Global Sensitivity Analysis with Dependence Measures

Sébastien da Veiga

Global Sensitivity Analysis with Dependence Measures

Sebastien Da Veiga[a*]
[a]*IFP Energies nouvelles, 1 & 4 avenue de Bois Préau, 92852 Rueil-Malmaison, France*

Abstract

Global sensitivity analysis with variance-based measures suffers from several theoretical and practical limitations, since they focus only on the variance of the output and handle multivariate variables in a limited way. In this paper, we introduce a new class of sensitivity indices based on dependence measures which overcomes these insufficiencies. Our approach originates from the idea to compare the output distribution with its conditional counterpart when one of the input variables is fixed. We establish that this comparison yields previously proposed indices when it is performed with Csiszár f-divergences, as well as sensitivity indices which are well-known dependence measures between random variables. This leads us to investigate completely new sensitivity indices based on recent state-of-the-art dependence measures, such as distance correlation and the Hilbert-Schmidt independence criterion. We also emphasize the potential of feature selection techniques relying on such dependence measures as alternatives to screening in high dimension.

1 Introduction

Since the early work of Sobol (Sobol', 1993), global sensitivity analysis (GSA) has received a lot of attention in the computer code experiments community. These days variance-based sensitivity indices are common tools in the analysis of complex physical phenomena. Several statistical estimators have been proposed (Cukier et al., 1973; Saltelli et al., 2010; Tarantola et al., 2006; Janon et al., 2013; Owen, 2013) and their asymptotic properties are now well understood (Tissot and Prieur, 2012; Da Veiga and Gamboa, 2013; Janon et al., 2013). In addition, the case of computationally expensive codes has been investigated thoroughly with the introduction of several dedicated surrogate models (Oakley and O'Hagan, 2004; Marrel et al., 2009; Da Veiga et al., 2009; Blatman and Sudret, 2010; Touzani and Busby, 2012; Durrande et al., 2012).

However, even if they are extremely popular and informative importance measures, variance-based indices suffer from theoretical and practical limitations. First, by definition they only study the impact of the input parameters on the variance of the output. Since this is a restricted summary of the output distribution, this measure happens to be inadequate for many case studies. Alternative approaches include for example density-based indices (Borgonovo, 2007), derivative-based measures (Sobol and Kucherenko, 2009), or goal-oriented dedicated indices (Fort et al., 2013). Second, variance-based indices do not generalize easily to the case of a multivariate output (Gamboa et al., 2013). Unfortunately, computer code outputs often consist of several scalars or even time-dependent curves, which limits severely the practical use of standard indices. Finally for high-dimensional problems, a preliminary screening procedure is usually mandatory before the analysis of the computer code or the modeling with a surrogate. The computational cost of GSA is in general too high to envision its use for screening purposes and more qualitative approaches are thus needed, e.g. the Morris method (Morris, 1991) or group screening (Moon et al., 2012).

In this paper, we propose a completely original point of view in order to overcome these limitations. Starting from the general framework of GSA and the concept of dissimilarity measure, we introduce a new

*Now at Snecma (Safran), Établisssement de Montereau, Aérodrome Melun-Villaroche, BP 1936, 77019 Melun cedex, France. Email: sebastien.daveiga@snecma.fr

class of sensitivity indices which comprises as a special case the density-based index of Borgonovo (2007). We propose an estimation procedure relying on density ratio estimation and show that it gives access to several different indices for the same computational cost. More importantly, we highlight that other special cases lead to well-known dependence measures, including the mutual information. This link motivates us to investigate the potential of recent state-of-the-art dependence measures as new sensitivity indices, such as the distance correlation (Székely et al., 2007) or the Hilbert-Schmidt independence criterion (Gretton et al., 2005a). An appealing property of such measures is that they can handle multivariate random variables very easily. We also discuss how feature selection methods based on these measures can effectively replace standard screening procedures.

The structure of the paper is as follows. In Section 2, we introduce the general GSA framework based on dissimilarity measures and discuss the use of Csiszár f-divergences. We also emphasize the link with mutual information and propose an estimation procedure. In Section 3, we give a review of some dependence measures and show how they can be used as new sensitivity indices. We also provide examples of feature selection techniques in which they are involved. Screening will then be seen as an equivalent to feature selection in machine learning. Finally, several numerical experiments are conducted in Section 4 on both analytical and industrial applications. In particular, we illustrate the potential of dependence measures for GSA.

2 From dissimilarity measures to sensitivity indices

Denote $Y = \eta(X^1, \dots, X^p)$ the computer code output which is a function of the p input random variables X^k, $k = 1, \dots, p$ where $\eta : \mathbb{R}^p \to \mathbb{R}$ is assumed to be continuous. In standard global sensitivity analysis, it is further assumed that the X^k have a known distribution and are independent. As pointed out by Baucells and Borgonovo (2013), a natural way of defining the impact of a given input X^k on Y is to consider a function which measures the similarity between the distribution of Y and that of $Y|X^k$. More precisely, the impact of X^k on Y is given by

$$S_{X^k} = \mathbb{E}_{X^k}\left(d(Y, Y|X^k)\right) \tag{1}$$

where $d(\cdot, \cdot)$ denotes a dissimilarity measure between two random variables. The advantage of such a formulation is that many choices for d are available, and we will see in what follows that some natural dissimilarity measures yield sensitivity indices related to well known quantities. However before going further, let us note that the naive dissimilarity measure

$$d(Y, Y|X^k) = \left(\mathbb{E}(Y) - \mathbb{E}(Y|X^k)\right)^2 \tag{2}$$

where random variables are compared only through their mean values produces the unnormalized Sobol first-order sensivity index $S^1_{X^k} = \mathrm{Var}\left(\mathbb{E}(Y|X^k)\right)$.

