k_l(x_l, x'_l)\right) k_{\mathcal{Y}}(y, y')$ *as in (10).*

The proof, which mainly relies on Mercer's theorem and on Theorem 4.1 from Kuo et al. (2010), is given in Appendix A.3. Once again, this decomposition resembles the ANOVA decomposition (1), where the conditional variances are replaced with HSIC dependence measures between subsets of inputs and the output.
Properly normalized HSIC-based indices can then be defined:

Definition 3 (HSIC-based sensitivity indices)**.** *In the frame of Theorem 4, let* $A \subseteq \mathcal{P}_d$*. The normalized HSIC-based sensitivity index associated to a subset* A *of input variables is defined as*

$$S_A^{\mathrm{HSIC}} = \frac{\mathrm{HSIC}_A}{\mathrm{HSIC}\left(\mathbf{X}, Y\right)},$$

while the total HSIC-based index associated to A *is*

$$S_A^{T,\mathrm{HSIC}} = \sum_{B \subseteq \mathcal{P}_d,\, B \cap A \neq \emptyset} S_B^{\mathrm{HSIC}} = 1 - \frac{\mathrm{HSIC}(\mathbf{X}_{-A}, Y)}{\mathrm{HSIC}\left(\mathbf{X}, Y\right)}.$$

From Theorem 4, we have the fundamental identity providing the interpretation of HSIC-based indices as percentage of the explained HSIC dependence measure between $\mathbf{X} = (X_1, \ldots, X_d)$ *and* Y*:*

$$\sum_{A \subseteq \mathcal{P}_d} S_A^{\mathrm{HSIC}} = 1.$$

Finally, a noteworthy asymptotic result yields a link between HSIC-based indices and MMD-based ones when the input kernel $k_{\mathcal{X}}$ degenerates to a dirac kernel, as elaborated in the following proposition.

Proposition 2. *For all subset* $A \subseteq \mathcal{P}_d$*, let us define a product RKHS* $\mathcal{H}_A = \mathcal{F}_A \times \mathcal{G}$ *with kernel* $k_A(\mathbf{x}_A, \mathbf{x}'_A) k_{\mathcal{Y}}(y, y')$*. We further assume that* $\forall \mathbf{x}_A \in \mathcal{X}_A$*,* $p_{\mathbf{X}_A}(\mathbf{x}_A) > 0$ *and that*

$$k_A(\mathbf{x}_A, \mathbf{x}'_A) = \frac{1}{\sqrt{p_{\mathbf{X}_A}(\mathbf{x}_A)} \sqrt{p_{\mathbf{X}_A}(\mathbf{x}'_A)}} \prod_{l \in A} \frac{1}{h} K\left(\frac{x_l - x'_l}{h}\right) \tag{13}$$

where $K : \mathbb{R} \to \mathbb{R}$ *is a symmetric kernel function satisfying* $\int_u K(u) du = 1$*, and* $h > 0$*. Then we have* $\forall A \subseteq \mathcal{P}_d$

$$\lim_{h \to 0} \mathrm{HSIC}(\mathbf{X}_A, Y) = \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)$$

where $\mathrm{HSIC}(\mathbf{X}_A, Y)$ *is defined with the product RKHS* $\mathcal{H}_A = \mathcal{F}_A \times \mathcal{G}$ *and* $\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})$ *with the RKHS* $\mathcal{G}$*.*

The proof is given in Appendix A.4. As a particular case of Proposition 2, one can for example choose a (normalized) Gaussian kernel for $k_{\mathcal{X}}$ with a standard deviation tending to 0, or the sinc kernel associated to the RKHS of band-limited continuous functions with a cutoff frequency tending to infinity. Obviously the result also holds if one uses different kernels K for each input $X_l \in \mathbf{X}_A$ in Eq. (13).

Although they may seem trivial, Proposition 2 and Corollary 1 actually justify our claim that both the MMD- and the HSIC-based sensitivity indices are natural generalizations of Sobol' indices, in the sense that a degenerate HSIC-based index with a dirac kernel for the input variables gives the MMD-based index which, in turn, is equal to the Sobol' index when using the dot product kernel for the output.

3.3 Kernel-embedding Shapley effects

In this section, we now discuss how the previously indices can still be valuable in the case where the input variables are no longer independent. In this setting, the Shapley effects introduced in the context of GSA by Owen (Owen, 2014) and based on Shapley values (Shapley, 1953) from game theory have appealing properties, since they provide a proper allocation of the output variance to each input variable, without requiring they are independent. We recall their definition below.

Definition 4 (Shapley effects (Shapley, 1953))**.** *For any $l = 1 \dots, d$, the Shapley effect of input X_l is given by*

$$Sh_l = \frac{1}{\operatorname{Var} Y} \frac{1}{p} \sum_{A \subseteq \mathcal{P}_d,\, A \not\ni l} \binom{p-1}{|A|}^{-1} \left\{ \operatorname{Var} \mathbb{E}\left(Y | \mathbf{X}_{A \cup \{l\}}\right) - \operatorname{Var} \mathbb{E}\left(Y | \mathbf{X}_A\right) \right\}. \tag{14}$$

This definition corresponds to the Shapley value (Shapley, 1953)

$$\phi_l = \frac{1}{p} \sum_{A \subseteq \mathcal{P}_d,\, A \not\ni l} \binom{p-1}{|A|}^{-1} \left\{ \operatorname{val}\left(A \cup \{l\}\right) - \operatorname{val}\left(A\right) \right\}$$

with value function $\operatorname{val} : \mathcal{P}_d \to \mathbb{R}_+$ *equal to* $\operatorname{val}(A) = \operatorname{Var} \mathbb{E}\left(Y | \mathbf{X}_A\right) / \operatorname{Var} Y$*. Moreover, we have the following decomposition*

$$\sum_{l=1}^{p} Sh_l = 1.$$

The only requirement is that the value function satisfies $\operatorname{val} : \mathcal{P}_d \to \mathbb{R}_+$ such that $\operatorname{val}(\emptyset) = 0$. Combining this result with the kernel-based sensitivity indices is consequently straightforward, which leads to the definition of *kernel-embedding Shapley effects*:

Definition 5 (Kernel-embedding Shapley effects)**.** *For any $l = 1 \dots, d$, we define*

(a) The MMD-Shapley effect

$$\begin{aligned} Sh_l^{\mathrm{MMD}} \;=\; & \frac{1}{\mathrm{MMD}^2_{\mathrm{tot}}} \frac{1}{p} \sum_{A \subseteq \mathcal{P}_d,\, A \not\ni l} \binom{p-1}{|A|}^{-1} \left\{ \mathbb{E}_{\mathbf{X}_{A \cup \{l\}}} \left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_{A \cup \{l\}}}) \right) \right. \\ & \left. - \mathbb{E}_{\mathbf{X}_A} \left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A}) \right) \right\} \end{aligned} \tag{15}$$

provided Assumption 1 holds.

(b) The HSIC-Shapley effect

$$Sh_l^{\mathrm{HSIC}} = \frac{1}{\mathrm{HSIC}\left(\mathbf{X}, Y\right)} \frac{1}{p} \sum_{A \subseteq \mathcal{P}_d,\, A \not\ni l} \binom{p-1}{|A|}^{-1} \left\{ \mathrm{HSIC}\left(\mathbf{X}_{A\cup\{l\}}, Y\right) - \mathrm{HSIC}\left(\mathbf{X}_A, Y\right) \right\} \quad (16)$$

provided Assumptions 2 and 3 hold.

We further have the decompositions

$$\sum_{l=1}^{p} Sh_l^{\mathrm{MMD}} = \sum_{l=1}^{p} Sh_l^{\mathrm{HSIC}} = 1.$$

Just like in the independent setting, kernel-embedding Shapley effects (15) and (16) can be seen as general kernelized versions of Shapley effects, since Proposition 2 and Corollary 1 are still valid when the inputs are dependent.

Remark 1. *In the machine learning community dedicated to the interpretability of black-box models, an importance measure called Kernel-Shap has been recently introduced (Lundberg and Lee, 2017). Although its naming resembles ours, they designate clearly separated approaches, since the Kernel-Shap measure is a local Shapley effect, and the "Kernel" denomination only refers to an estimation procedure without any links to RKHS.*

3.4 Enhancing traditional GSA with kernels

Beyond their theoretical interest in themselves, the kernel-ANOVA decompositions and the associated sensitivity indices also appear powerful from a practical point of view when one carefully examines the potential of using kernels. We give below some insights on how they could enhance traditional GSA studies in several settings.

Categorical model outputs and target sensitivity analysis. In some applications, the model output Y is categorical, meaning that $\mathcal{Y} = \{1, \ldots, K\}$ when the output can take K levels. A simple common instance involves two levels, corresponding to a failure/success situation. Similarly even if Y is not categorical, the objective may be to measure the impact of each input on the fact that the output reaches disjoint regions of interest $\mathcal{R}_1, \ldots, \mathcal{R}_K \subset \mathcal{Y}$, as for example in the case where one focuses on events $\{t_{i+1} > Y > t_i\}$ for thresholds t_i, $i = 1, \ldots, K$. Such an objective is called *target sensitivity analysis* (TSA, see Marrel and Chabridon (2020)) and can be reformulated in a categorical framework by the change of variable $Z = i$ if $Y \in \mathcal{R}_i$.

The case where $\mathcal{Y} = \{0, 1\}$ (or equivalently $Z = \mathbb{1}_{\{Y \in \mathcal{R}\}}$) is frequent in TSA. A straightforward approach is to use Sobol' indices with a 0/1 output, yielding a first-order Sobol index equal to

$$S_l^{\mathrm{TSA}} = \frac{\mathbb{E}_{X_l}\left(\mathbb{P}(Y=1|X_l) - \mathbb{P}(Y=1)\right)^2}{\mathbb{P}(Y=1)(1-\mathbb{P}(Y=1))} \quad (17)$$

see Li et al. (2012). But to the best of our knowledge, no systematic procedure is available when the number of levels is greater than two. Without resorting yet to our kernel-based indices, there are at least two roads, which actually lead to the same indices:

(a) The *one-versus-all* approach, where we compute several Sobol' indices by repeatedly considering $Z = \mathbb{1}_{\{Y=i\}}$ for all $i = 1 \dots, K$. We thus have a collection of indices

$$S_l^{\mathrm{TSA},[i]} = \frac{\mathbb{E}_{X_l}\left(\mathbb{P}(Y=i|X_l) - \mathbb{P}(Y=i)\right)^2}{\mathbb{P}(Y=i)(1-\mathbb{P}(Y=i))},$$

and we can aggregate them by normalizing each of them by its own variance, yielding

$$S_l^{\mathrm{TSA}} = \frac{\sum_{i=1}^K \mathbb{P}(Y=i)(1-\mathbb{P}(Y=i))S_l^{\mathrm{TSA},[i]}}{\sum_{i=1}^K \mathbb{P}(Y=i)(1-\mathbb{P}(Y=i))} = \frac{\sum_{i=1}^K \mathbb{E}_{X_l}\left(\mathbb{P}(Y=i|X_l) - \mathbb{P}(Y=i)\right)^2}{\sum_{i=1}^K \mathbb{P}(Y=i)(1-\mathbb{P}(Y=i))}.$$

(b) The *one-hot encoding* approach, which consists in encoding the categorical output into a multivariate vector of 0/1 variables and use the aggregated Sobol' indices defined in Gamboa et al. (2013) on these transformed variables. More precisely if $\mathcal{Y} = \{1, \dots, K\}$, Y is encoded as a K-dimensional vector $(Z_1 = \mathbb{1}_{Y=1}, \dots, Z_K = \mathbb{1}_{Y=K})$. The aggregated Sobol' indices are then

$$S_l^{\mathrm{TSA}} = \frac{\sum_{i=1}^K \mathrm{Var}\,\mathbb{E}(Z_i|X_l)}{\sum_{i=1}^K \mathrm{Var}\,Z_i} = \frac{\sum_{i=1}^K \mathbb{E}_{X_l}\left(\mathbb{P}(Y=i|X_l) - \mathbb{P}(Y=i)\right)^2}{\sum_{i=1}^K \mathbb{P}(Y=i)(1-\mathbb{P}(Y=i))},$$

which is exactly the index obtained with the one-versus-all approach.

As for the kernel-based indices, the process is less cumbersome, since the only ingredient that requires attention is the choice of a kernel $k_{\mathcal{Y}}(\cdot,\cdot)$ adapted to categorical outputs, which has already been investigated in the kernel literature (Song et al., 2007, 2012). We focus here on the simple *dirac kernel* defined as $k_{\mathcal{Y}}(y, y') = \delta(y, y')$ for categorical values $y, y' \in \{1, \dots, K\}$, and the corresponding kernel-based indices are then

- The first-order MMD-based index:

$$\begin{aligned} S_l^{\mathrm{MMD}} &= \frac{\mathbb{E}_{X_l}\sum_{i=1}^K\sum_{j=1}^K \delta(i,j)\mathbb{P}(Y=i|X_l)\mathbb{P}(Y=j|X_l) - \sum_{i=1}^K\sum_{j=1}^K \delta(i,j)\mathbb{P}(Y=i)\mathbb{P}(Y=j)}{\sum_{i=1}^K \mathbb{P}(Y=i) - \sum_{i=1}^K\sum_{j=1}^K \delta(i,j)\mathbb{P}(Y=i)\mathbb{P}(Y=j)} \\ &= \frac{\mathbb{E}_{X_l}\sum_{i=1}^K \mathbb{P}(Y=i|X_l)^2 - \sum_{i=1}^K \mathbb{P}(Y=i)^2}{\sum_{i=1}^K \mathbb{P}(Y=i) - \sum_{i=1}^K \mathbb{P}(Y=i)^2} \\ &= \frac{\sum_{i=1}^K \mathbb{E}_{X_l}\left(\mathbb{P}(Y=i|X_l) - \mathbb{P}(Y=i)\right)^2}{\sum_{i=1}^K \mathbb{P}(Y=i)(1-\mathbb{P}(Y=i))}, \end{aligned}$$

  where we retrieve again the one-versus-all Sobol' index.

- The first-order HSIC-based index:

$$\begin{aligned} S_l^{\mathrm{HSIC}} &= \int_{\mathcal{X}_l\times\mathcal{X}_l} \sum_{i=1}^K\sum_{j=1}^K k_{\{l\}}(x,x')\delta(i,j)\left[p_{X_l|Y=i}(x) - p_{X_l}(x)\right] \\ &\quad \left[p_{X_l|Y=j}(x') - p_{X_l}(x')\right]\mathbb{P}(Y=i)\mathbb{P}(Y=j)dxdx' \\ &= \sum_{i=1}^K \mathbb{P}(Y=i)^2 \int_{\mathcal{X}_l\times\mathcal{X}_l} k_{\{l\}}(x,x')\left[p_{X_l|Y=i}(x) - p_{X_l}(x)\right]\left[p_{X_l|Y=i}(x') - p_{X_l}(x')\right]dxdx' \\ &= \sum_{i=1}^K \mathbb{P}(Y=i)^2 \mathrm{MMD}^2\left(\mathrm{P}_{X_l|Y=i}, \mathrm{P}_{X_l}\right), \end{aligned}$$

thus extending the result of Spagnol et al. (2019) to any number of levels.

- The MMD- and HSIC- Shapley effects using one of the above indices as building block.