2.1 Csiszár f-divergences

Assuming all input random variables have an absolutely continuous distribution with respect to the Lebesgue measure on $\mathbb{R}$, the f-divergence (Csiszár, 1967) between Y and $Y|X^k$ is given by

$$d_f(Y||Y|X^k) = \int_{\mathbb{R}} f\left(\frac{p_Y(y)}{p_{Y|X^k}(y)}\right) p_{Y|X^k}(y)dy$$

where f is a convex function such that $f(1) = 0$ and p_Y and $p_{Y|X^k}$ are the probability distribution functions of Y and $Y|X^k$, respectively. Standard choices for function f include for example

- Kullback-Leibler divergence: $f(t) = -\ln(t)$ or $f(t) = t\ln(t)$;
- Hellinger distance: $f(t) = \left(\sqrt{t} - 1\right)^2$;

- Total variation distance: $f(t) = |t-1|$;
- Pearson χ^2 divergence: $f(t) = (t-1)^2$ or $f(t) = t^2 - 1$;
- Neyman χ^2 divergence: $f(t) = (t-1)^2/t$ or $f(t) = (1-t^2)/t$.

Plugging this dissimilarity measure in (1) yields the following sensitivity index:

$$S^f_{X^k} = \int_{\mathbb{R}^2} f\left(\frac{p_Y(y)p_{X^k}(x)}{p_{X^k,Y}(x,y)}\right) p_{X^k,Y}(x,y)dxdy \tag{3}$$

where p_{X^k} and $p_{X^k,Y}$ are the probability distribution functions of X^k and (X^k, Y), respectively. First of all, note that inequalities on Csiszár f-divergences imply that such sensitivity indices are positive and equal zero when Y and X_k are independent. Also, it is important to note that given the form of $S^f_{X^k}$, it is invariant under any smooth and uniquely invertible transformation of the variables X^k and Y, see the proof for mutual information in Kraskov et al. (2004). This is a major advantage over variance-based Sobol sensitivity indices, which are only invariant under linear transformations.

It is easy to see that the total variation distance with $f(t) = |t-1|$ gives a sensivity index equal to the one proposed by Borgonovo (2007):

$$S^f_{X^k} = \int_{\mathbb{R}^2} |p_Y(y)p_{X^k}(x) - p_{X^k,Y}(x,y)|dxdy.$$

In addition, the Kullback-Leibler divergence with $f(t) = -\ln(t)$ yields

$$S^f_{X^k} = \int_{\mathbb{R}^2} p_{X^k,Y}(x,y) \ln\left(\frac{p_{X^k,Y}(x,y)}{p_Y(y)p_{X^k}(x)}\right) dxdy,$$

that is the mutual information $I(X^k; Y)$ between X^k and Y. A normalized version of this sensitivity index was studied by Krzykacz-Hausmann (2001). Similarly, the Neyman χ^2 divergence with $f(t) = (1-t^2)/t$ leads to

$$S^f_{X^k} = \int_{\mathbb{R}^2} \left(\frac{p_{X^k,Y}(x,y)}{p_Y(y)p_{X^k}(x)} - 1\right)^2 p_Y(y)p_{X^k}(x)dxdy,$$

which is the so-called squared-loss mutual information between X^k and Y (or mean square contingency). These results show that some previously proposed sensitivity indices are actually special cases of more general indices defined through Csiszár f-divergences. To the best of our knowledge, this is the first work in which this link is highlighted. Moreover, the specific structure of equation (3) makes it possible to envision more efficient tools for the estimation of these sensitivity indices. Indeed, it only involves approximating a density ratio rather than full densities. This point is investigated in the next subsection. But more importantly, we see that special choices for f define sensivity indices that are actually well-known dependence measures such as the mutual information. This paves the way for completely new sensitivity indices based on recent state-of-the-art dependence measures, see Section 3.

2.2 Estimation

Coming back to equation (3), the goal is to estimate

$$S^f_{X^k} = \int_{\mathbb{R}^2} f\left(\frac{1}{r(x,y)}\right) p_{X^k,Y}(x,y)dxdy = \mathbb{E}_{(X^k,Y)} f\left(\frac{1}{r(X^k,Y)}\right)$$

where $r(x,y) = p_{X^k,Y}(x,y)/(p_Y(y)p_{X^k}(x))$ is the ratio between the joint density of (X^k, Y) and the marginals. Of course, straightforward estimation is possible if one estimates the densities $p_{X^k,Y}(x,y)$, $p_{X^k}(x)$

and $p_Y(y)$ with e.g. kernel density estimators. However, it is well known that density estimation suffers from the curse of dimensionality. This limits the possible multivariate extensions we discuss in the next subsection. Besides, since only the ratio function $r(x, y)$ is needed, we expect more robust estimates by focusing only on it.

Let us assume now that we have a sample $(X_i^k, Y_i)_{i=1,\dots,n}$ of (X^k, Y), the idea is to build first an estimate $\hat{r}(x, y)$ of the ratio. The final estimator $\hat{S}^f_{X^k}$ of $S^f_{X^k}$ will then be given by

$$\hat{S}^f_{X^k} = \frac{1}{n}\sum_{i=1}^{n} f\left(\frac{1}{\hat{r}(X_i^k, Y_i)}\right). \tag{4}$$

Powerful estimating methods for ratios include e.g. maximum-likelihood estimation (Suzuki et al., 2008), unconstrained least-squares importance fitting (Kanamori et al., 2009), among others (see Sugiyama et al., 2012). A k-nearest neighbors strategy dedicated to mutual information is also discussed in Kraskov et al. (2004).

2.3 Multivariate extensions

Given our approach focusing only on densities, it is straightforward to extend the definition of the sensitivity index in equation (3) to any number of input and output variables. We can then study the impact of a given group of input variables $X^u = \left\{X^k\right\}_{k\in u}$ where u is a subset of $\{1, \dots, p\}$ on a multivariate output $Y \in \mathbb{R}^q$ with the sensitivity index given by

$$S^f_{X^u} = \int_{\mathbb{R}^{|u|}\times\mathbb{R}^q} f\left(\frac{p_Y(y)p_{X^u}(x)}{p_{X^u,Y}(x, y)}\right) p_{X^u,Y}(x, y)dxdy.$$

This favorable generalization was already mentioned for the special cases of the total-variation distance and mutual information by Borgonovo (2007) and Auder and Iooss (2008), respectively. However, in the high-dimensional setting, estimation of such sensitivity indices is infeasible since even the ratio trick detailed above fails. This is thus a severe limitation for screening purposes. We examine efficient alternatives in Section 3.