Interestingly, it has been shown that Eq. (17) can also been written, up to a constant, as the Pearson χ^2 divergence between $\mathrm{P}_{X_l|Y=1}$ and P_{X_l} (Perrin and Defaux, 2019; Spagnol, 2020). This means that $S_l^{\mathrm{TSA}} = S_l^{\mathrm{MMD}}$ and S_l^{HSIC} essentially have the same interpretation as weighted sums of distances between the initial input distributions and the conditional input distributions (when restricted to an output level), with the Pearson χ^2 divergence and the MMD distance, respectively. But we will see in Section 4 that the estimation of HSIC-based sensitivity indices is much less prone to the curse of dimensionality and does not require density estimation, as opposed to S_l^{TSA} (Perrin and Defaux, 2019). Finally, note that another kernel for categorical variables has also been proposed (Song et al., 2007, 2012), but this is actually a normalized dirac kernel which would only modify the weights in the indices above.

Beyond scalar model outputs. In many numerical simulation models, some of the outputs are curves representing the temporal evolution of physical quantities of the system such as temperatures, pressures, etc. One can also encounter industrial applications which involve spatial outputs (Marrel et al., 2008),. In such cases, the two main approaches in GSA are (a) the ubiquitous point of view, where one sensitivity index is computed for each time step or each spatial location (Terraz et al., 2017) and (b) the dimension reduction angle, in which one preliminary projects the output into a low-dimensional vector space and then calculates aggregated sensitivity indices for this new multivariate vector (Lamboni et al., 2011; Gamboa et al., 2013).

However, the kernel perspective for such structured outputs can bring new insights for GSA. Indeed the kernel literature has already proposed several ways to handle curves or images in regression or classification tasks. For instance the PCA-kernel (Ferraty and Vieu, 2006) can be used as an equivalent of (b), such as illustrated in Da Veiga (2015). But more interestingly, kernels dedicated to times series were designed, such as the global alignment kernel (Cuturi, 2011) inspired by the dynamic time-warping kernel (Sakoe and Chiba, 1978). Such kernel could be employed in industrial applications where one is interested by the impact of an input variable on the shape of the output curve. On the other hand, for dealing with spatial outputs similar to images such as in Marrel et al. (2008), one may consider a kernel based on image classification (Harchaoui and Bach, 2007) which would be better suited to analyze the impact of inputs on the change of the shapes appearing inside the image output.

Finally, numerical models involving graphs as inputs or outputs (*e.g.* electricity networks or molecules) may now be tractable with GSA by employing kernels specifically tailored for graphs (Gärtner et al., 2003; Ramon and Gärtner, 2003).

Stochastic numerical models. On occasions one has to deal with *stochastic simulators*, where internally the numerical model relies on random draws to compute the output. Typical industrial applications include models dedicated to the optimization of maintenance costs, where random failures are simulated during the system life cycle, or molecular modeling to predict macroscopic properties based on statistical mechanics, where several microstates of the system are generated at random (Moutoussamy et al., 2015). For fixed values of the input variables, the output is therefore a probability distribution, meaning that $\mathcal{Y} \subset \mathcal{M}_1^+$ the set of probability measures. In this setting

GSA aims at measuring how changes in the inputs modify the output probability distribution, which is clearly out of the traditional scope of GSA.

Once again the kernel point of view makes it possible to easily recycle the MMD- and the HSIC-based sensitivity indices in this context since they only require the definition of a kernel $k_{\mathcal{Y}}(\cdot,\cdot)$ on probability distributions. This can be achieved through one of the two following kernels:

$$k_{\mathcal{Y}}(\mathrm{P},\mathrm{Q}) = \sigma^2 e^{-\lambda \mathrm{MMD}^2(\mathrm{P},\mathrm{Q})} \tag{18}$$

introduced in Song (2008) or

$$k_{\mathcal{Y}}(\mathrm{P},\mathrm{Q}) = \sigma^2 e^{-\lambda W_2^2(\mathrm{P},\mathrm{Q})}$$

discussed in Bachoc et al. (2017) where $\mathrm{P},\mathrm{Q} \in \mathcal{M}_1^+$, W_2 is the Wasserstein distance and $\sigma^2, \lambda > 0$ are parameters.

4 Estimation

The properly normalized kernel-based sensitivity indices being defined above, we now discuss their estimation. The HSIC-based index is first examined as we only consider already proposed estimators. On the other hand, the MMD-based index is analyzed more thoroughly since several estimators can be envisioned given its close links with Sobol' indices. Finally we investigate the estimation of kernel-embedding Shapley effects.

4.1 HSIC-based index estimation

We start by observing that if Assumption 3 holds, for any subset $A \subseteq \mathcal{P}_d$ we have $\mathbb{E}_{\mathbf{X}_A} k_A(\mathbf{X}_A, \mathbf{x}'_A) = 1$ for all $\mathbf{x}'_A \in \mathcal{X}_A$, which means that HSIC in Eq. (8) simplifies into:

$$\mathrm{HSIC}(\mathbf{X}_A, Y) = \mathbb{E}_{\mathbf{X}_A,\mathbf{X}'_A,Y,Y'} k_A(\mathbf{X}_A,\mathbf{X}'_A) k_{\mathcal{Y}}(Y,Y') - \mathbb{E}_{Y,Y'} k_{\mathcal{Y}}(Y,Y').$$

Given a sample $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$, $i = 1,\dots,n$ and following Song et al. (2007); Gretton et al. (2008) two estimators $\mathrm{HSIC}_u(\mathbf{X}_A, Y)$ and $\mathrm{HSIC}_b(\mathbf{X}_A, Y)$ based on U- and V-statistics, respectively, can be introduced:

$$\begin{aligned}
\mathrm{HSIC}_u(\mathbf{X}_A, Y) &= \frac{1}{n(n-1)} \sum_{i,j=1,\, i\neq j}^{n} \left(k_A(\mathbf{x}_A^{(i)}, \mathbf{x}_A^{(j)}) - 1\right) k_{\mathcal{Y}}(y^{(i)}, y^{(j)}) \\
\mathrm{HSIC}_b(\mathbf{X}_A, Y) &= \frac{1}{n^2} \sum_{i,j=1}^{n} \left(k_A(\mathbf{x}_A^{(i)}, \mathbf{x}_A^{(j)}) - 1\right) k_{\mathcal{Y}}(y^{(i)}, y^{(j)})
\end{aligned}$$

where we assume that $\mathrm{P}_{\mathbf{X}_A}$ is known and is used to compute analytically the zero-mean kernels in Eq. (3.2.2). The study of the version of the above estimators when the sample $\left(\mathbf{x}^{(i)}\right)_{i=1,\dots,n}$ also serves to estimate k_A is left as future work. Both $\mathrm{HSIC}_u(\mathbf{X}_A, Y)$ and $\mathrm{HSIC}_b(\mathbf{X}_A, Y)$ converge in probability to $\mathrm{HSIC}(\mathbf{X}_A, Y)$ with rate $1/\sqrt{n}$, and one can show (Song et al., 2007) that if we assume that k_A and $k_{\mathcal{Y}}$ are bounded almost everywhere by 1 and are nonnegative, with probability at least $1-\delta$ we have

$$|\mathrm{HSIC}_u(\mathbf{X}_A, Y) - \mathrm{HSIC}(\mathbf{X}_A, Y)| \leq 8\sqrt{\log(2/\delta)/n}.$$

The asymptotic distributions of $\mathrm{HSIC}_u(\mathbf{X}_A, Y)$ and $\mathrm{HSIC}_b(\mathbf{X}_A, Y)$ have also been studied in the case where $\mathbf{X}_A$ and Y are dependent, see Song et al. (2007) and Gretton et al. (2008).

It is worth mentioning that here the number of model evaluations is n, which is independent from the input dimension, meaning that all HSIC-based sensitivity indices can be computed with only a given sample $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$, $i = 1, \ldots, n$.

4.2 MMD-based index estimation

MMD-based indices are close generalizations of Sobol' indices, since they involve computing the expectation of a conditional quantity (a MMD distance with a conditional probability for the former and a conditional variance for the latter). This is the reason why estimation procedures developed for Sobol' indices can be adapted to the MMD ones. The first two estimators discussed below assume that one can easily sample the computer model for any input values (to be determined by the estimation procedure), as opposed to the next two ones which can be defined with any given sample $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$, $i = 1, \ldots, n$.

4.2.1 Double-loop Monte-Carlo

The first naive estimator consists in systematically resampling the conditional distribution $\mathrm{P}_{Y|\mathbf{X}_A=\mathbf{x}_A}$ for many values of $\mathbf{x}_A$, as detailed in Algorithm 1 below.

Algorithm 1 Double-loop estimator of $\mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)$

Sample $\mathbf{x}^{(j)}$ from $\mathrm{P}_{\mathbf{X}}$ and compute $y^{(j)} = \eta(\mathbf{x}^{(j)})$ for $j = 1, \ldots, m$.
for $i = 1 \ldots, n$ **do**
  Outer-loop
  Sample $\mathbf{x}_A^{(i)}$ from $\mathrm{P}_{\mathbf{X}_A}$;
  for $j = 1 \ldots, m$ **do**
    Inner-loop
    Sample $\mathbf{x}_{-A}^{(j)}$ from $\mathrm{P}_{\mathbf{X}_{-A}}$ (if inputs are independent) or from $\mathrm{P}_{\mathbf{X}_{-A}|\mathbf{X}_A=\mathbf{x}_A^{(i)}}$ (otherwise);
    Compute $\tilde{y}^{(j)} = \eta(\mathbf{x}')$ where $\mathbf{x}'_A = \mathbf{x}_A^{(i)}$ and $\mathbf{x}'_{-A} = \mathbf{x}_{-A}^{(j)}$;
  end for
  Compute

$$M^{(i)} = \frac{1}{n^2} \sum_{j,j'=1}^{m} k_{\mathcal{Y}}\left(y^{(j)}, y^{(j')}\right) + \frac{1}{n^2} \sum_{j,j'=1}^{m} k_{\mathcal{Y}}\left(\tilde{y}^{(j)}, \tilde{y}^{(j')}\right) - \frac{2}{n^2} \sum_{j,j'=1}^{m} k_{\mathcal{Y}}\left(y^{(j)}, \tilde{y}^{(j')}\right)$$

  the estimator of $\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A=\mathbf{x}_A^{(i)}})$;
end for
$\mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)$ is finally estimated by $\frac{1}{n}\sum_{i=1}^{n} M^{(i)}$.

For each MMD-based index of a subset of variables $\mathbf{X}_A$ the total number of model evaluations is $(n+1)m$, which means for example that all first-order MMD-based sensitivity indices are computed at a cost of $p(n+1)m$ model evaluations. It is however possible to design better sampling strategies to compute first-order and total indices if the inputs are independent, as explained in the next section.

4.2.2 Pick-freeze estimators

We begin by recalling the definition of the pick-freeze estimators for Sobol' indices.

Lemma 1 (Pick-freeze formulation of Sobol indices (Janon et al., 2014)). *Assume* $\mathbf{X}$ *and* $\mathbf{X}'$ *are two independent copies of the input vector, the inputs being independent. For any subset* $A \subseteq \mathcal{P}_d$ *define* $\mathbf{X}^{\sim A}$ *the vector assembled from* $\mathbf{X}$ *and* $\mathbf{X}'$ *such that* $\mathbf{X}_A^{\sim A} = \mathbf{X}_A$ *and* $\mathbf{X}_{-A}^{\sim A} = \mathbf{X}'_A$*. Now if we denote* $Y = \eta(\mathbf{X})$ *and* $Y^{\sim A} = \eta(\mathbf{X}^{\sim A})$*, we have*

$$\begin{aligned} \operatorname{Var}\mathbb{E}\left(Y|\mathbf{X}_A\right) &= \operatorname{Cov}\left(Y, Y^{\sim A}\right), \\ S_A^T &= 1 - \frac{\operatorname{Cov}\left(Y, Y^{\sim -A}\right)}{\operatorname{Var} Y}. \end{aligned}$$

In the particular case of $A = \{l\}$, the first-order and total indices S_l and S_l^T can be estimated by collecting estimators $\hat{V}_l$, $\hat{V}_{-l}$ and $\hat{V}$ of $\operatorname{Cov}\left(Y, Y^{\sim l}\right)$, $\operatorname{Cov}\left(Y, Y^{\sim -l}\right)$ and $\operatorname{Var} Y$, respectively. Such estimators have been first studied in Homma and Saltelli (1996), but we focus on the ones introduced by Saltelli et al. (2010) which write

$$\begin{aligned} \hat{V}_l &= \frac{1}{n}\sum_{i=1}^n \eta(\mathbf{x}^{(i)}) \left\{\eta(\mathbf{x}^{\sim l,(i)}) - \eta(\mathbf{x}'^{(i)})\right\}, \\ \hat{V}_{-l} &= \frac{1}{n}\sum_{i=1}^n \eta(\mathbf{x}'^{(i)}) \left\{\eta(\mathbf{x}^{\sim l,(i)}) - \eta(\mathbf{x}^{(i)})\right\}, \\ \hat{V} &= \frac{1}{n}\sum_{i=1}^n \eta(\mathbf{x}^{(i)})^2 - \left(\frac{1}{n}\sum_{i=1}^n \eta(\mathbf{x}^{(i)})\right)^2 \end{aligned}$$

where $\mathbf{x}^{(i)}$ and $\mathbf{x}'^{(i)}$ denote independent samples of $\mathbf{X}$ and $\mathbf{x}^{\sim l,(i)}$ is a vector such that $\mathbf{x}_l^{\sim l,(i)} = \mathbf{x}_l^{(i)}$ and $\mathbf{x}_{-l}^{\sim l,(i)} = \mathbf{x}_{-l}^{\prime(i)}$. The total number of model evaluations to estimate both S_l and S_l^T is thus $(p+2)n$, which is much less than the amount required by the previously introduced double-loop estimator.

We now build upon these estimators to design equivalent ones for the first-order and total MMD-based sensitivity indices. The main ingredient is to state an equivalent of Lemma 1 for the MMD.

Lemma 2 (Pick-freeze formulation of MMD-based indices). *With the same notations and assumptions as in Lemma 1, we have*

$$\mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right) = \mathbb{E}k_{\mathcal{Y}}\left(Y, Y^{\sim A}\right) - \mathbb{E}k_{\mathcal{Y}}\left(Y, Y'\right).$$

Proof. Since Y and $Y^{\sim A}$ are conditionally independent on $\mathbf{X}_A$ with the same distribution, we can write $\mathbb{E}k_{\mathcal{Y}}\left(Y, Y^{\sim A}\right) = \mathbb{E}_{\mathbf{X}_A}\mathbb{E}\left[k_{\mathcal{Y}}\left(Y, Y^{\sim A}\right)|\mathbf{X}_A\right] = \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\zeta,\zeta'\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\zeta, \zeta')$. □

Estimators $\widehat{\mathrm{MMD}}_l^2$ for $\mathbb{E}_{X_l}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$ and $\widehat{\mathrm{MMD}}_{-l}^2$ for $\mathbb{E}_{X_{-l}}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|X_{-l}})\right)$ are therefore given by

$$\begin{aligned} \widehat{\mathrm{MMD}}_l^2 &= \frac{1}{n}\sum_{i=1}^n \left\{k_{\mathcal{Y}}\left(\eta(\mathbf{x}^{(i)}), \eta(\mathbf{x}^{\sim l,(i)})\right) - k_{\mathcal{Y}}\left(\eta(\mathbf{x}^{(i)}), \eta(\mathbf{x}'^{(i)})\right)\right\} \\ \widehat{\mathrm{MMD}}_{-l}^2 &= \frac{1}{n}\sum_{i=1}^n \left\{k_{\mathcal{Y}}\left(\eta(\mathbf{x}'^{(i)}), \eta(\mathbf{x}^{\sim l,(i)})\right) - k_{\mathcal{Y}}\left(\eta(\mathbf{x}^{(i)}), \eta(\mathbf{x}'^{(i)})\right)\right\} \end{aligned}$$

Similarly the normalization constant $\mathrm{MMD}^2_{\mathrm{tot}} = \mathbb{E}k_{\mathcal{Y}}(Y,Y) - \mathbb{E}k_{\mathcal{Y}}(Y,Y')$ is estimated by

$$\widehat{\mathrm{MMD}}^2_{\mathrm{tot}} = \frac{1}{n}\sum_{i=1}^{n} k_{\mathcal{Y}}\left(\eta(\mathbf{x}^{(i)}),\eta(\mathbf{x}^{(i)})\right) - \frac{1}{n^2}\sum_{i,j=1}^{n} k_{\mathcal{Y}}\left(\eta(\mathbf{x}^{(i)}),\eta(\mathbf{x}^{(j)})\right).$$

All these estimators can actually be recovered by using Mercer's theorem $k_{\mathcal{Y}}(y,y') = \sum_{r=1}^{\infty}\phi_r(y)\phi_r(y')$ and plugging the Sobol' estimators of $\mathrm{Cov}\left(\phi_r(Y),\phi_r(Y^{\sim l})\right)$, $\mathrm{Cov}\left(\phi_r(Y),\phi_r(Y^{\sim -l})\right)$ and $\mathrm{Var}\,\phi_r(Y)$ for all $r>1$. Once again all first-order and total MMD-based sensitivity indices can be estimated with a total cost of $(p+2)n$ model evaluations, and by the strong law of large numbers it is straightforward to show that both $\widehat{\mathrm{MMD}}^2_l/\widehat{\mathrm{MMD}}^2_{\mathrm{tot}}$ and $\widehat{\mathrm{MMD}}^2_{-l}/\widehat{\mathrm{MMD}}^2_{\mathrm{tot}}$ are consistent.

4.2.3 First-order index estimation with ranks

The two previous estimators, although simple, necessitate specific sampling schemes (double-loop Monte-Carlo or pick-freeze) which may not be amenable in practice. In addition first-order MMD indices estimation call for a number of model evaluations which increases with the number of input variables d. Recently, Gamboa et al. (2020) introduced new estimators of first-order Sobol' indices based on ranking and inspired by the work of Chatterjee (2020). In particular, for any pair of random variables (V,Y) and measurable bounded functions f and g, they propose a universal estimation procedure for expectations of the form

$$\mathbb{E}\left(\mathbb{E}[f(Y)|V]\mathbb{E}[g(Y)|V]\right)$$

using only a given sample $(v^{(i)},y^{(i)})_{i=1,\dots,n}$ and an estimator given by

$$\frac{1}{n}\sum_{i=1}^{n} f(y^{(i)})g(y^{(\sigma_n(i))})$$

where σ_n is a random permutation with no fixed point and measurable with respect to the σ-algebra generated by $(v^{(1)},\dots,v^{(n)})$. First-order Sobol' indices are then estimated using $f(x)=g(x)=x$ and the permutation $\sigma_n = N$ defined as in Chatterjee (2020):

$$N(i) = \begin{cases} \pi^{-1}(\pi(i)+1) & \text{if } \pi(i)+1 \leq n \\ \pi^{-1}(1) & \text{otherwise} \end{cases} \tag{19}$$

where $\pi(i)$ is the rank of $V^{(i)}$ in the sample $(V^{(1)},\dots,V^{(n)})$. All first-order indices are finally obtained with a given sample by considering one after the other the pairs (X_l,Y) with their own permutation based on the sample ranks of X_l.

Interestingly, it is possible to generalize this result to the first-order MMD indices with the following proposition.

Proposition 3 (Generalization of Proposition 3.2 from Gamboa et al. (2020))**.** *Let $k(\cdot,\cdot)$ be a measurable bounded kernel and $(v^{(i)},y^{(i)})_{i=1,\dots,n}$ an iid sample from a pair of random variables (V,Y). Consider a random permutation with no fixed point and measurable with respect to the σ-algebra generated by $(v^{(1)},\dots,v^{(n)})$ such that for any $i=1,\dots,n$, $v^{(\sigma_n(i))}\to v^{(i)}$ as $n\to\infty$ with probability one. Then the estimator*

$$\chi_n(V,Y,k) = \frac{1}{n}\sum_{i=1}^{n} k(y^{(i)},y^{(\sigma_n(i))})$$

converges almost surely to

$$\chi(V, Y, k) = \mathbb{E}_V \mathbb{E}_{\xi,\xi' \sim \mathrm{P}_{Y|V}} k_{\mathcal{Y}}(\xi, \xi')$$

as $n \to \infty$.

The proof relies again on Mercer's theorem and is given in Appendix A.5. The estimators of $\mathbb{E}_{X_l}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|X_l})\right)$ and $\mathrm{MMD}^2_{\mathrm{tot}}$ are finally given by

$$\widehat{\mathrm{MMD}}_l^2 = \frac{1}{n}\sum_{i=1}^{n} k_{\mathcal{Y}}\left(y^{(i)}, y^{(\sigma_n^l(i))}\right) - \frac{1}{n^2}\sum_{i,j=1}^{n} k_{\mathcal{Y}}\left(y^{(i)}, y^{(j)}\right) \tag{20}$$

$$\widehat{\mathrm{MMD}}_{\mathrm{tot}}^2 = \frac{1}{n}\sum_{i=1}^{n} k_{\mathcal{Y}}\left(y^{(i)}, y^{(i)}\right) - \frac{1}{n^2}\sum_{i,j=1}^{n} k_{\mathcal{Y}}\left(y^{(i)}, y^{(j)}\right) \tag{21}$$

for a sample $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$, $i = 1, \ldots, n$ and where σ_n^l is the permutation defined in Eq. (19) with a ranking performed on the sample $\left(x_l^{(i)}\right)_{i=1,\ldots,n}$.

4.2.4 Higher-order index estimation with nearest-neighbors

The ranking approach introduced above can actually be generalized to estimate higher-order sensitivity indices by replacing ranking (in dimension 1) by nearest-neighbors (in arbitrary dimension), since they define a permutation with the same properties as required in Proposition 3. This was proposed independently by Azadkia and Chatterjee (2019) in the context of a dependence measure and by Broto et al. (2020) for Shapley effects estimation. Here we adopt the formalism of Broto et al. (2020), where they introduce $j_A^*(i, m)$ the index such that the sample point $\mathbf{x}_A^{(j_A^*(i,m))}$ of the subset $A \subseteq \mathcal{P}_d$ of input variables is the m-th nearest neighbor of the sample point $\mathbf{x}_A^{(i)}$ in a sample of the inputs $\left(\mathbf{x}^{(i)}\right)_{i=1,\ldots,n}$. Then their nearest-neighbor estimator $\hat{V}_A^{\mathrm{knn}}$ of $\mathrm{Var}\,\mathbb{E}(Y|\mathbf{X}_A)$ is given by

$$\hat{V}_A^{\mathrm{knn}} = \frac{1}{n_A}\sum_{j=1}^{n_A} \eta\left(\mathbf{x}^{(j_A^*(s(j),1))}\right)\eta\left(\mathbf{x}^{(j_A^*(s(j),2))}\right) - \left(\frac{1}{n}\sum_{i=1}^{n}\eta\left(\mathbf{x}^{(i)}\right)\right)^2$$

where $s(j)$, $j = 1, \ldots, n_A$ is a sample of uniformly distributed integers in $\{1, \ldots, n\}$, with $n_A \leq n$. The choice of using a subsample $s(j)$ is motivated by the authors so that their framework is general enough for the different aggregation procedures they propose for Shapley effects and for their consistency proofs. Several numerical experimentations not reported here also show that using all the samples instead of subsamples yield biased estimators, so we follow the procedure of Broto et al. (2020). Once again this estimator can be generalized to MMD-based indices, where $\mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)$ is estimated by

$$\widehat{\mathrm{MMD}}_A^2 = \frac{1}{n_A}\sum_{j=1}^{n_A} k_{\mathcal{Y}}\left(y^{(j_A^*(s(j),1))}, y^{(j_A^*(s(j),2))}\right) - \frac{1}{n^2}\sum_{i,j=1}^{n} k_{\mathcal{Y}}\left(y^{(i)}, y^{(j)}\right)$$

where we denote $y^{(i)} = \eta\left(\mathbf{x}^{(i)}\right)$. The consistency of this estimator directly follows from the consistency of $\hat{V}_A^{\mathrm{knn}}$ from Broto et al. (2020) and Mercer's theorem. Since $j_A^*(s(j), 1) = s(j)$, the estimator is identical to the ranking-based one in (20) where the permutation from rankings is simply replaced by the index of the nearest neightbor not including itself $j_A^*(s(j), 2)$.

4.3 Shapley effect estimation

The last estimation task concerns kernel-embedding Shapley effects set forth in Definition 5. Of course a straightforward approach consists in using any of the estimators discussed before in the general formulation of the MMD- or HSIC-Shapley effects. But a closer inspection actually reveals that although this is easy for the HSIC-Shapley effects since both $\mathrm{HSIC}_u(\mathbf{X}_A, Y)$ and $\mathrm{HSIC}_b(\mathbf{X}_A, Y)$ can be computed for all subsets $A \subseteq \mathcal{P}_d$ with only one sample $\left(\mathbf{x}^{(i)}, y^{(i)}\right)$, $i = 1, \ldots, n$, the MMD-Shapley effects require estimators of $\mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)$ which do not involve two many calls to the numerical model. Among the estimators introduced in Section 4.2, only the one based on nearest neighbors has a computational cost independent of the number of input variables. This is exactly the framework proposed in Broto et al. (2020) for the variance-based Shapley effects.

However, as pointed out in Song et al. (2016) in the case of variance-based Shapley effects, a double-loop Monte-Carlo estimator of the value function $\mathrm{val}(A) = \mathrm{Var}\,\mathbb{E}\left(Y|\mathbf{X}_A\right)/\mathrm{Var}\,Y$ can be heavily biased. They show that another value function $\mathrm{val}'(A) = \mathbb{E}\mathrm{Var}\left(Y|\mathbf{X}_{-A}\right)/\mathrm{Var}\,Y$ behaves better and gives rise to the exact same Shapley effects (Theorem 1 in Song et al. (2016)). This is why Broto et al. (2020) also introduced a nearest neighbor estimator of $\mathbb{E}\mathrm{Var}\left(Y|\mathbf{X}_{-A}\right)$ given by

$$\hat{E}_A^{\mathrm{knn}} = \frac{1}{n_A}\sum_{j=1}^{n_A}\left\{\frac{1}{n_I - 1}\sum_{i=1}^{n}\left[y^{(j^*_{-A}(s(j),i))} - \frac{1}{n_I}\sum_{i=1}^{n} y^{(j^*_{-A}(s(j),i))}\right]^2\right\}$$

where this time n_I nearest neighbors are used. In a nutshell, the nearest neighbors are used as if they were independent samples from $P_{Y|\mathbf{X}_A=\mathbf{x}^{(s(j))}}$, which explains why we compute their empirical variance in the formula above. In order to follow the same road for the estimation of MMD-Shapley effects, we first need an equivalent of Theorem 1 from Song et al. (2016) for a new value function related to the MMD.

Lemma 3 (Other formulation of MMD-Shapley effects)**.** *The Shapley values obtained with value function* $\mathrm{val}'(A) = \mathbb{E}_{\mathbf{X}_{-A}}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_{-A}}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_{-A}}} k_{\mathcal{Y}}(\xi,\xi')\right]/\mathrm{MMD}^2_{\mathrm{tot}}$ *are exactly equal to the MMD-Shapley effects from Definition 5 with value function* $\mathrm{val}(A) = \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)/\mathrm{MMD}^2_{\mathrm{tot}}$.

The proof is based on the generalization of the law of total variance for the generalized variance $\mathrm{MMD}^2_{\mathrm{tot}}$ and is given in Appendix A.6. A nearest neighbor estimator $\widehat{\mathrm{EMMD}}^2_A$ of

$$\mathbb{E}_{\mathbf{X}_{-A}}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_{-A}}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_{-A}}} k_{\mathcal{Y}}(\xi,\xi')\right]$$

is then given by

$$\begin{aligned}\widehat{\mathrm{EMMD}}^2_A \;=\; & \frac{1}{n_A}\sum_{j=1}^{n_A}\left\{\frac{1}{n_I}\sum_{i=1}^{n_I} k_{\mathcal{Y}}\left(y^{(j^*_{-A}(s(j),i))}, y^{(j^*_{-A}(s(j),i))}\right)\right. \\ & \left. -\frac{1}{n_I^2}\sum_{i,i'=1}^{n_I} k_{\mathcal{Y}}\left(y^{(j^*_{-A}(s(j),i))}, y^{(j^*_{-A}(s(j),i'))}\right)\right\}\end{aligned}$$

and the MMD-Shapley effect estimator is

$$\widehat{Sh}_l^{\mathrm{MMD}} = \frac{1}{\widehat{\mathrm{MMD}}^2_{\mathrm{tot}}}\frac{1}{p}\sum_{A\subseteq\mathcal{P}_d,\, A\not\ni l}\binom{p-1}{|A|}^{-1}\left\{\widehat{\mathrm{EMMD}}^2_{A\cup\{l\}} - \widehat{\mathrm{EMMD}}^2_A\right\}.$$

where $\widehat{\mathrm{MMD}}^2_{\mathrm{tot}}$ is estimated as in Eq. (21).

As a side-note, when the number of input variables is large, the number of terms involved in Shapley effects severely increases and the computational cost to assemble all the terms (even if one uses estimators relying on a given sample only) becomes prohibitive. For such cases it is possible to use a formulation of Shapley effects involving a sum on permutations of $\{1, \ldots, d\}$ instead of a sum on subsets of $\mathcal{P}_d$, which makes it possible to add another level of approximation by computing the sum on a random sample of permutations instead of on all of them (Castro et al., 2009). Obviously since this trick does not depend on the value function used inside the Shapley values, it can also be used for our kernel-embedding Shapley effects.

5 Experiments

In this section we illustrate the behavior of the kernel-based sensitivity indices on several test cases representative of typical GSA industrial applications. In particular, we address the following numerical model categories: a standard scalar output model, a stochastic simulator, a model with a time-series output and a multi-class categorical output simulator with dependent inputs. All the results presented here are reproducible with the R code provided in the supplementary material.