Moreover, note that extending the naive dissimilarity measure (2) to the multivariate output case naturally leads to consider $d(Y, Y|X^k) = \|\mathbb{E}(Y) - \mathbb{E}(Y|X^k)\|_2^2$. Straightforward calculations reveal that the corresponding sensitivity index is then the sum of Sobol first-order sensitivity indices on each output. Gamboa et al. (2013) showed that this multivariate index is the only one possessing desired invariance properties in the variance-based index family.

2.4 On the use of other dissimilarity measures

We focused above on Csiszár f-divergences but other dissimilarity measures exist to compare probability distributions. In particular, integral probability metrics (IPM, Müller, 1997) are a popular family of distance measures on probabilities given by

$$\gamma_{\mathcal{F}}(\mathbb{P}, \mathbb{Q}) = \sup_{f\in\mathcal{F}} \left| \int_S f d\mathbb{P} - \int_S f d\mathbb{Q} \right| \tag{5}$$

for two probability measures $\mathbb{P}$ and $\mathbb{Q}$ and where $\mathcal{F}$ is a class of real-valued bounded measurable functions on S. Just as the choice of function f in Csiszár f-divergences gives rise to different measures, the choice of $\mathcal{F}$ generates different IPMs, e.g. the Wasserstein distance, the Dudley metric or the total variation distance. It is interesting to note that Csiszár f-divergences and IPMs are very distinct classes of measures, since they only intersect at the total variation distance (Sriperumbudur et al., 2012). Unfortunately, plugging the general expression (5) of an IPM in equation (1) no longer yields a closed-form expression for a sensitivity index. However, we plan to study such indices in a future work since estimation of IPMs appears to be easier than for

Csiszár f-divergences and is independent of the dimensionality of the random variables (Sriperumbudur et al., 2012).

Finally, let us mention the recent work of Fort et al. (2013) on goal-oriented measures, where they introduce a new class of sensitivity indices

$$S^{\psi}_{X^k} = \mathbb{E}\psi(Y;\theta^*) - \mathbb{E}_{(X^k,Y)}\psi(Y;\theta_k(X^k)) \tag{6}$$

where ψ is the contrast function associated to the features of interest $\theta^* = \arg\min_\theta \mathbb{E}\psi(Y;\theta)$ and $\theta_k(x) = \arg\min_\theta \mathbb{E}(\psi(Y;\theta)|X^k = x)$ of Y and Y conditionally to $X^k = x$, respectively (note that we only give here the unnormalized version of the index). It is easy to check that (6) is a special case of (1).

3 Dependence measures and feature selection

Given two random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$, dependence measures aim at quantifying the dependence between X and Y in arbitrary dimension, with the property that the measure equals zero if and only if X and Y are independent. In particular, they are useful when one wants to design a statistical test for independence. Here, we focus on the long-known mutual information criterion, as well as on the novel distance correlation measure (Székely et al., 2007). Recently, Sejdinovic et al. (2013) showed that it shares deep links with distances between embeddings of distributions to reproducing kernel Hilbert spaces (RKHS) and especially the Hilbert-Schmidt independence criterion (HSIC, Gretton et al., 2005a) which will also be discussed. Finally, we will review feature selection techniques introduced in machine learning which make use of these dependence measures.

3.1 Mutual information

Mutual information (MI, Shannon, 1948) is a symmetric measure of dependence which is related to the entropy. Assuming X and Y are two random vectors which are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^p$ and $\mathbb{R}^q$ with density functions $p_X(x)$ and $p_Y(y)$, respectively, one can define their marginal entropy:

$$H(X) = -\int_{\mathbb{R}^p} p_X(x)\ln(p_X(x))dx$$

and $H(Y)$ similarly. Denoting $p_{X,Y}(x,y)$ their joint density function, the joint entropy between X and Y writes

$$H(X,Y) = -\int_{\mathbb{R}^{p+q}} p_{X,Y}(x,y)\ln(p_{X,Y}(x,y))dxdy.$$

MI is then formally defined as

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= \int_{\mathbb{R}^{p+q}} p_{X,Y}(x,y)\ln\left(\frac{p_{X,Y}(x,y)}{p_Y(y)p_X(x)}\right)dxdy. \end{aligned}$$

Interestingly, MI equals zero if and only if X and Y are independent. This implies that MI is able to detect nonlinear dependencies between random variables, unlike the correlation coefficient. It is also easy to check that $I(X;Y) \geq 0$ with Jensen's inequality. Further note that it is not a distance since it does not obey the triangle inequality. A simple modified version yielding a distance, the variation of information (VI), is given by

$$VI(X;Y) = H(X,Y) - I(X;Y) = H(X) + H(Y) - 2I(X;Y).$$

Another variant is the squared-loss mutual information (SMI, Suzuki et al., 2009):

$$SMI(X;Y) = \int_{\mathbb{R}^{p+q}} \left(\frac{p_{X,Y}(x,y)}{p_Y(y)p_X(x)} - 1\right)^2 p_Y(y)p_X(x)dxdy \tag{7}$$

which is again a dependence measure verifying $SMI(X;Y) \geq 0$ with equality if and only if X and Y are independent. Applications of MI, VI and SMI include independent component analysis (Hyvärinen and Oja, 2000), image registration (Pluim et al., 2003) and hierarchical clustering (Meilă, 2007), among many others.

In the context of global sensitivity analysis, we have seen in Section 2.1 that MI and SMI arise as sensitivity indices when specific Csiszár f-divergences are chosen to evaluate the dissimilarity between the output Y and the conditional output $Y|X^k$ where X^k is one of the input variables. We will then study the two following sensitivity indices:

$$S_{X^k}^{MI} = I(X^k;Y) = \int_{\mathbb{R}^2} p_{X^k,Y}(x,y) \ln\left(\frac{p_{X^k,Y}(x,y)}{p_Y(y)p_{X^k}(x)}\right) dxdy \tag{8}$$

and

$$S_{X^k}^{SMI} = SMI(X^k;Y) = \int_{\mathbb{R}^2} \left(\frac{p_{X^k,Y}(x,y)}{p_Y(y)p_{X^k}(x)} - 1\right)^2 p_Y(y)p_{X^k}(x)dxdy. \tag{9}$$

A normalized version of $S_{X^k}^{MI}$ given by $I(X^k;Y)/H(Y)$ has already been proposed by Krzykacz-Hausmann (2001) and compared to Sobol sensitivity indices by Auder and Iooss (2008).