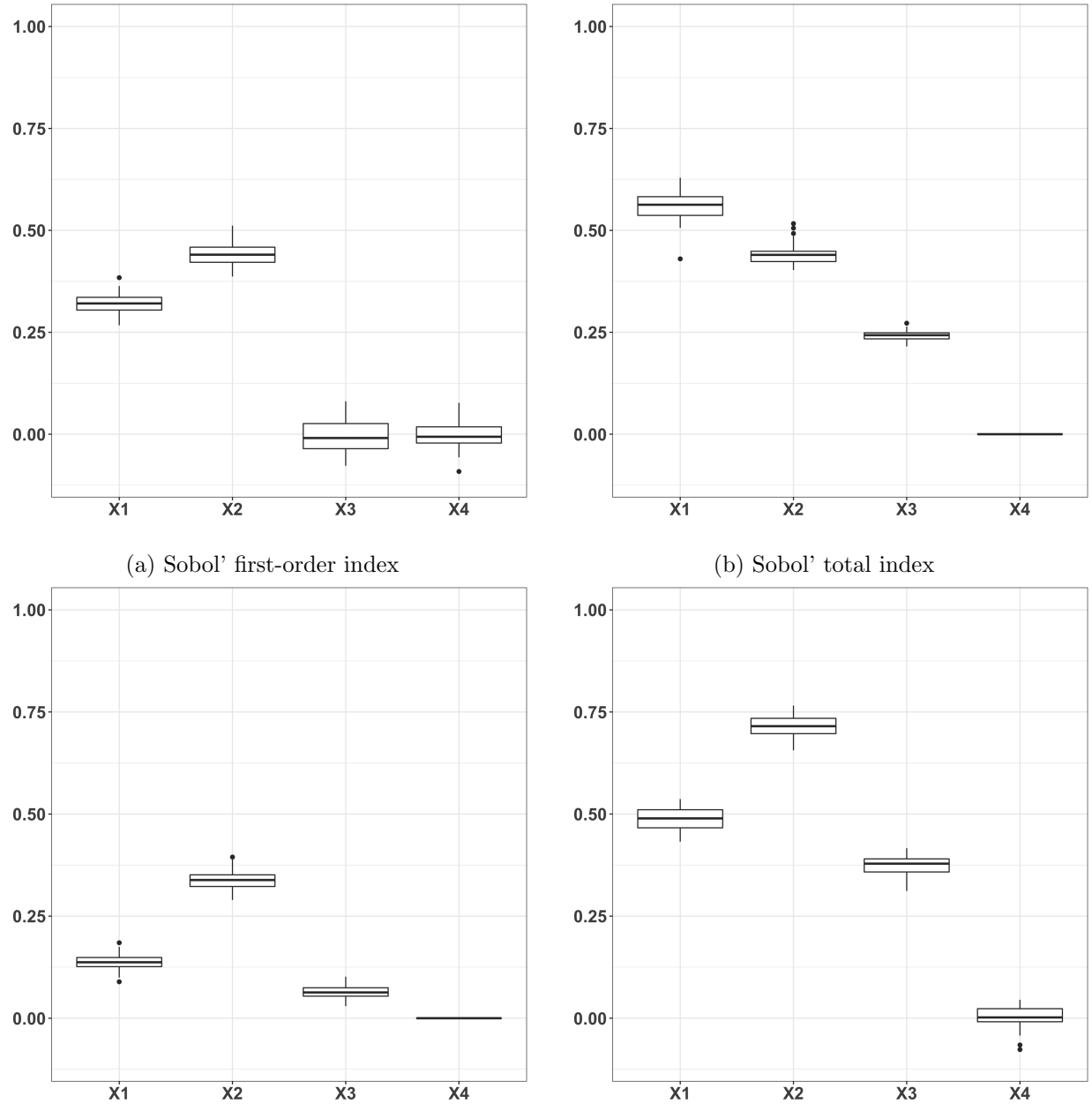
5.1 Standard scalar output model

To exemplify the additional insight provided by these indices we first consider a classical GSA test case, the Ishigami function (Ishigami and Homma, 1990) where the output Y is given by

$$Y = \sin(X_1) + 7\sin(X_2)^2 + X_3^4 \sin(X_1)$$

where $X_l \sim \mathcal{U}(-\pi, \pi)$ for $l = 1, \ldots, 4$, meaning that we add a dummy input variable X_4 for analysis purposes.

We start by computing the traditional Sobol' first-order and total sensitivity indices using a pick-freeze estimator as in Section 4.2.2 with a sample size $n = 1000$ and we repeat this calculation 50 times. For each replication the total number of calls to the numerical model is thus $(p + 2)n = 6000$. We then use the same pick-freeze procedure to estimate the MMD-based first-order and total indices with the exact same samples. For the output we use a Gaussian kernel $k_{\mathcal{Y}}(y, y') = \exp(-\frac{1}{2\sigma^2}(y - y')^2)$ where σ is chosen as the median of the pairwise distances between the output samples. Results are given in Figure 1. First note that, as is well known, the first-order Sobol' index of X_3 is zero, while its total index is around 0.25 due to its interaction with X_1. X_2 is also an important variable, which does not have any interaction since its total Sobol' index is equal to its first-order one. As expected X_4 is correctly detected as non-important. The MMD-based indices however bring a different insight: from a probability distribution perspective, one can observe that interactions are much more present since there is a large gap between total and first-order indices for all inputs (except X_4 of course). In addition, this time X_3 is detected to have a main effect: indeed even though it does not impact the output conditional mean, it influences the tails of the output conditional distribution when it is close to $-2\pi/2\pi$ as was already illustrated in Da Veiga (2016). This shows that MMD-based indices capture other types of input influence than Sobol' ones.



(a) Sobol' first-order index

(b) Sobol' total index

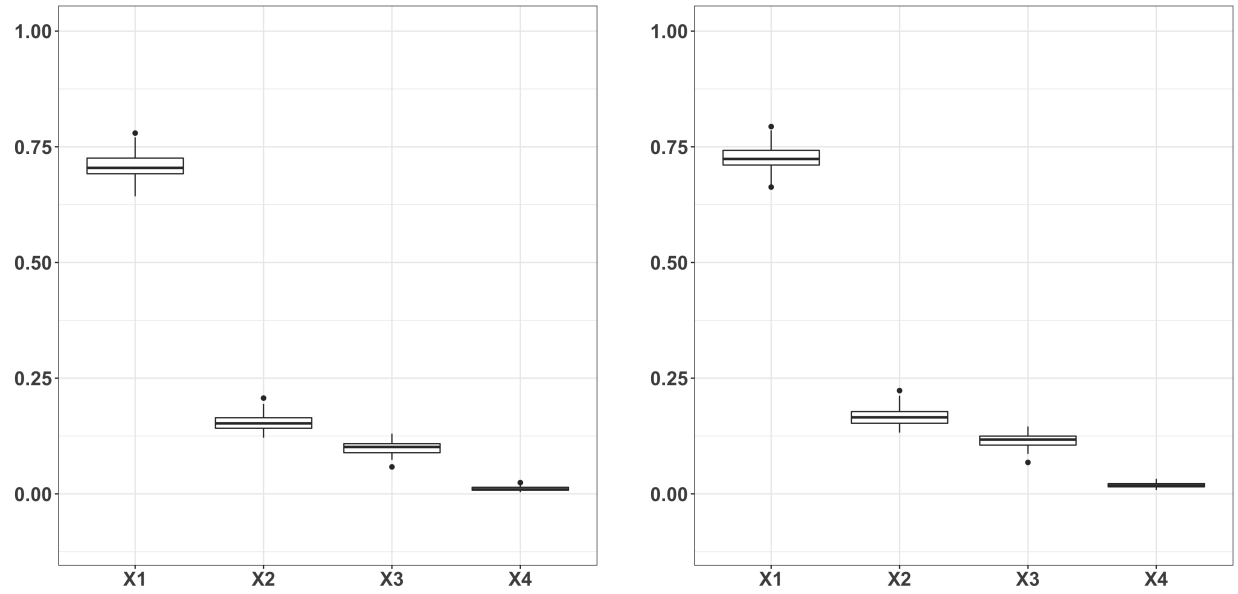
(c) MMD-based first-order index

(d) MMD-based total index

Figure 1: Ishigami test case. First-order (a) and total (b) Sobol' indices and first-order (c) and total (d) MMD-based indices with pick-freeze estimators, $n = 1000$, 50 replicates.

To take a different view at the inputs/output relationship we also estimate HSIC-based first-order and total indices using the V-statistic of Section 4.1. Again for the output we use the same Gaussian kernel as above, while we use the Sobolev kernel from Eq. (12) for the inputs. Since they are uniform it is easy to renormalize them to satisfy the zero-mean kernel condition. We use only one sample of size $n = 1000$ and estimates obtained with 50 replications are reported in Figure 2.

Interestingly, we observe first that with HSIC we no longer detect any interaction: our intuition is that first-order HSIC indices already aggregate a very large family of potential influences and thus interactions may only appear with highly complicated inputs/output link functions. This is supported by the fact that HSIC indices rank the inputs the exact same way at total Sobol' indices. Another appealing property is that to compute all HSIC indices we only need a given sample of moderate size, which is interesting from a screening perspective for GSA on very time-consuming numerical models.



(a) HSIC-based first-order index

(b) HSIC-based total index

Figure 2: Ishigami test case. First-order (a) and total (b) HSIC-based indices with V-statistic estimator, $n = 1000$, 50 replicates.

5.2 Stochastic simulator

Our second illustration is a more original setting for GSA which consists of a stochastic simulator where the numerical model outputs a probability distribution, or rather a sample from a probability distribution in practice, for a fixed value of the input variables. Here we use a test case proposed in Moutoussamy et al. (2015) which involves five input variables and writes

$$Y = (X_1 + 2X_2 + U_1)\sin(3X_3 - 4X_4 + N) + U_2 + 5X_5 B + \sum_{i=1}^{5} iX_i$$

where $X_1, \ldots, X_5 \sim \mathcal{U}(0,1)$ are the input variables and $U_1 \sim \mathcal{U}(0,1)$, $U_2 \sim \mathcal{U}(1,2)$, $N \sim \mathcal{N}(0,1)$ and $B \sim$ Bernoulli$(1/2)$ are additional random variables which are responsible for the simulator stochasticity. Note that we modify the constant in front of $X_5 B$ to lessen the effect of X_5 as compared to Moutoussamy et al. (2015). An example of the output distribution for 20 random fixed values of the input variables obtained each time with a sample of size 100 for the stochastic ones is

given in Figure 3.

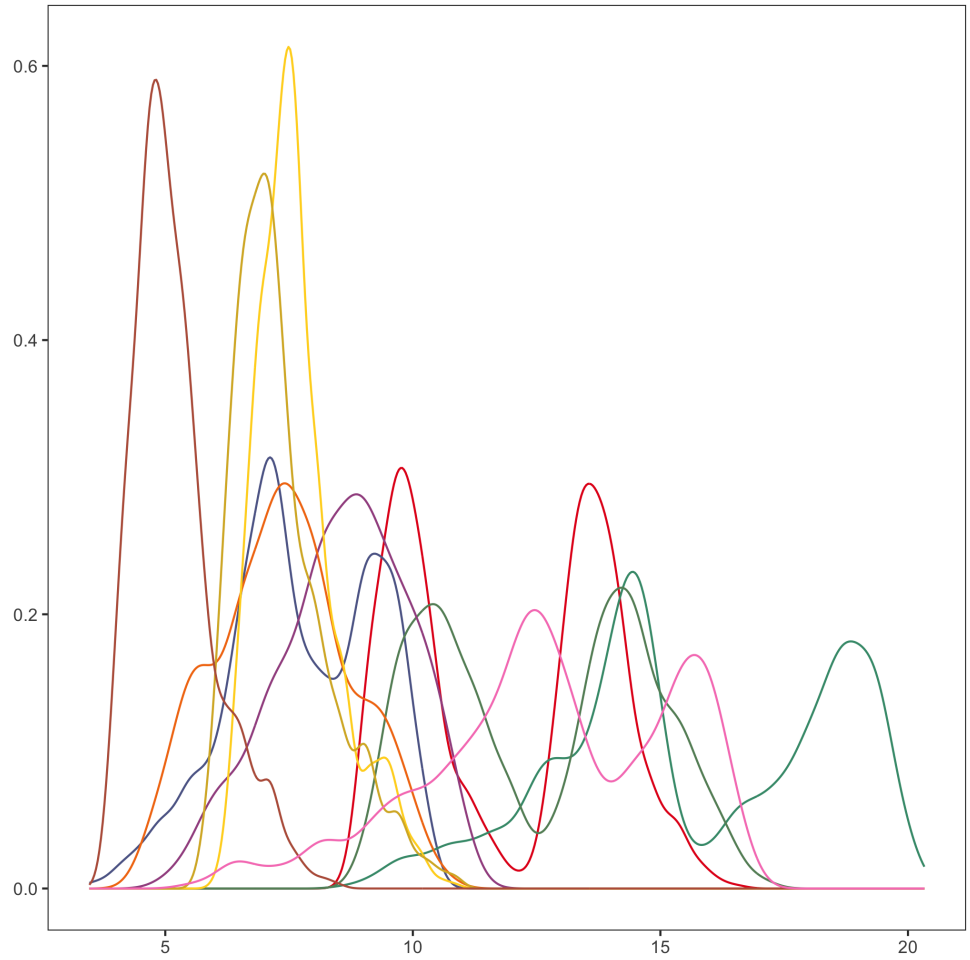
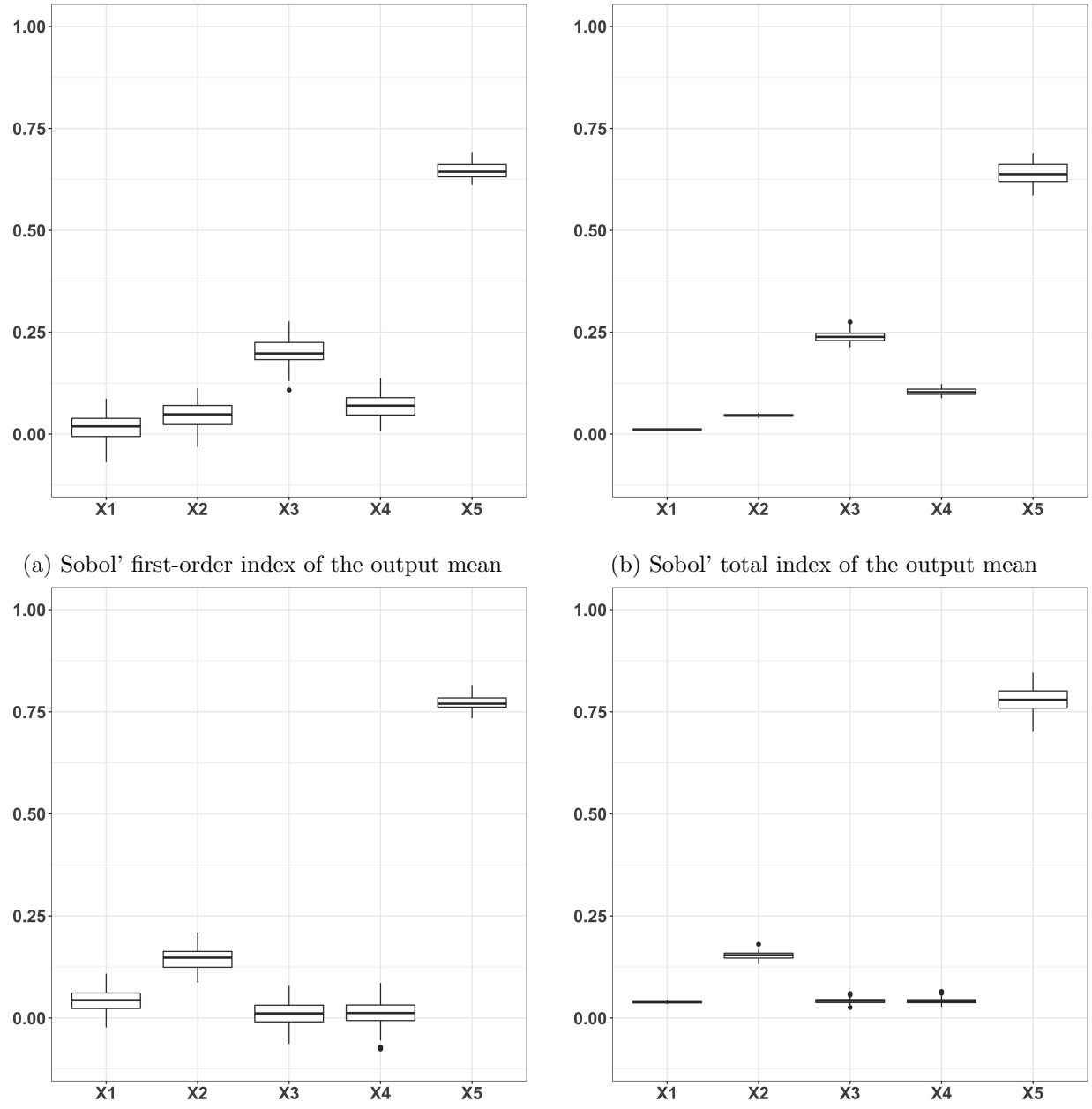


Figure 3: Stochastic simulator test case. Output probability distribution for 20 values of the input variables chosen at random. The distribution is estimated with a kernel-density estimator.