3.2 Distance correlation

The distance correlation was introduced by Székely et al. (2007) to address the problem of testing dependence between two random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$. It is based on the concept of distance covariance which measures the distance between the joint characteristic function of (X,Y) and the product of the marginal characteristic functions.

More precisely, denote ϕ_X and ϕ_Y the characteristic function of X and Y, respectively, and $\phi_{X,Y}$ their joint characteristic function. For a complex-valued function $\phi(\cdot)$, we also denote $\bar{\phi}$ the complex conjugate of ϕ and $|\phi|^2 = \phi\bar{\phi}$. The distance covariance (dCov) $\mathcal{V}(X,Y)$ between X and Y with finite first moment is then defined as a weighted L_2-distance between $\phi_{X,Y}$ and $\phi_X\phi_Y$ given by

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \|\phi_{X,Y} - \phi_X\phi_Y\|_w^2 \\ &= \int_{\mathbb{R}^{p+q}} |\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2 w(t,s)dtds\end{aligned} \tag{10}$$

where the weight function $w(t,s) = (c_p c_q \|t\|_2^{1+p}\|s\|_2^{1+q})^{-1}$ with constants $c_l = \pi^{(1+l)/2}/\Gamma((1+l)/2)$ for $l \in \mathbb{N}$ is chosen to ensure invariance properties, see Székely et al. (2007). The distance correlation (dCor) $\mathcal{R}(X,Y)$ between X and Y is then naturally defined as

$$\mathcal{R}^2(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y)}} \tag{11}$$

if $\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y) > 0$ and equals 0 otherwise. Important properties of the distance correlation introduced in (11) include that $0 \leq \mathcal{R}(X,Y) \leq 1$ and $\mathcal{R}(X,Y) = 0$ if and only if X and Y are independent. Interestingly, the distance covariance in (10) can be computed in terms of expectations of pairwise Euclidean distances, namely

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}_{X,X',Y,Y'}\|X-X'\|_2\|Y-Y'\|_2 \\ &\quad + \mathbb{E}_{X,X'}\|X-X'\|_2\mathbb{E}_{Y,Y'}\|Y-Y'\|_2 \\ &\quad - 2\mathbb{E}_{X,Y}\left[\mathbb{E}_{X'}\|X-X'\|_2\mathbb{E}_{Y'}\|Y-Y'\|_2\right]\end{aligned} \tag{12}$$

where (X',Y') is an i.i.d. copy of (X,Y). Concerning estimation, let $(X_i,Y_i)_{i=1,\dots,n}$ be a sample of the

random vector (X, Y). Following equation (12), an estimator $\mathcal{V}_n^2(X, Y)$ of $\mathcal{V}^2(X, Y)$ is then given by

$$\begin{aligned}\mathcal{V}_n^2(X,Y) &= \frac{1}{n^2}\sum_{i,j=1}^{n}\|X_i - X_j\|_2\|Y_i - Y_j\|_2 \\ &+ \frac{1}{n^2}\sum_{i,j=1}^{n}\|X_i - X_j\|_2\frac{1}{n^2}\sum_{i,j=1}^{n}\|Y_i - Y_j\|_2 \\ &- \frac{2}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}\|X_i - X_j\|_2\frac{1}{n}\sum_{j=1}^{n}\|Y_i - Y_j\|_2\right]. \end{aligned} \tag{13}$$

Denoting $a_{ij} = \|X_i - X_j\|_2$, $\bar{a}_{i\cdot} = \sum_j a_{ij}/n$, $\bar{a}_{\cdot j} = \sum_i a_{ij}/n$, $\bar{a}_{\cdot\cdot} = \sum_{ij} a_{ij}/n^2$, $A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}$ and similarly $B_{ij} = b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}$ for $b_{ij} = \|Y_i - Y_j\|_2$, Székely et al. (2007) show that equation (13) can be written as

$$\mathcal{V}_n^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij}$$

and is also equal to equation (10) if one uses the empirical characteristic functions computed on the sample $(X_i, Y_i)_{i=1,\dots,n}$. The empirical distance correlation $\mathcal{R}_n(X, Y)$ is then

$$\mathcal{R}_n^2(X,Y) = \frac{\mathcal{V}_n^2(X,Y)}{\sqrt{\mathcal{V}_n^2(X,X)\mathcal{V}_n^2(Y,Y)}}$$

and satisfies $0 \leq \mathcal{R}_n(X, Y) \leq 1$. Although $\mathcal{V}_n^2(X, Y)$ is a consistent estimator of $\mathcal{V}^2(X, Y)$, it is easy to see that it is biased. Székely and Rizzo (2013*b*) propose an unbiased version of $\mathcal{V}_n^2(X, Y)$ and a specific correction for the high-dimensional case $p, q \gg 1$ is investigated in Székely and Rizzo (2013*a*). Further note that Székely et al. (2007) also study $\mathcal{V}^{2(\alpha)}(X, Y)$ defined as

$$\begin{aligned}\mathcal{V}^{2(\alpha)}(X,Y) &= \int_{\mathbb{R}^{p+q}}|\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)|^2 w_\alpha(t,s)dtds \\ &= \mathbb{E}_{X,X',Y,Y'}\|X - X'\|_2^\alpha\|Y - Y'\|_2^\alpha \\ &+ \mathbb{E}_{X,X'}\|X - X'\|_2^\alpha\mathbb{E}_{Y,Y'}\|Y - Y'\|_2^\alpha \\ &- 2\mathbb{E}_{X,Y}\left[\mathbb{E}_{X'}\|X - X'\|_2^\alpha\mathbb{E}_{Y'}\|Y - Y'\|_2^\alpha\right] \end{aligned} \tag{14}$$