Leaving aside for now the whole output distribution, we first place ourselves in a standard GSA deterministic setting by first analyzing the input influence on both the output mean and standard deviation (with respect to U_1, U_2, N and B). We thus compute Sobol' indices for these two outputs of interest with a pick-freeze estimator with a sample of size $n = 1000$ and perform 50 replications, see Figure 4. It shows that interactions are negligible, and that X_5 is clearly the most influential input by far: it explains alone 65% of the output mean variability and 75% of the output standard deviation variability. This is expected since X_5 is coupled with B, which creates the multi-modal feature of the output distribution. The output mean variability also depends on X_3 and X_4 to some lesser extent, and the output standard deviation variability on X_2.

We now make use of the kernel framework to compute MMD- and HSIC-based sensitivity indices



(a) Sobol' first-order index of the output mean

(b) Sobol' total index of the output mean

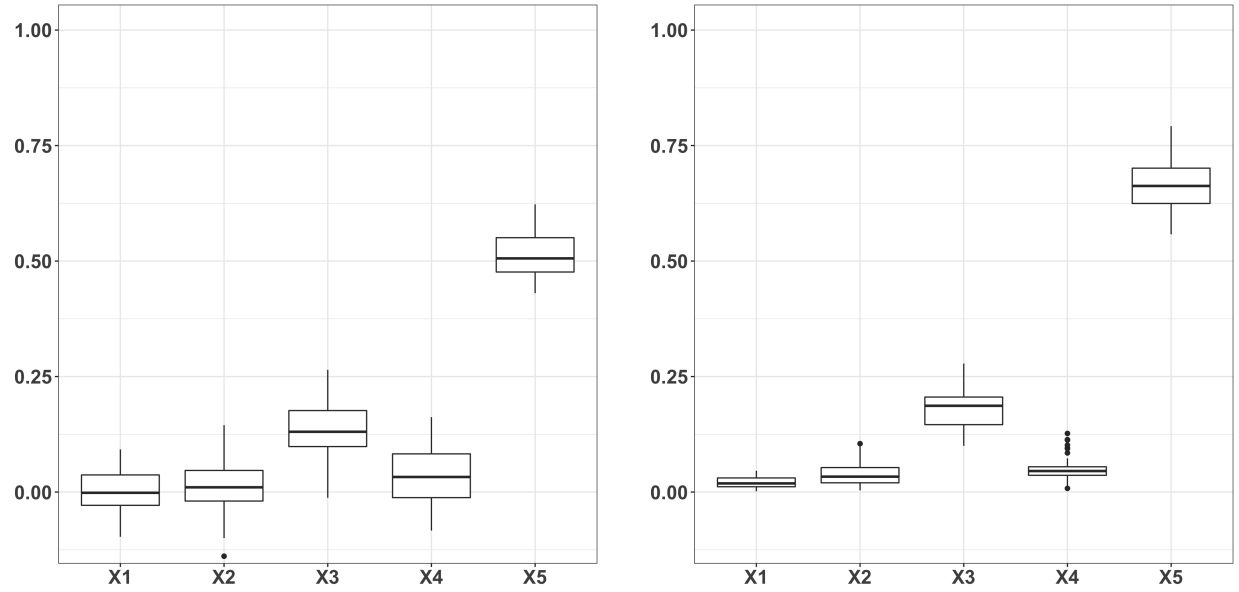
(c) Sobol' first-order index of the output standard deviation

(d) Sobol' total index of the output standard deviation

Figure 4: Stochastic simulator test case. First-order (a) and total (b) Sobol' indices of the output mean and first-order (c) and total (d) Sobol' indices of the output standard deviation with pick-freeze estimators, $n = 1000$, 50 replicates.

which can accommodate directly the output distribution thanks to the specific kernels discussed in Section 3.4. More precisely we use the kernel of Eq. (18) with $\sigma^2 = 1$ and λ chosen as the median of the MMD^2 computed on the preliminary sample used for visualization in Figure 3, with a kernel

$k_{\mathcal{Y}}(y, y') = \exp(-\frac{1}{2\tau^2}(y - y')^2)$ and τ chosen as the median of the pairwise distances between the output samples. We only compute first-order indices here and use for illustration the rank estimator of the MMD index from Section 4.2.3 while for HSIC we use again the Sobolev kernel. Results with 50 replications and a sample of size $n = 200$ are given in Figure 5. Both indices coincide and identify X_5 as the most important input variable, as well as a small influence of X_3, X_2 and X_4 while X_1 is non-important: considering the whole output distribution variability via the specific kernel is comparable to an aggregation of the variability on the output mean and standard deviation (and other moments we did not compute above).



(a) MMD first-order index

(b) HSIC first-order index

Figure 5: Stochastic simulator test case. First-order MMD (a) and HSIC (b) indices of the output distribution with rank and V-statistic estimators, respectively, $n = 200$, 50 replicates.

5.3 Functional output

Another commonly encountered industrial application is a physics-based numerical simulator involving functional outputs, such as curves representing the evolution over time of some system characteristics (*e.g.* pressure, temperature, ...). To illustrate how time-series kernels can easily handle GSA on such systems we build a simplified compartmental epidemiological model inspired by previous works on COVID-19 (Magal and Webb, 2020; Charpentier et al., 2020; Di Domenico et al., 2020). Our model is a straightforward Susceptible - Infected - Recovered (SIR) model (Kermack and McKendrick, 1927) which is slightly modified, in the sense that it accounts for two different types of infectious people: the reported cases, which we assume are isolated and can non longer contaminate others, and the unreported cases who can infect others. A summary of this compartment model proposed by Magal and Webb (2020) is given in Figure 6.

S consists of the susceptible individuals who are not yet infected. During the epidemic spread they are infected depending on the time-dependent transmission rate $\tau(t)$. Once infected they

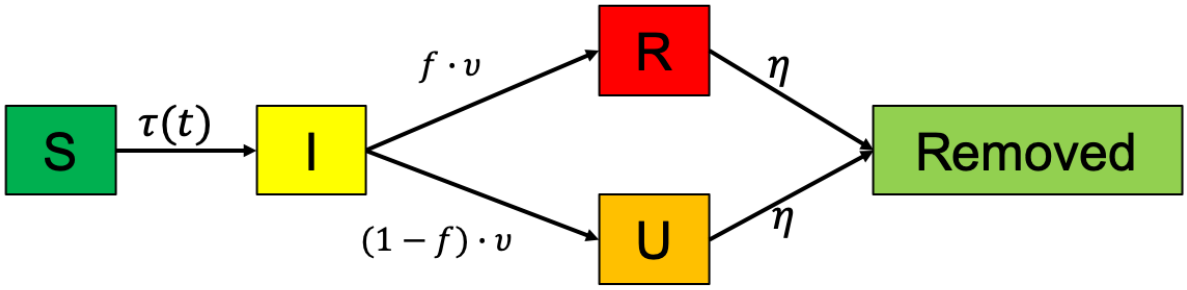


Figure 6: Functional simulator test case. The modified SIR model with 4 compartments following Magal and Webb (2020) .

mode to compartment I where are the asymptomatic infectious individuals. After a period of η days they become symptomatic and a fraction f of them is detected and go to compartment R, while the rest of them are undetected and go to compartment U. After a recovering period of η days symptomatic people from R and U recover and go to the last compartment. Observe that this is a highly simplified representation of the epidemic where we do not account for hospitalizations, testing strategies or deaths: our goal here is not to be representative of COVID-19 but rather exemplify how GSA can be applied to such models.

The dynamics of the evolution of individuals from a compartment to another is modeled with the following system of ordinary differential equations:

$$\begin{aligned}
\frac{dS}{dt} &= -\tau S(I+U) \\
\frac{dI}{dt} &= \tau S(I+U) - \nu I \\
\frac{dR}{dt} &= f\nu I - \eta R \\
\frac{dU}{dt} &= (1-f)\nu I - \eta U
\end{aligned}$$

The transmission rate is chosen according to Magal and Webb (2020) where they propose a parametric form given by $\tau(t) = \tau_0 \exp(-\mu \max(t - N, 0))$. The underlying assumption is that before the epidemic outbreak the transmission rate is constant equal to τ_0 and it then decreases with an exponential decay with rate μ once social distancing and lockdown start to have an effect after N days. They further assume that the cumulative number of reported cases $CR(t)$ is approximately

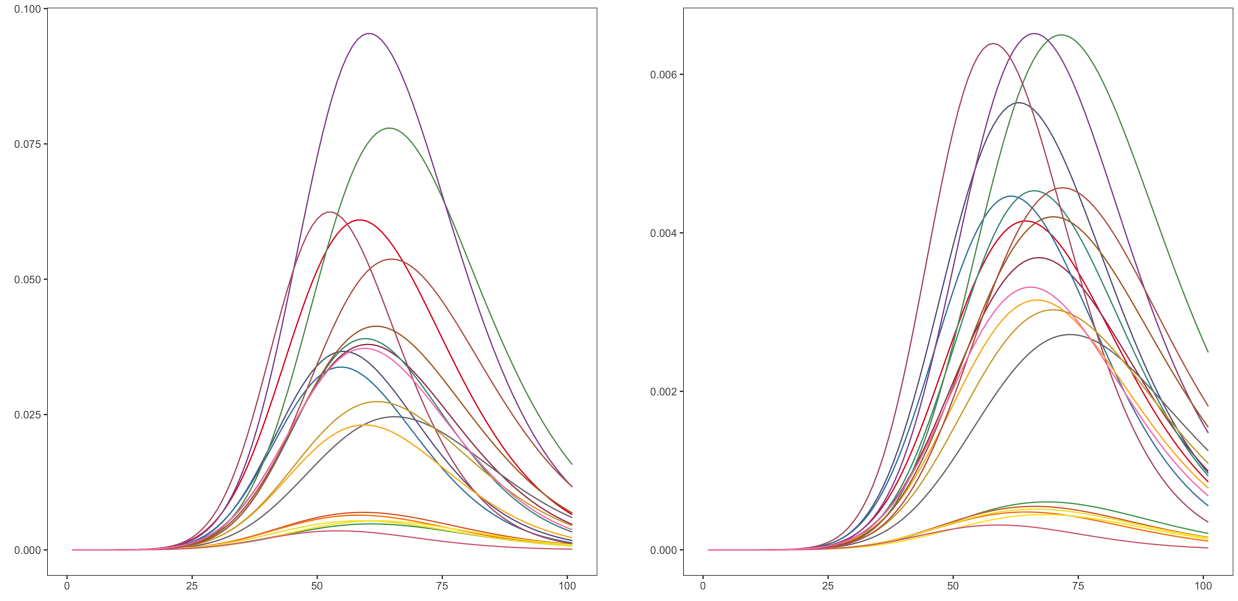
$$CR(t) = \chi_1 \exp(\chi_2 t) - 1$$

where χ_1 and χ_2 are to be estimated on data. From this assumption they get the value of the initial conditions I_0, U_0, R_0

$$I_0 = \frac{\chi_2}{f\nu},\ U_0 = \frac{(1-f)\nu}{\eta + \chi_2} I_0,\ R_0 = 1$$

and with in our case $S_0 = 66.99 \times 10^6$ is the initial susceptible population (here in France). From a GSA perspective we then assume that we have uncertainty on the following 6 input variables: τ_0, μ,

N (transmission rate), η, ν (days until symptoms and recovery) and χ_2 (which impacts the initial conditions). f is assumed to be fixed at a fraction equal to 0.1. We assign uniform distributions to the input variables with ranges consistent with the values from Magal and Webb (2020), *i.e.* $\tau_0 \sim \mathcal{U}(5.9 \times 10^{-9}, 6.1 \times 10^{-9})$, $\mu \sim \mathcal{U}(0.028, 0.036)$, $N \sim \mathcal{U}(8, 15)$, $1/\eta \sim \mathcal{U}(5, 9)$, $1/\nu \sim \mathcal{U}(5, 9)$ and $\chi_2 \sim \mathcal{U}(0.32, 0.4)$. An example of the dynamics of compartments I and R for 20 values of the inputs chosen at random according to these uniform distributions is given in Figure 7.



(a) Infectious cases

(b) Reported cases

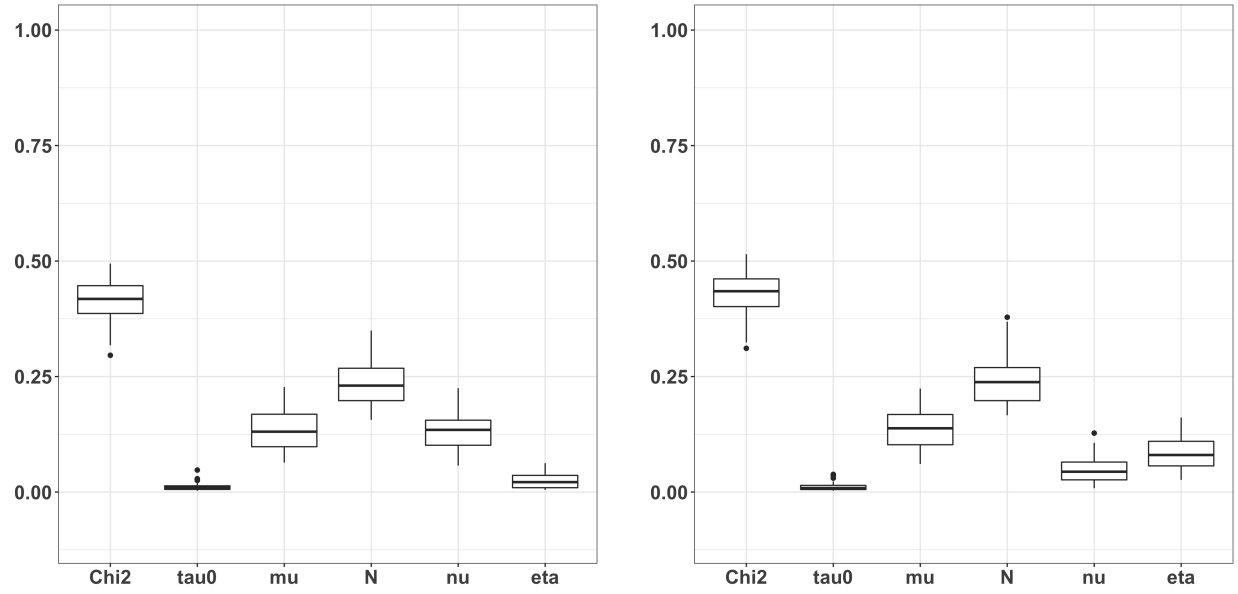
Figure 7: Functional simulator test case. Output dynamics over time for compartment I (left) and R (right) for 20 values of the input variables chosen at random. They are both normalized by the total population S_0.