with the new weight function $w_\alpha(t, s) = (C(p, \alpha)C(q, \alpha)\|t\|_2^{\alpha+p}\|s\|_2^{\alpha+q})^{-1}$ and constants $C(l, \alpha) = \frac{2\pi^{l/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha\Gamma((l+\alpha)/2)}$ as soon as $\mathbb{E}(\|X\|_2^\alpha + \|Y\|_2^\alpha) < \infty$ and $0 < \alpha < 2$. Distance covariance is retrieved for $\alpha = 1$. The very general case of X and Y living in metric spaces has been examined by Lyons (2013). More precisely, let $(\mathcal{X}, \rho_\mathcal{X})$ and $(\mathcal{Y}, \rho_\mathcal{Y})$ be metric spaces of negative type (see Lyons, 2013), the generalized distance covariance

$$\begin{aligned}\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y) &= \mathbb{E}_{X,X',Y,Y'}\rho_\mathcal{X}(X,X')\rho_\mathcal{Y}(Y,Y') \\ &+ \mathbb{E}_{X,X'}\rho_\mathcal{X}(X,X')\mathbb{E}_{Y,Y'}\rho_\mathcal{Y}(Y,Y') \\ &- 2\mathbb{E}_{X,Y}\left[\mathbb{E}_{X'}\rho_\mathcal{X}(X,X')\mathbb{E}_{Y'}\rho_\mathcal{Y}(Y,Y')\right] \end{aligned} \tag{15}$$

characterizes independence between $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$.

Coming back to sensitivity analysis, just like we defined a new index based on mutual information, we can finally introduce an index based on distance correlation, i.e.

$$S_{X^k}^{dCor} = \mathcal{R}(X^k, Y) \tag{16}$$

which will measure the dependence between an input variable X^k and the output Y. Since distance correlation is designed to detect nonlinear relationships, we except this index to quantify effectively the impact

of X^k on Y. Besides, considering that distance covariance is defined in arbitrary dimension, this index generalizes easily to the multivariate case:

$$S_{X^u}^{dCor} = \mathcal{R}(X^u, Y)$$

for evaluating the impact of a group of inputs X^u on a multivariate output Y.

Remark 1 *The limiting case $\alpha \to 2$ in (14) interestingly leads to $\mathcal{V}^{2(2)}(X,Y) = Cov(X,Y)^2$, see Székely et al. (2007). This turns out to be another original way for defining a new sensitivity index. Indeed, recall that Sobol first-oder sensitivity index actually equals $Cov(Y, Y_{X^k})/Var(Y)$ where Y_{X^k} is an independent copy of Y obtained by fixing X^k, see Janon et al. (2013). The idea is then to replace the covariance (obtained with $\alpha \to 2$) by dCov (with $\alpha = 1$):*

$$S_{X^k}^{dCorPF} = \mathcal{R}(Y, Y_{X^k}), \tag{17}$$

where PF stands for pick-and-freeze, since this index generalizes the pick-and-freeze estimator proposed by Janon et al. (2013) and is able to detect nonlinear dependencies, unlike the correlation coefficient.

3.3 HSIC

3.3.1 Definition

The Hilbert-Schmidt independence criterion proposed by Gretton et al. (2005a) builds upon kernel-based approaches for detecting dependence, and more particularly on cross-covariance operators in RKHSs. Here, we only give a brief summary and introduction on this topic and refer the reader to Berlinet and Thomas-Agnan (2004); Gretton et al. (2005a); Smola et al. (2007) for details.

Let the random vector $X \in \mathcal{X}$ have distribution P_X and consider a RKHS $\mathcal{F}$ of functions $\mathcal{X} \to \mathbb{R}$ with kernel $k_{\mathcal{X}}$ and dot product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$. Similarly, we can also define a second RKHS $\mathcal{G}$ of functions $\mathcal{Y} \to \mathbb{R}$ with kernel $k_{\mathcal{Y}}$ and dot product $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ associated to the random vector $Y \in \mathcal{Y}$ with distribution P_Y. By definition, the cross-covariance operator C_{XY} associated to the joint distribution P_{XY} of (X,Y) is the linear operator $\mathcal{G} \to \mathcal{F}$ defined for every $f \in \mathcal{F}$ and $g \in \mathcal{G}$ as

$$\langle f, C_{XY} g \rangle_{\mathcal{F}} = \mathbb{E}_{XY}[f(X)g(Y)] - \mathbb{E}_X f(X) \mathbb{E}_Y g(Y).$$

In a nutshell, the cross-covariance operator generalizes the covariance matrix by representing higher order correlations between X and Y through nonlinear kernels. For every linear operator $C : \mathcal{G} \to \mathcal{F}$ and provided the sum converges, the Hilbert-Schmidt norm of C is given by

$$\|C\|_{HS}^2 = \sum_{k,l} \langle u_k, C v_l \rangle_{\mathcal{F}}$$

where u_k and v_l are orthonormal bases of $\mathcal{F}$ and $\mathcal{G}$, respectively. This is simply the generalization of the Frobenius norm on matrices. The HSIC criterion is then defined as the Hilbert-Schmidt norm of the cross-covariance operator:

$$\begin{aligned} HSIC(X,Y)_{\mathcal{F},\mathcal{G}} &= \|C_{XY}\|_{HS}^2 \\ &= \mathbb{E}_{X,X',Y,Y'} k_{\mathcal{X}}(X,X') k_{\mathcal{Y}}(Y,Y') \\ &+ \mathbb{E}_{X,X'} k_{\mathcal{X}}(X,X') \mathbb{E}_{Y,Y'} k_{\mathcal{Y}}(Y,Y') \\ &- 2\mathbb{E}_{X,Y} \left[\mathbb{E}_{X'} k_{\mathcal{X}}(X,X') \mathbb{E}_{Y'} k_{\mathcal{Y}}(Y,Y') \right] \end{aligned} \tag{18}$$

where the last equality in terms of kernels is proven in Gretton et al. (2005a). An important property of $HSIC(X,Y)_{\mathcal{F},\mathcal{G}}$ is that it equals 0 if and only if X and Y are independent, as long as the associated RKHSs $\mathcal{F}$ and $\mathcal{G}$ are universal, i.e. they are dense in the space of continuous functions with respect to the infinity norm (Gretton et al., 2005b). Examples of kernels generating universal RKHSs are e.g. the Gaussian and the Laplace kernels (Sriperumbudur et al., 2009).