For GSA we rely on the global-alignment kernel of Cuturi (2011) designed for time-series, which searches for all their alignments and averages them, and use it inside our first-order HSIC indices for the output, whereas we still employ the Sobolev kernel for the inputs. The results obtained with 50 repetitions with a sample of size $n = 200$ and the V-statistic estimator are reported in Figure 8.

For both compartments the most influential input is χ_2, as expected since it influences the initial conditions, and then N and μ related to the transmission rate. ν has also an impact for compartment I but not η, which is coherent with the ordinary differential equations, and one can see on the contrary that η influences compartment R.

5.4 Multi-class output with dependent inputs

Finally we investigate a numerical model with both a categorical output (to make use of the discussion from Section 3.4) and dependent inputs (to analyze kernel-embedding Shapley effects from Section 3.3). We build upon the famous wine quality data set (Cortez et al., 2009) of the UCI repository (Dua and Graff, 2017). This dataset consists of 4898 observations of wine qualities



(a) First-order HSIC index for compartment I

(b) First-order HSIC index for compartment R

Figure 8: Functional simulator test case. First-order HSIC index for compartments I (left) and R (right) with V-statistics estimator, $n = 200$, 50 replicates.

(categorical variable with levels 0 to 10 corresponding to a score) associated to 11 features obtained with physicochemical tests. In order to place ourselves in a standard computer experiments setting (*i.e.* a numerical simulator and uncertain inputs with given probability distribution) we use this dataset to design a GSA scenario detailed in the following steps:

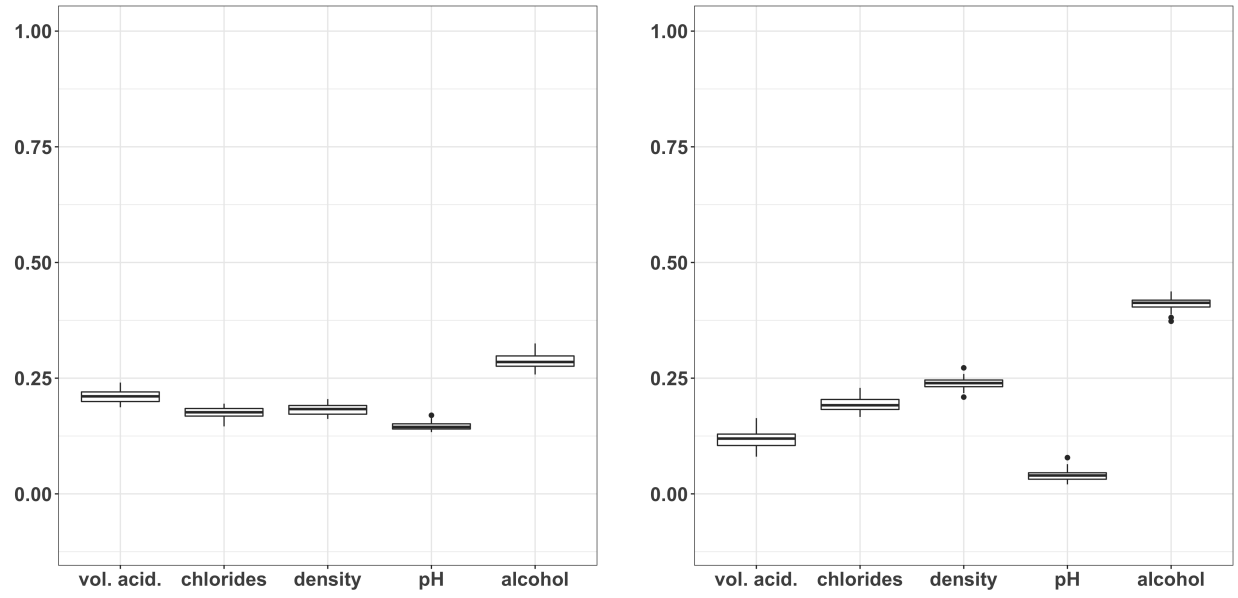
1. We regroup wine quality scores into only 3 categories: low (score less than 5), medium (score equal to 6) and high (score higher than 7) in order to have a balanced dataset. We also use a small subsample of size 600 of white wine only from the initial 4898 observations for faster estimation of the input dependence structure;

2. We estimate a random forest model between the wine quality and the 11 inputs from this transformed dataset and compute variable importance for each input. The variable importance score is used to select only 4 important features among the initial 11 ones (volatile acidity, chlorides, density and alcohol). This is absolutely not a mandatory step, but we choose to do so for both a faster computation of Shapley effects and estimation of the input dependence structure. A new random forest model is finally built with these 4 input variables only, and the predictor serves as our numerical simulation model;

3. The samples from the 4 input variables identified above are used to estimate a vine copula structure which models their dependence (Czado, 2019). Once the vine copula is estimated, it is then easy to generate new input samples as much as required.

MMD- and HSIC-Shapley effects are then computed with a sample size of $n = 1000$ with a dirac categorical kernel for the output and a Sobolev kernel for the inputs in the HSIC case. For

MMD we use the nearest-neighbor estimator of Section 4.3 and for HSIC the V-statistic estimator and we repeat the estimation 50 times, see Figure 9.



(a) MMD-Shapley effect

(b) HSIC-Shapley effect

Figure 9: Multi-class output test case. MMD- (a) and HSIC- (b) Shapley effects with nearest-neighbor and V-statistic estimators, respectively, $n = 1000$, 50 replicates.

Both kernel-embedding Shapley effects identify alcohol as the most influential input, which was expected from the variable importance scores computed with the random forest. However, the MMD-Shapley effects do not discriminate as clearly the input variables as HSIC. We suspect that there may be remaining estimation bias coming from the nearest-neighbor estimators which we plan to carefully examine in future work.

6 Conclusion

In this paper we discussed two moment-independent sensitivity indices which generalize Sobol' ones by relying on the RKHS embedding of probability distributions. These MMD- and HSIC-based sensitivity indices are shown to admit an ANOVA-decomposition, which makes it possible to properly define input interactions and their natural normalization constant. To the best of our knowledge this is the first time such a result is proved for sensitivity indices apart from Sobol' ones. We also defined kernel-embedding Shapley effects which are built upon these indices for the case where the input variables are no longer independent. As discussed through several GSA applications with categorical outputs or stochastic simulators, this opens the path for new powerful and general GSA approaches by means of kernels adapted to the task at hand. Finally, several estimators have been introduced, including new ones inspired by recent advances in Sobol' indices and Shapley effects estimation.

However, there is still room for improvement in the theoretical understanding of theses indices. First, we extensively used Mercer's theorem and it would be interesting to extend our results when

it no longer holds. We also assume a kernel product form for HSIC indices, whereas the theorem used in our proof allows for more general kernels. From an estimation perspective, we did not exhibit here any central limit theorem, although this would be an important step enabling to statistically test whether indices are zero or not. But this is not at all an easy task, which may be tackled via the functional delta method combined with Mercer's theorem. On the other hand, some bias can be observed in the nearest neighbor estimators, which should be analyzed carefully in future work. Finally, further practical experimentations should be performed to better understand the behavior of these new indices. We think in particular to the choice of the kernel hyperparameters, and the investigation of invariant kernels for outputs given as curves or images.

A Proofs

A.1 Proof of Theorem 3

Proof. The theorem is proved in the case where Mercer's theorem holds, *i.e.*, the output is assumed to be such that $Y \in \mathcal{Y}$ with $\mathcal{Y}$ a compact set and $k_{\mathcal{Y}}$ has the representation

$$k_{\mathcal{Y}}(y, y') = \sum_{r=1}^{\infty} \phi_r(y)\phi_r(y')$$

as in Eq. (9). Consider now the random variable $W = \sum_{r=1}^{\infty} \eta^{[r]}(\mathbf{X})$ where $\eta^{[r]}(\mathbf{X}) = \phi_r(Y) = \phi_r(\eta(\mathbf{X}))$. To prove the theorem, two formulations of $\operatorname{Var} W$ are exhibited. First, since the functions ϕ_r are orthogonal in $\mathbb{L}^2(\mathcal{Y})$ and using the absolute convergence of the series, we have

$$\begin{aligned}
\operatorname{Var} W &= \sum_{r=1}^{\infty} \operatorname{Var} \phi_r(Y) \\
&= \sum_{r=1}^{\infty} \mathbb{E}\left(\phi_r(Y)\phi_r(Y)\right) - \sum_{r=1}^{\infty} \mathbb{E}\left(\phi_r(Y)\phi_r(Y')\right) \\
&= \mathbb{E}\left(\sum_{r=1}^{\infty} \phi_r(Y)\phi_r(Y)\right) - \mathbb{E}\left(\sum_{r=1}^{\infty} \phi_r(Y)\phi_r(Y')\right) \\
&= \mathbb{E}k(Y, Y) - \mathbb{E}k(Y, Y').
\end{aligned}$$

On the other hand, using the variance decomposition (1) for each $\eta^{[r]}(\mathbf{X}) = \phi_r(\eta(\mathbf{X}))$ we get

$$\begin{aligned}
\operatorname{Var} W &= \sum_{r=1}^{\infty} \operatorname{Var} \eta^{[r]}(\mathbf{X}) \\
&= \sum_{r=1}^{\infty} \sum_{A \subseteq \mathcal{P}_d} \sum_{B \subset A} (-1)^{|A|-|B|} \operatorname{Var} \mathbb{E}\left(\eta^{[r]}(\mathbf{X}) | \mathbf{X}_B\right) \\
&= \sum_{A \subseteq \mathcal{P}_d} \sum_{B \subset A} (-1)^{|A|-|B|} \sum_{r=1}^{\infty} \operatorname{Var} \mathbb{E}\left(\phi_r(Y) | \mathbf{X}_B\right) \\
&= \sum_{A \subseteq \mathcal{P}_d} \sum_{B \subset A} (-1)^{|A|-|B|} \mathbb{E}_{\mathbf{X}_B}\left(\operatorname{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_B})\right)
\end{aligned}$$

using again the absolute continuity and the expansion of $\mathbb{E}_{\mathbf{X}_B}\left(\operatorname{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_B})\right)$ obtained in Eq. (9). The theorem follows by equating both formulations of $\operatorname{Var} W$. □

A.2 Proof of Proposition 1

Proof. We simply add and subtract $\mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi')$:

$$\begin{aligned}
\mathrm{MMD}^2_{\mathrm{tot}} &= \mathbb{E}_{\zeta\sim \mathrm{P}_Y} k_{\mathcal{Y}}(\zeta,\zeta) - \mathbb{E}_{\zeta,\zeta'\sim \mathrm{P}_Y} k_{\mathcal{Y}}(\zeta,\zeta') \\
&= \mathbb{E}_{\zeta\sim \mathrm{P}_Y} k_{\mathcal{Y}}(\zeta,\zeta) - \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi') + \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y,\mathrm{P}_{Y|\mathbf{X}_A})\right) \\
&= \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\mathbf{X}_A}\mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi') + \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y,\mathrm{P}_{Y|\mathbf{X}_A})\right) \\
&= \mathbb{E}_{\mathbf{X}_A}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi')\right] + \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y,\mathrm{P}_{Y|\mathbf{X}_A})\right).
\end{aligned}$$

□

A.3 Proof of Theorem 4

Proof. We first rewrite HSIC between $\mathbf{X}$ and Y from Eq. (8) as a multivariate integral, assuming $\mathrm{P}_{\mathbf{X}Y}$ is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}\times\mathcal{Y}$:

$$\begin{aligned}
\mathrm{HSIC}(\mathbf{X},Y) &= \int_{\mathcal{X}\times\mathcal{X}}\int_{\mathcal{Y}\times\mathcal{Y}} k_{\mathcal{X}}(\mathbf{x},\mathbf{x}')k_{\mathcal{Y}}(y,y')\left[p_{\mathbf{X}Y}(\mathbf{x},y) - p_{\mathbf{X}}(\mathbf{x})p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}Y}(\mathbf{x}',y') - p_{\mathbf{X}}(\mathbf{x}')p_Y(y')\right] d\mathbf{x}d\mathbf{x}'dydy' \qquad (22)
\end{aligned}$$

where $p_{\mathbf{X}Y}$, $p_{\mathbf{X}}$ and p_Y are the probability density functions of $(\mathbf{X},Y)$, $\mathbf{X}$ and Y, respectively. As in Theorem 3 we further assume Mercer's theorem holds, which means that

$$k_{\mathcal{Y}}(y,y') = \sum_{r=1}^{\infty}\phi_r(y)\phi_r(y').$$

For each r, we then define the function

$$g^{[r]}(\mathbf{x}) = \int_{\mathcal{X}}\int_{\mathcal{Y}} k_{\mathcal{X}}(\mathbf{x},\mathbf{x}')\phi_r(y')\left[p_{\mathbf{X}Y}(\mathbf{x}',y') - p_{\mathbf{X}}(\mathbf{x}')p_Y(y')\right] d\mathbf{x}'dy',$$

noting that $g^{[r]}\in\mathcal{F}$ from Assumption 2. It is then straightforward to show that

$$\begin{aligned}
\|g^{[r]}\|^2_{\mathcal{F}} &= \int_{\mathcal{X}\times\mathcal{X}}\int_{\mathcal{Y}\times\mathcal{Y}} k_{\mathcal{X}}(\mathbf{x},\mathbf{x}')\phi_r(y)\phi_r(y')\left[p_{\mathbf{X}Y}(\mathbf{x},y) - p_{\mathbf{X}}(\mathbf{x})p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}Y}(\mathbf{x}',y') - p_{\mathbf{X}}(\mathbf{x}')p_Y(y')\right] d\mathbf{x}d\mathbf{x}'dydy',
\end{aligned}$$

which means that

$$\mathrm{HSIC}(\mathbf{X},Y) = \sum_{r=1}^{\infty}\|g^{[r]}\|^2_{\mathcal{F}} \qquad (23)$$

since in Mercer's theorem we have the absolute convergence of the series. Now the idea is to write an orthogonal decomposition (in $\mathcal{F}$) for each function $g^{[r]}$, which will finally provide a decomposition for HSIC through Eq. (23).