It is interesting to note the similarity between the generalized distance covariance of equation (15) and the HSIC criterion (18). Actually, Sejdinovic et al. (2013) recently studied the deep connection between these approaches and show that

$$\mathcal{V}^2_{\rho_\mathcal{X},\rho_\mathcal{Y}}(X,Y) = 4\ HSIC(X,Y)_{\mathcal{F},\mathcal{G}}$$

if the kernels $k_\mathcal{X}$ and $k_\mathcal{Y}$ generate the metrics $\rho_\mathcal{X}$ and $\rho_\mathcal{Y}$, respectively (see Sejdinovic et al., 2013). In particular, the standard distance covariance (12) is retrieved with the (universal) kernel $k(z,z') = \frac{1}{2}(\|z\|_2 + \|z'\|_2 - \|z-z'\|_2)$ which generates the metric $\rho(z,z') = \|z-z'\|_2$.

Assume now that $(X_i,Y_i)_{i=1,\dots,n}$ is a sample of the random vector (X,Y) and denote $K_\mathcal{X}$ and $K_\mathcal{Y}$ the Gram matrices with entries $K_\mathcal{X}(i,j) = k_\mathcal{X}(X_i,X_j)$ and $K_\mathcal{Y}(i,j) = k_\mathcal{Y}(Y_i,Y_j)$. Gretton et al. (2005a) propose the following consistent estimator for $HSIC(X,Y)_{\mathcal{F},\mathcal{G}}$:

$$HSIC_n(X,Y)_{\mathcal{F},\mathcal{G}} = \frac{1}{n^2}\mathrm{Tr}\left(K_\mathcal{X} H K_\mathcal{Y} H\right)$$

where H is the centering matrix such that $H(i,j) = \delta_{ij} - \frac{1}{n}$. Besides, it is easy to check that $HSIC_n(X,Y)_{\mathcal{F},\mathcal{G}}$ can be expressed just like the empirical distance covariance (13):

$$\begin{aligned} HSIC_n(X,Y)_{\mathcal{F},\mathcal{G}} = &\frac{1}{n^2}\sum_{i,j=1}^n k_\mathcal{X}(X_i,X_j)k_\mathcal{Y}(Y_i,Y_j) \\ &+ \frac{1}{n^2}\sum_{i,j=1}^n k_\mathcal{X}(X_i,X_j)\frac{1}{n^2}\sum_{i,j=1}^n k_\mathcal{Y}(Y_i,Y_j) \\ &- \frac{2}{n}\sum_{i=1}^n\left[\frac{1}{n}\sum_{j=1}^n k_\mathcal{X}(X_i,X_j)\frac{1}{n}\sum_{j=1}^n k_\mathcal{Y}(Y_i,Y_j)\right]. \end{aligned}$$

An unbiased estimator is also introduced by Song et al. (2012).

We can finally propose a sensitivity index generalizing (16):

$$S_{X^k}^{HSIC_{\mathcal{F},\mathcal{G}}} = \mathcal{R}(X^k,Y)_{\mathcal{F},\mathcal{G}} \tag{19}$$

where the kernel-based distance correlation is given by

$$\mathcal{R}^2(X,Y)_{\mathcal{F},\mathcal{G}} = \frac{HSIC(X,Y)_{\mathcal{F},\mathcal{G}}}{\sqrt{HSIC(X,X)_{\mathcal{F},\mathcal{F}}HSIC(Y,Y)_{\mathcal{G},\mathcal{G}}}}$$

and the kernels inducing $\mathcal{F}$ and $\mathcal{G}$ have to be chosen within the class of universal kernels. The multivariate extension of $S_{X^k}^{HSIC_{\mathcal{F},\mathcal{G}}}$ is straightforward. The impact of the choice of kernels has previously been studied by Sriperumbudur et al. (2009) in the context of independence hypothesis tests.

Remark 2 *Instead of working with the cross-covariance operator C_{XY}, Fukumizu et al. (2007) work with the normalized cross-covariance operator (NOCCO) V_{XY} defined as $C_{XY} = C_{YY}^{1/2}V_{XY}C_{XX}^{1/2}$, see Fukumizu et al. (2007) for the existence of this representation. Just as the HSIC criterion, the associated measure of dependence is given by $I^{NOCCO}(X,Y) = \|V_{XY}\|^2_{HS}$. Interestingly, $I^{NOCCO}(X,Y)$ is independent of the choice of kernels and is actually equal to the squared-loss mutual information (7) under some assumptions, see Fukumizu et al. (2008). Despite the advantage of being kernel-free, using I^{NOCCO} in practice unfortunately requires to work with an estimator with a regularization parameter, which has to be selected (Fukumizu et al., 2007). Nevertheless, it is still interesting to use this approach for approximating SMI efficiently, since dimensionality limitations related to density function estimation no longer apply.*

Remark 3 *The pick-and-freeze estimator defined in Remark 1 can be readily generalized with kernels:*

$$S_{X^k}^{HSIC_{\mathcal{F},\mathcal{G}}PF} = \mathcal{R}(Y,Y_{X^k})_{\mathcal{G},\mathcal{G}} \tag{20}$$

where this time only the kernel acting on $\mathcal{Y}$ needs to be specified.

3.3.2 Going beyond $Y \in \mathbb{R}^q$

The kernel point of view in HSIC also provides an elegant and powerful framework for dealing with categorical inputs and outputs, as well as functional ones.