The orthogonal decomposition of $g^{[r]}$ is obtained with Theorem 4.1 from Kuo et al. (2010). First, recall that we have from the first part of Assumption 3:

$$k_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{p} \left(1 + k_l(x_l, x_l')\right) = \sum_{A \subseteq \mathcal{P}_d} \prod_{l \in A} k_l(x_l, x_l') := \sum_{A \subseteq \mathcal{P}_d} k_A(\mathbf{x}_A, \mathbf{x}_A') \tag{24}$$

which corresponds to Eq. (4.1) in Kuo et al. (2010). We then introduce a set of commuting projections $\{P_l\}_{l=1}^{p}$ on $\mathcal{F}$ given by

$$P_l(f) = \int_{\mathcal{X}_l} f(x_1, \ldots, x_{l-1}, t, x_{l+1}, \ldots, x_d) p_{X_l}(t) dt \tag{25}$$

for all $f \in \mathcal{F}$. From the second part of Assumption 3, one has for all subset $A \subseteq \mathcal{P}_d$ and $\mathbf{x}_A \in \mathcal{X}_A$

$$P_l\left(k_A(\cdot, \mathbf{x}_A)\right) = \prod_{l' \in A, l' \neq l} k_{l'}(\cdot, x_{l'}) \int_{\mathcal{X}_l} k_l(t, x_l) p_{X_l}(t) dt = 0$$

for $l \in A$, meaning that Eq. (4.5) from Kuo et al. (2010) is satisfied. From Theorem 4.1 from Kuo et al. (2010), we can now state that $g^{[r]}(\mathbf{x})$ has an unique orthogonal decomposition given by

$$g^{[r]} = \sum_{A \subseteq \mathcal{P}_d} g_A^{[r]}$$

where

$$g_A^{[r]} = \sum_{B \subseteq A} (-1)^{|A|-|B|} P_{-B}(g^{[r]})$$

with $P_{-B} = \prod_{l \notin B} P_l$. Since the decomposition is orthogonal, we further have

$$\|g^{[r]}\|_{\mathcal{F}}^2 = \sum_{A \subseteq \mathcal{P}_d} \|g_A^{[r]}\|_{\mathcal{F}}^2$$

and

$$\|g_A^{[r]}\|_{\mathcal{F}}^2 = \sum_{B \subseteq A} (-1)^{|A|-|B|} \|P_{-B}(g^{[r]})\|_{\mathcal{F}}^2.$$

The last part is to expand $\|P_{-B}(g^{[r]})\|_{\mathcal{F}}^2$. We first write the projection:

$$\begin{aligned}
P_{-B}(g^{[r]}) &= \int_{\mathcal{X}}\int_{\mathcal{Y}}\int_{\mathcal{X}_{-B}} k_{\mathcal{X}}(\mathbf{x},\mathbf{x}')p_{\mathbf{X}_{-B}}(\mathbf{x}_{-B})\phi_r(y')\left[p_{\mathbf{X}Y}(\mathbf{x}',y') - p_{\mathbf{X}}(\mathbf{x}')p_Y(y')\right]d\mathbf{x}'dy'd\mathbf{x}_{-B} \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}}\left(\prod_{l\notin B}\int_{\mathcal{X}_l}\left(1+k_l(x_l,x_l')\right)p_{X_l}(x_l)dx_l\right)\prod_{l\in B}\left(1+k_l(x_l,x_l')\right)\phi_r(y') \\
&\quad \left[p_{\mathbf{X}Y}(\mathbf{x}',y') - p_{\mathbf{X}}(\mathbf{x}')p_Y(y')\right]d\mathbf{x}'dy' \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}}\prod_{l\in B}\left(1+k_l(x_l,x_l')\right)\phi_r(y')\left[p_{\mathbf{X}Y}(\mathbf{x}',y') - p_{\mathbf{X}}(\mathbf{x}')p_Y(y')\right]d\mathbf{x}'dy' \\
&= \int_{\mathcal{X}_B}\int_{\mathcal{X}_{-B}}\int_{\mathcal{Y}}\prod_{l\in B}\left(1+k_l(x_l,x_l')\right)\phi_r(y')\left[p_{\mathbf{X}Y}(\mathbf{x}',y') - p_{\mathbf{X}}(\mathbf{x}')p_Y(y')\right]d\mathbf{x}_B'd\mathbf{x}_{-B}'dy' \\
&= \int_{\mathcal{X}_B}\int_{\mathcal{Y}}\prod_{l\in B}\left(1+k_l(x_l,x_l')\right)\phi_r(y')\left(\int_{\mathcal{X}_{-B}}\left[p_{\mathbf{X}Y}(\mathbf{x}',y') - p_{\mathbf{X}}(\mathbf{x}')p_Y(y')\right]d\mathbf{x}'_{-B}\right)d\mathbf{x}_B'dy' \\
&= \int_{\mathcal{X}_B}\int_{\mathcal{Y}}\prod_{l\in B}\left(1+k_l(x_l,x_l')\right)\phi_r(y')\left[p_{\mathbf{X}_B Y}(\mathbf{x}_B',y') - p_{\mathbf{X}_B}(\mathbf{x}_B')p_Y(y')\right]d\mathbf{x}_B'dy' \\
&= \int_{\mathcal{X}_B}\int_{\mathcal{Y}} k_B(\mathbf{x}_B,\mathbf{x}_B')\phi_r(y')\left[p_{\mathbf{X}_B Y}(\mathbf{x}_B',y') - p_{\mathbf{X}_B}(\mathbf{x}_B')p_Y(y')\right]d\mathbf{x}_B'dy'
\end{aligned}$$

and its norm then equals

$$\begin{aligned}
\|P_{-B}(g^{[r]})\|_{\mathcal{F}}^2 &= \int_{\mathcal{X}_B\times\mathcal{X}_B}\int_{\mathcal{Y}\times\mathcal{Y}} k_B(\mathbf{x}_B,\mathbf{x}_B')\phi_r(y)\phi_r(y')\left[p_{\mathbf{X}_B Y}(\mathbf{x}_B,y) - p_{\mathbf{X}_B}(\mathbf{x}_B)p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}_B Y}(\mathbf{x}_B',y') - p_{\mathbf{X}_B}(\mathbf{x}_B')p_Y(y')\right]d\mathbf{x}_B d\mathbf{x}_B' dy dy'.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\mathrm{HSIC}(\mathbf{X},Y) &= \sum_{r=1}^{\infty}\|g^{[r]}\|_{\mathcal{F}}^2 \\
&= \sum_{A\subseteq\mathcal{P}_d}\sum_{r=1}^{\infty}\|g_A^{[r]}\|_{\mathcal{F}}^2 \\
&= \sum_{A\subseteq\mathcal{P}_d}\sum_{B\subseteq A}(-1)^{|A|-|B|}\sum_{r=1}^{\infty}\|P_{-B}(g^{[r]})\|_{\mathcal{F}}^2
\end{aligned}$$

and the proof follows from

$$\begin{aligned}
\sum_{r=1}^{\infty} \|P_{-B}(g^{[r]})\|_{\mathcal{F}}^2 &= \sum_{r=1}^{\infty} \int_{\mathcal{X}_B\times\mathcal{X}_B} \int_{\mathcal{Y}\times\mathcal{Y}} k_B(\mathbf{x}_B,\mathbf{x}'_B)\phi_r(y)\phi_r(y') \left[p_{\mathbf{X}_B Y}(\mathbf{x}_B,y) - p_{\mathbf{X}_B}(\mathbf{x}_B)p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}_B Y}(\mathbf{x}'_B,y') - p_{\mathbf{X}_B}(\mathbf{x}'_B)p_Y(y')\right] d\mathbf{x}_B d\mathbf{x}'_B dy dy' \\
&= \int_{\mathcal{X}_B\times\mathcal{X}_B} \int_{\mathcal{Y}\times\mathcal{Y}} k_B(\mathbf{x}_B,\mathbf{x}'_B) \left(\sum_{r=1}^{\infty} \phi_r(y)\phi_r(y')\right) \left[p_{\mathbf{X}_B Y}(\mathbf{x}_B,y) - p_{\mathbf{X}_B}(\mathbf{x}_B)p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}_B Y}(\mathbf{x}'_B,y') - p_{\mathbf{X}_B}(\mathbf{x}'_B)p_Y(y')\right] d\mathbf{x}_B d\mathbf{x}'_B dy dy' \\
&= \int_{\mathcal{X}_B\times\mathcal{X}_B} \int_{\mathcal{Y}\times\mathcal{Y}} k_B(\mathbf{x}_B,\mathbf{x}'_B) k_{\mathcal{Y}}(y,y') \left[p_{\mathbf{X}_B Y}(\mathbf{x}_B,y) - p_{\mathbf{X}_B}(\mathbf{x}_B)p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}_B Y}(\mathbf{x}'_B,y') - p_{\mathbf{X}_B}(\mathbf{x}'_B)p_Y(y')\right] d\mathbf{x}_B d\mathbf{x}'_B dy dy' \\
&= \mathrm{HSIC}(\mathbf{X}_B, Y).
\end{aligned}$$

□

A.4 Proof of Proposition 2

Proof. We begin with the integral formulation of HSIC as in Eq. (22) and plug the kernel defined in Eq. (13):

$$\begin{aligned}
\mathrm{HSIC}(\mathbf{X}_A, Y) &= \int_{\mathcal{X}_A\times\mathcal{X}_A} \int_{\mathcal{Y}\times\mathcal{Y}} k_A(\mathbf{x}_A,\mathbf{x}'_A) k_{\mathcal{Y}}(y,y') \left[p_{\mathbf{X}_A Y}(\mathbf{x}_A,y) - p_{\mathbf{X}_A}(\mathbf{x}_A)p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}_A Y}(\mathbf{x}'_A,y') - p_{\mathbf{X}_A}(\mathbf{x}'_A)p_Y(y')\right] d\mathbf{x}_A d\mathbf{x}'_A dy dy' \\
&= \int_{\mathcal{X}_A\times\mathcal{X}_A} \int_{\mathcal{Y}\times\mathcal{Y}} \frac{1}{\sqrt{p_{\mathbf{X}_A}(\mathbf{x}_A)}\sqrt{p_{\mathbf{X}_A}(\mathbf{x}'_A)}} \prod_{l\in A} \frac{1}{h} K\left(\frac{x_l - x'_l}{h}\right) k_{\mathcal{Y}}(y,y') \\
&\quad \left[p_{\mathbf{X}_A Y}(\mathbf{x}_A,y) - p_{\mathbf{X}_A}(\mathbf{x}_A)p_Y(y)\right] \left[p_{\mathbf{X}_A Y}(\mathbf{x}'_A,y') - p_{\mathbf{X}_A}(\mathbf{x}'_A)p_Y(y')\right] d\mathbf{x}_A d\mathbf{x}'_A dy dy'.
\end{aligned}$$

We then use a change of variables $u_l = (x_l - x'_l)/h$, which leads to

$$\begin{aligned}
\mathrm{HSIC}(\mathbf{X}_A, Y) &= \int_{\mathcal{X}_A\times\mathcal{X}_A} \int_{\mathcal{Y}\times\mathcal{Y}} \frac{1}{\sqrt{p_{\mathbf{X}_A}(\mathbf{x}_A)}\sqrt{p_{\mathbf{X}_A}(\mathbf{x}_A - h\mathbf{u}_A)}} \prod_{l\in A} K(u_l)\, k_{\mathcal{Y}}(y,y') \\
&\quad \left[p_{\mathbf{X}_A Y}(\mathbf{x}_A,y) - p_{\mathbf{X}_A}(\mathbf{x}_A)p_Y(y)\right] \\
&\quad \left[p_{\mathbf{X}_A Y}(\mathbf{x}_A - h\mathbf{u}_A,y') - p_{\mathbf{X}_A}(\mathbf{x}_A - h\mathbf{u}_A)p_Y(y')\right] d\mathbf{x}_A d\mathbf{u}_A dy dy'.
\end{aligned}$$

Now we let $h \to 0$:

$$\begin{aligned}
\lim_{h\to 0} \mathrm{HSIC}(\mathbf{X}_A, Y) &= \int_{\mathcal{X}_A\times\mathcal{X}_A}\int_{\mathcal{Y}\times\mathcal{Y}} \frac{1}{\sqrt{p_{\mathbf{X}_A}(\mathbf{x}_A)}\sqrt{p_{\mathbf{X}_A}(\mathbf{x}_A)}} \prod_{l\in A} K(u_l)\, k_{\mathcal{Y}}(y,y') \\
&\quad [p_{\mathbf{X}_A Y}(\mathbf{x}_A, y) - p_{\mathbf{X}_A}(\mathbf{x}_A) p_Y(y)] \\
&\quad \left[p_{\mathbf{X}_A Y}(\mathbf{x}_A, y') - p_{\mathbf{X}_A}(\mathbf{x}_A) p_Y(y')\right] d\mathbf{x}_A d\mathbf{u}_A dy dy' \\
&= \int_{\mathcal{X}_A} \prod_{l\in A} K(u_l)\, d\mathbf{u}_A \int_{\mathcal{X}_A}\int_{\mathcal{Y}\times\mathcal{Y}} \frac{1}{p_{\mathbf{X}_A}(\mathbf{x}_A)} k_{\mathcal{Y}}(y,y') \\
&\quad [p_{\mathbf{X}_A Y}(\mathbf{x}_A, y) - p_{\mathbf{X}_A}(\mathbf{x}_A) p_Y(y)] \left[p_{\mathbf{X}_A Y}(\mathbf{x}_A, y') - p_{\mathbf{X}_A}(\mathbf{x}_A) p_Y(y')\right] d\mathbf{x}_A dy dy' \\
&= \int_{\mathcal{X}_A}\int_{\mathcal{Y}\times\mathcal{Y}} k_{\mathcal{Y}}(y,y') \left[p_{Y|\mathbf{X}_A=\mathbf{x}_A}(y) - p_Y(y)\right] \\
&\quad \left[p_{Y|\mathbf{X}_A=\mathbf{x}_A}(y') - p_Y(y')\right] p_{\mathbf{X}_A}(\mathbf{x}_A) d\mathbf{x}_A dy dy'
\end{aligned}$$

where we have used $\int_u K(u)du = 1$. Proposition 2 then follows by noting that the last equation equals $\mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right)$ thanks to the integral formulation of the MMD.

□

A.5 Proof of Proposition 3

Proof. Assuming Mercer's theorem holds, we have $k(y,y') = \sum_{r=1}^{\infty} \phi_r(y)\phi_r(y')$ and

$$\begin{aligned}
\mathbb{E}(\chi_n) &= \sum_{r=1}^{\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\phi_r\left(y^{(i)}\right) \phi_r\left(y^{(\sigma_n((i))}\right)\right] && (26) \\
&= \sum_{r=1}^{\infty} \mathbb{E}\left[\phi_r\left(y^{(1)}\right) \phi_r\left(y^{(\sigma_n((1))}\right)\right] \\
&\to \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[\phi_r(Y)|V\right] \mathbb{E}\left[\phi_r(Y)|V\right]\right] && (27) \\
&= \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[\phi_r(Y)|V\right]^2\right] \\
&= \mathbb{E}_V \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|V}} k_{\mathcal{Y}}(\xi,\xi') && (28)
\end{aligned}$$

□

where (26) is obtained by the absolute convergence of Mercer's series, (27) by applying Eq. (34) in the proof of Proposition 3.2 from Gamboa et al. (2020) to $f = g = \phi_r$ (which is bounded since k is bounded) and the absolute convergence of Mercer's series and (28) with Eq. 9. The Mac Diarmid's concentration inequality given in Theorem A.1 in the proof of Proposition 3.2 from Gamboa et al. (2020) is unchanged and concludes the proof.