The categorical case is common practice in feature selection, since the target output is often represented as labels. Appropriate kernels include for example $k_{\mathcal{Y}}(y, y') = \delta_{yy'}/n_y$ where n_y is the number of samples with label y, see e.g. Song et al. (2012); Yamada et al. (2013). From a GSA perspective, this implies that we can evaluate the impact of the inputs on level sets of the output by a simple change of variable $Z = \mathbb{1}\{Y > t\}$ for a given threshold t. We can note the resemblance with the approach of Fort et al. (2013) if one uses a contrast function adapted to exceedance probabilities.

As a matter of fact, it is also possible to design dedicated semi-metrics for functional data which can be incorporated in the definition of the kernels, see e.g. Ferraty and Vieu (2006). For example, let $\Delta(\cdot, \cdot)$ be such a semi-metric defined on $\mathcal{Y} \times \mathcal{Y}$ when the output variable is of functional type. The kernel associated to $\mathcal{Y}$ is then given by $k_{\mathcal{Y}}(y, y') = k(\Delta(y, y'))$ where k is a kernel acting on $\mathbb{R}$. The same scheme applies to functional inputs as well, see Ginsbourger et al. (2012) for an illustration in the context of surrogate modeling where the semi-metric is a cheap and simplified computer code. However, a theoretical shortcoming lies in our current inability to check if such semi-metric kernels are universal, which implies that we can not claim that independence can be detected. Despite this deficiency, we show in Section 4 that from a practical perspective, the use of a semi-metric based on principal components can efficiently deal with a functional output given as a 2D map.

3.4 Feature selection as an alternative to screening

In machine learning, feature selection aims at identifying relevant features (among a large set) with respect to a prediction task. The goal is to detect irrelevant or redundant features which may increase the prediction variance without reducing its bias. As a matter of fact, this closely resembles the objective of factor screening in GSA. The main difference is that in GSA, input variables are usually assumed to be independent, whereas in feature selection redundant features, i.e. highly dependent factors, precisely have to be filtered out. This apparently naive distinction actually makes feature selection an interesting alternative to screening when some input variables are correlated. But it is important to note that it is also a powerful option even in the independent case. We do not intend here to give an exhaustive review of feature selection techniques, but rather detail some approaches which make use of the dependence measures we recapped above. We hope that it will illustrate how they can be used as new screening procedures in high dimensional problems.

Literature on feature selection is abundant and entails many approaches. In the high dimensional setting, model-based techniques include for example the Lasso (Tibshirani, 1996) or sparse additive models (Ravikumar et al., 2009), see Fan and Lv (2010) for a selective overview. Generalizations for the ultra-high dimensional case usually replace penalty-based techniques to focus on marginal regression, where an underlying model is still assumed (e.g. linear Fan and Lv (2008) or non-parametric additive Fan et al. (2011)). Another line of work for the ultra-high dimensional setting are model-free methods, where only dependence measures are used to identify relevant features. Except for the very specific HSIC Lasso technique (Yamada et al., 2013), here we only focus on pure dependence-based approaches.

Let us first introduce the concept of Max-Dependency (Peng et al., 2005). Denote $X^1, \ldots, X^p$ the set of available features, Y the target output to predict and $D(\cdot, \cdot)$ any measure quantifying the dependence between two random vectors. The Max-Dependency scheme for feature selection involves finding m features $X^{i_1}, \ldots, X^{i_m}$ which jointly have the largest dependency with Y, i.e. one has to solve the following optimization problem

$$\max_{\{i_1,\ldots,i_m\} \subset \{1,\ldots,p\}} D(\left\{X^{i_1}, \ldots, X^{i_m}\right\}, Y). \tag{21}$$

Solving (21) is however computationally infeasible when m and p are large for cardinality reasons. Near-optimal solutions are then usually found by iterative procedures, where features are added one at a time in the subset $X^{i_1}, \ldots, X^{i_m}$ (forward selection). On the other hand, the dependence measure $D(\cdot, \cdot)$ must also

be robust to dimensionality, which is hard to achieve in practice when the number of samples is less than m. Consequently, marginal computations which only involve $D(X^k, Y)$ terms are usually preferred. The Max-Relevance criterion (Peng et al., 2005) serves in this context as a proxy to Max-Dependency, where the optimization problem writes

$$\max_{\{i_1,\dots,i_m\}\subset\{1,\dots,p\}} \frac{1}{m}\sum_{k=1}^{m} D(X^{i_k}, Y). \tag{22}$$

But when the features are dependent, it is likely that this criterion will select redundant features. To limit this effect, one can add a condition of Min-Redudancy expressed as

$$\min_{\{i_1,\dots,i_m\}\subset\{1,\dots,p\}} \frac{1}{m^2}\sum_{k,l=1}^{m} D(X^{i_k}, X^{i_l}). \tag{23}$$

The final scheme combining (22) and (23), called minimal-redundancy-maximal-relevance (mRMR), is given by

$$\max_{\{i_1,\dots,i_m\}\subset\{1,\dots,p\}} \frac{1}{m}\sum_{k=1}^{m} D(X^{i_k}, Y) - \frac{1}{m^2}\sum_{k,l=1}^{m} D(X^{i_k}, X^{i_l}). \tag{24}$$

Forward and backward procedures for mRMR are investigated by Peng et al. (2005) where $D(\cdot,\cdot)$ is chosen as the mutual information. Similarly, forward and backward approaches where MI is replaced with HSIC is introduced by Song et al. (2012). A purely marginal point of view is studied by Li et al. (2012) where the authors propose the dCor criterion (11). In a nutshell, the dCor measure is computed between Y and each factor X^k, $k = 1,\dots,p$ and only the features with dCor above a certain threshold are retained. A sure screening property of this approach is also proven. Balasubramanian et al. (2013) extend this work by considering a modified version of the HSIC dependence measure (supremum of HSIC over a family of universal kernels, denoted sup-HSIC). Even if the sure screening procedure of this generalized method is proven, the authors mention that every feature selection technique based on marginal computations fails at detecting features that may be marginally uncorrelated with the output but are in fact jointly correlated with it. As a result, they propose the following iterative approach:

1. Compute the marginal sup-HSIC measures between Y and each feature X^k, $k = 1,\dots,p$ and select the inputs with a measure above a given threshold. Let $u \subset \{1,\dots,p\}$ be the subset of selected features.