A.6 Proof of Lemma 3

Proof. We follow closely the proof of Theorem 1 from Song et al. (2016). We only need to prove that for a subset $A \subseteq \mathcal{P}_d$ such that $l \notin A$, then

$$\mathrm{val}(A \cup \{l\}) - \mathrm{val}(A) = \mathrm{val}'(B \cup \{l\}) - \mathrm{val}'(B) \tag{29}$$

where $B = \mathcal{P}_d \backslash (A \cup \{l\})$. We first need the generalized law of total variance for $\mathrm{MMD}^2_{\mathrm{tot}}$ from Proposition 1:

$$\mathrm{MMD}^2_{\mathrm{tot}} = \mathbb{E}_{\mathbf{X}_A}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi')\right] + \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right).$$

Now we prove the equality (29) for val and val′ defined in Lemma 3, except that here we work without the denominator $\mathrm{MMD}^2_{\mathrm{tot}}$ for better readability (this does not change the proof since this same constant appears in both value functions).

$$\begin{aligned}
\mathrm{val}(A \cup \{l\}) - \mathrm{val}(A) &= \mathbb{E}_{\mathbf{X}_{A\cup\{l\}}}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_{A\cup\{l\}}})\right) - \mathbb{E}_{\mathbf{X}_A}\left(\mathrm{MMD}^2(\mathrm{P}_Y, \mathrm{P}_{Y|\mathbf{X}_A})\right) \\
&= \left\{\mathrm{MMD}^2_{\mathrm{tot}} - \mathbb{E}_{\mathbf{X}_{A\cup\{l\}}}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_{A\cup\{l\}}}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_{A\cup\{l\}}}} k_{\mathcal{Y}}(\xi,\xi')\right]\right\} \\
&\quad - \left\{\mathrm{MMD}^2_{\mathrm{tot}} - \mathbb{E}_{\mathbf{X}_A}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi')\right]\right\} \\
&= \mathbb{E}_{\mathbf{X}_A}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_A}} k_{\mathcal{Y}}(\xi,\xi')\right] \\
&\quad - \mathbb{E}_{\mathbf{X}_{A\cup\{l\}}}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_{A\cup\{l\}}}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_{A\cup\{l\}}}} k_{\mathcal{Y}}(\xi,\xi')\right] \\
&= \mathbb{E}_{\mathbf{X}_{-(B\cup\{l\})}}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_{-(B\cup\{l\})}}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_{-(B\cup\{l\})}}} k_{\mathcal{Y}}(\xi,\xi')\right] \\
&\quad - \mathbb{E}_{\mathbf{X}_{-B}}\left[\mathbb{E}_{\xi\sim \mathrm{P}_{Y|\mathbf{X}_{-B}}} k_{\mathcal{Y}}(\xi,\xi) - \mathbb{E}_{\xi,\xi'\sim \mathrm{P}_{Y|\mathbf{X}_{-B}}} k_{\mathcal{Y}}(\xi,\xi')\right] \\
&= \mathrm{val}'(B \cup \{l\}) - \mathrm{val}'(B)
\end{aligned}$$

where we have used the generalized law of total variance for the second equality. The rest of the proof is identical to the one of Theorem 1 from Song et al. (2016).

□

References

Antoniadis, A. (1984), 'Analysis of variance on function spaces', *Math. Operationsforsch. u. Statist., ser. statist.* **15**, 59–71.

Aubin, J. P. (2000), *Applied Functional Analysis*, 2^{nd} edn, New York: Wiley-Interscience.

Azadkia, M. and Chatterjee, S. (2019), 'A simple measure of conditional dependence', *arXiv preprint arXiv:1910.12327* .

Bachoc, F., Gamboa, F., Loubes, J.-M. and Venet, N. (2017), 'A gaussian process regression model for distribution inputs', *IEEE Transactions on Information Theory* **64**(10), 6620–6637.

Baucells, M. and Borgonovo, E. (2013), 'Invariant probabilistic sensitivity analysis', *to appear in Management Science* .

Borgonovo, E. (2007), 'A new uncertainty importance measure', *Reliability Engineering & System Safety* **92**(6), 771–784.

Broto, B., Bachoc, F. and Depecker, M. (2020), 'Variance reduction for estimation of shapley effects and adaptation to unknown input distribution', *SIAM/ASA Journal on Uncertainty Quantification* **8**(2), 693–716.

Castro, J., Gómez, D. and Tejada, J. (2009), 'Polynomial calculation of the shapley value based on sampling', *Computers & Operations Research* **36**(5), 1726–1730.

Charpentier, A., Elie, R., Laurière, M. and Tran, V. C. (2020), 'Covid-19 pandemic control: balancing detection policy and lockdown intervention under icu sustainability', *arXiv preprint arXiv:2005.06526* .

Chastaing, G., Gamboa, F., Prieur, C. et al. (2012), 'Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis', *Electronic Journal of Statistics* **6**, 2420–2448.

Chatterjee, S. (2020), 'A new coefficient of correlation', *Journal of the American Statistical Association* pp. 1–21.

Chwialkowski, K., Strathmann, H. and Gretton, A. (2016), A kernel test of goodness of fit, *in* 'ICML'.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009), 'Modeling wine preferences by data mining from physicochemical properties', *Decision Support Systems* **47**(4), 547–553.

Cuturi, M. (2011), Fast global alignment kernels, *in* 'Proceedings of the 28th international conference on machine learning (ICML-11)', pp. 929–936.

Czado, C. (2019), 'Analyzing dependent data with vine copulas', *Lecture Notes in Statistics, Springer* .

Da Veiga, S. (2015), 'Global sensitivity analysis with dependence measures', *Journal of Statistical Computation and Simulation* **85**(7), 1283–1305.

Da Veiga, S. (2016), New perspectives for sensitivity analysis, *in* 'Proceedings of Mascot-Num 2016 conference', Toulouse, France. `https://mascot2016.sciencesconf.org/resource/page/id/2`.

Da Veiga, S. and Gamboa, F. (2013), 'Efficient estimation of sensitivity indices', *Journal of Nonparametric Statistics* **25**(3), 573–595.

Da Veiga, S., Wahl, F. and Gamboa, F. (2009), 'Local polynomial estimation for sensitivity analysis on models with correlated inputs', *Technometrics* **51**(4), 452–463.

Di Domenico, L., Pullano, G., Sabbatini, C. E., Boëlle, P.-Y. and Colizza, V. (2020), 'Expected impact of lockdown in île-de-france and possible exit strategies', *medRxiv* .

Ditlevsen, O. and Madsen, H., eds (1996), *Structural reliability methods*, Wiley & Sons.

Dua, D. and Graff, C. (2017), 'Uci machine learning repository'.

Durrande, N., Ginsbourger, D., Roustant, O. and Carraro, L. (2012), 'Anova kernels and rkhs of zero mean functions for model-based sensitivity analysis', *Journal of Multivariate Analysis* **115**, 57–67.

Ferraty, F. and Vieu, P. (2006), *Nonparametric functional data analysis: theory and practice*, Springer.

Fort, J.-C., Klein, T. and Rachdi, N. (2016), 'New sensitivity analysis subordinated to a contrast', *Communications in Statistics-Theory and Methods* **45**(15), 4349–4364.

Gamboa, F., Gremaud, P., Klein, T. and Lagnoux, A. (2020), 'Global sensitivity analysis: a new generation of mighty estimators based on rank statistics', *hal-02474902v3* .

Gamboa, F., Janon, A., Klein, T. and Lagnoux, A. (2013), 'Sensitivity indices for multivariate outputs', *Comptes Rendus Mathematique* **351**(7-8), 307–310.

Gärtner, T., Flach, P. and Wrobel, S. (2003), On graph kernels: Hardness results and efficient alternatives, *in* 'Learning theory and kernel machines', Springer, pp. 129–143.

Gorham, J. and Mackey, L. (2015), 'Measuring sample quality with stein's method', *Advances in Neural Information Processing Systems* **28**, 226–234.

Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. (2005a), Measuring statistical dependence with hilbert-schmidt norms, *in* S. Jain, H. Simon and E. Tomita, eds, 'Algorithmic Learning Theory', Vol. 3734 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 63–77.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B. and Smola, A. J. (2008), A kernel statistical test of independence, *in* 'Advances in neural information processing systems', pp. 585–592.

Gretton, A., Herbrich, R., Smola, A., Bousquet, O. and Schölkopf, B. (2005b), 'Kernel methods for measuring independence', *The Journal of Machine Learning Research* **6**, 2075–2129.

Harchaoui, Z. and Bach, F. (2007), Image classification with segmentation graph kernels, *in* '2007 IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 1–8.

Hoeffding, W. (1948), 'A class of statistics with asymptotically normal distributions', *Annals of Mathematical Statistics* **19**, 293–325.

Homma, T. and Saltelli, A. (1996), 'Importance measures in global sensitivity analysis of nonlinear models', *Reliability Engineering & System Safety* **52**(1), 1–17.

Iooss, B. and Prieur, C. (2019), 'Shapley effects for sensitivity analysis with dependent inputs: comparisons with Sobol' indices, numerical estimation and applications', *International Journal for Uncertainty Quantification* **9**, 493–514,.

Ishigami, T. and Homma, T. (1990), An importance quantification technique in uncertainty analysis for computer models, *in* '[1990] Proceedings. First International Symposium on Uncertainty Modeling and Analysis', IEEE, pp. 398–403.

Janon, A., Klein, T., Lagnoux, A., Nodet, M. and Prieur, C. (2014), 'Asymptotic normality and efficiency of two sobol index estimators', *ESAIM: Probability and Statistics* **18**, 342–364.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K. and Gretton, A. (2017), A linear-time kernel goodness-of-fit test, *in* 'Advances in Neural Information Processing Systems', pp. 262–271.

Kermack, W. O. and McKendrick, A. G. (1927), 'A contribution to the mathematical theory of epidemics', *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character* **115**(772), 700–721.

Kuo, F., Sloan, I., Wasilkowski, G. and Woźniakowski, H. (2010), 'On decompositions of multivariate functions', *Mathematics of computation* **79**(270), 953–966.

Lamboni, M., Monod, H. and Makowski, D. (2011), 'Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models', *Reliability Engineering & System Safety* **96**, 450–459.

Li, L., Lu, Z., Feng, J. and Wang, B. (2012), 'Moment-independent importance measure of basic variable and its state dependent parameter solution', *Structural Safety* **38**, 40–47.

Lundberg, S. M. and Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Advances in neural information processing systems', pp. 4765–4774.

Magal, P. and Webb, G. (2020), 'Predicting the number of reported and unreported cases for the covid-19 epidemic in south korea, italy, france and germany', *medRxiv* .
URL: *https://www.medrxiv.org/content/early/2020/03/24/2020.03.21.20040154*

Mara, T. A., Tarantola, S. and Annoni, P. (2015), 'Non-parametric methods for global sensitivity analysis of model output with dependent inputs', *Environmental modelling & software* **72**, 173–183.

Marrel, A. and Chabridon, V. (2020), 'Statistical developments for target and conditional sensitivity analysis: application on safety studies for nuclear reactor', *hal-02541142* .

Marrel, A., Iooss, B., Van Dorpe, F. and Volkova, E. (2008), 'An efficient methodology for modeling complex computer codes with Gaussian processes', *Computational Statistics and Data Analysis* **52**, 4731–4744.

Maume-Deschamps, V. and Niang, I. (2018), 'Estimation of quantile oriented sensitivity indices', *Statistics & Probability Letters* **134**, 122–127.

Moutoussamy, V., Nanty, S. and Pauwels, B. (2015), 'Emulators for stochastic simulation codes', *ESAIM: Proceedings and Surveys* **48**, 116–155.

Muandet, K., Fukumizu, K., Dinuzzo, F. and Schölkopf, B. (2012), 'Learning from distributions via support measure machines', *Advances in neural information processing systems* **25**, 10–18.

Oates, C. J., Girolami, M. and Chopin, N. (2017), 'Control functionals for monte carlo integration', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(3), 695–718.

Owen, A. (2014), 'Sobol' indices and Shapley value', *SIAM/ASA Journal on Uncertainty Quantification* **2**, 245–251.

Perrin, G. and Defaux, G. (2019), 'Efficient estimation of reliability-oriented sensitivity indices', *Journal of Scientific Computing* **80**(3).

Plischke, E., Rabitti, G. and Borgonovo, E. (2020), 'Computing shapley effects for sensitivity analysis', *arXiv preprint arXiv:2002.12024* .

Rahman, S. (2016), 'The f-sensitivity index', *SIAM/ASA Journal on Uncertainty Quantification* **4**(1), 130–162.

Ramon, J. and Gärtner, T. (2003), Expressivity versus efficiency of graph kernels, *in* 'Proceedings of the first international workshop on mining graphs, trees and sequences', pp. 65–74.

Sakoe, H. and Chiba, S. (1978), 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE transactions on acoustics, speech, and signal processing* **26**(1), 43–49.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M. and Tarantola, S. (2010), 'Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index', *Computer physics communications* **181**(2), 259–270.

Saltelli, A., Tarantola, S. and Chan, K.-S. (1999), 'A quantitative model-independent method for global sensitivity analysis of model output', *Technometrics* **41**(1), 39–56.

Shapley, L. (1953), A value for n-persons game, *in* H. Kuhn and A. Tucker, eds, 'Contributions to the theory of games II, Annals of mathematic studies', Princeton University Press, Princeton, NJ.

Smola, A., Gretton, A., Song, L. and Schölkopf, B. (2007), A hilbert space embedding for distributions, *in* 'Algorithmic Learning Theory', Vol. 4754, Springer, pp. 13–31.

Sobol', I. (1993), 'Sensitivity estimates for non linear mathematical models', *Mathematical Modelling and Computational Experiments* **1**, 407–414.

Solís, M. (2019), 'Non-parametric estimation of the first-order sobol indices with bootstrap bandwidth', *Communications in Statistics-Simulation and Computation* pp. 1–16.

Song, E., Nelson, B. L. and Staum, J. (2016), 'Shapley effects for global sensitivity analysis: Theory and computation', *SIAM/ASA Journal on Uncertainty Quantification* **4**(1), 1060–1083.

Song, L. (2008), Learning via Hilbert Space Embedding of Distributions, PhD thesis, University of Sydney.

Song, L., Smola, A., Gretton, A., Bedo, J. and Borgwardt, K. (2012), 'Feature selection via dependence maximization', *The Journal of Machine Learning Research* **13**, 1393–1434.

Song, L., Smola, A., Gretton, A., Borgwardt, K. M. and Bedo, J. (2007), Supervised feature selection via dependence estimation, *in* 'Proceedings of the 24th international conference on Machine learning', pp. 823–830.

Spagnol, A. (2020), Kernel-based sensitivity indices for high-dimensional optimization problems, PhD thesis, University Lyon, France.

Spagnol, A., Le Riche, R. and Da Veiga, S. (2019), 'Global sensitivity analysis for optimization with variable selection', *SIAM/ASA Journal on Uncertainty Quantification* **7**(2), 417–443.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Lanckriet, G. R. and Schölkopf, B. (2009), Kernel choice and classifiability for rkhs embeddings of probability distributions, *in* 'Advances in neural information processing systems', pp. 1750–1758.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B. and Lanckriet, G. R. (2010), 'Hilbert space embeddings and metrics on probability measures', *The Journal of Machine Learning Research* **11**, 1517–1561.

Stein, C. et al. (1972), A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, *in* 'Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory', The Regents of the University of California.

Steinwart, I., Hush, D. and Scovel, C. (2006), 'An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels', *IEEE Transactions on Information Theory* **52**(10), 4635–4643.

Szabó, Z., Sriperumbudur, B. K., Póczos, B. and Gretton, A. (2016), 'Learning theory for distribution regression', *The Journal of Machine Learning Research* **17**(1), 5272–5311.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), 'Measuring and testing dependence by correlation of distances', *The Annals of Statistics* **35**(6), 2769–2794.

Terraz, T., Ribes, A., Fournier, Y., Iooss, B. and Raffin, B. (2017), Large scale in transit global sensitivity analysis avoiding intermediate files, *in* 'Proceedings the International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing)', Denver, USA.

Wahba, G., Wang, Y., Gu, C., Klein, R., Klein, B. et al. (1995), 'Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy: the 1994 neyman memorial lecture', *The Annals of Statistics* **23**(6), 1865–1895.