2. Compute sup-HSIC between Y and $\left(X^u, X^k\right)$ for each $k \notin u$. Augment u with features having a measure greater than the sup-HSIC criterion between Y and X^u.

3. Repeat until the subset of selected features stabilizes or when its cardinality reaches a given maximum value.

As pointed out previously, another drawback of marginal computations which is not taken care of by the above scheme is that redundant variables are not eliminated. But Balasubramanian et al. (2013) design another iterative procedure to deal with this case. Finally, let us note that in the examples of Section 4, we will only study the above iterative technique since we focus on independent input factors. We plan to investigate in particular the full mRMR approach for problems with correlated inputs in a future work.

Instead of working with forward and backward approaches, Yamada et al. (2013) propose a combination of the Lasso and the HSIC dependence measure. Denote $\tilde{K}^k_{\mathcal{X}} = HK^k_{\mathcal{X}}H$ for $k = 1,\dots,p$ and $\tilde{K}_{\mathcal{Y}} = HK_{\mathcal{Y}}H$ the centered Gram matrices computed from a sample $\left(X^1_i,\dots,X^p_i,Y_i\right)_{i=1,\dots,n}$ of $\left(X^1,\dots,X^p,Y\right)$ following the notations of Section 3.3. The HSIC Lasso solves the following optimization problem

$$\min_{\alpha\in\mathbb{R}^p} \frac{1}{2}\|\tilde{K}_{\mathcal{Y}} - \sum_{k=1}^{p}\alpha_k\tilde{K}^k_{\mathcal{X}}\|^2_{\text{Frob}} + \lambda\|\alpha\|_1 \tag{25}$$

with constraints $\alpha_1, \ldots, \alpha_p \geq 0$ and where $\|\cdot\|_{\text{Frob}}$ stands for the Frobenius norm and λ is a regularization parameter. Interestingly, the first term of equation (25) expands as

$$\frac{1}{2n^2}\|\tilde{K}_{\mathcal{Y}} - \sum_{k=1}^{p} \alpha_k \tilde{K}_{\mathcal{X}}^k\|_{\text{Frob}}^2 = \frac{1}{2} HSIC_n(Y,Y) - \sum_{k=1}^{p} \alpha_k HSIC_n(X^k, Y) + \frac{1}{2} \sum_{k,l=1}^{p} \alpha_k \alpha_l HSIC_n(X^k, X^l)$$

using that $\tilde{K}_{\mathcal{X}}^k$, $\tilde{K}_{\mathcal{Y}}$ are symmetric and H is idempotent, which highlights the strong correspondence with the mRMR criterion (24). The authors show that (25) can be recast as a standard Lasso program and propose a dual augmented Lagrangian algorithm to solve the optimization problem. They also discuss a variant based on the I^{NOCCO} dependence measure.

Remark 4 *We mentioned before that feature selection techniques based on dependence measures have been particularly designed for the ultra-high dimensional case, which is not the common setting of screening problems in GSA. Nevertheless, we illustrate in Section 4 that they perform remarkably well on complex benchmark functions, while requiring very few samples of the output variable. This reveals their high potential for preliminary screening of expensive computer codes.*

4 Experiments

In this Section, we finally assess the performance of all the new sensitivity indices introduced before on a series of benchmark analytical functions and two industrial applications. All benchmark functions can be found in the Virtual Library of Simulation Experiments available at `http://www.sfu.ca/~ssurjano/index.html`. For easier comparison, we first summarize the proposed indices in Table 1 (SI stands for sensitivity index and see Tarantola et al. (2006) for RBD-FAST).

Index	Origin	Notes
S_X^1	Sobol first-order SI	Normalized version (RBD-FAST est.)
S_X^{TOT}	Sobol total SI	Normalized version (RBD-FAST est.)
S_X^f (eq. (3))	Csiszár f-divergences	Includes as special case Borgonovo (2007), S_X^{MI} (eq. (8)) and S_X^{SMI} (eq. (9))
S_X^{dCor} (eq. (16))	Distance correlation	
S_X^{dCorPF} (eq. (17))	Pick-and-freeze distance correlation	Generalization of Janon et al. (2013)
$S_X^{HSIC_{\mathcal{F},\mathcal{G}}}$ (eq. (19)	HSIC	Can be extended to categorical/functional data
$S_X^{HSIC_{\mathcal{F},\mathcal{G}}PF}$ (eq. (20)	Pick-and-freeze HSIC	Generalization of Janon et al. (2013)

Table 1: Summary of sensitivity indices.

4.1 Analytical functions

Standard GSA. For the first experiments, we focus on GSA problems where the dimensionality is not too large (less than 10 input variables). The objective is to compare the information given by the new indices with Sobol first-order and total indices.

- Linkletter et al. (2006) decreasing function

$$\eta_1(X) = 0.2X_1 + \frac{0.2}{2}X_2 + \frac{0.2}{4}X_3 + \frac{0.2}{8}X_4 + \frac{0.2}{16}X_5 + \frac{0.2}{32}X_6 + \frac{0.2}{64}X_7 + \frac{0.2}{128}X_8$$

  with $X_i \sim \mathcal{U}(0,1)$, $i = 1, \ldots, 10$.

We compute the sensitivity indices based on Csiszár f-divergences, dCor, pick-and-freeze dCor, HSIC and pick-and-freeze HSIC (Gaussian kernels) with a sample of size $n = 500$ and we repeat this calculation 100 times. Here we use a simple kernel density estimator since we only study first-order indices. Results are given in Figure 1. Analytical first-order SIs are $S^1_{X_k} = S^{TOT}_{X_k} = \frac{3}{4}\left(\frac{1}{4}\right)^{i-1} / (1 - (\frac{1}{4})^{10})$ for $i = 1, \ldots, 10$, which is coherent with the estimates at the top left. As expected, indices given by Csiszár f-divergences, S^{dCor}_X, S^{dCorPF}_X, $S^{HSIC_{\mathcal{F},\mathcal{G}}}_X$ and $S^{HSIC_{\mathcal{F},\mathcal{G}}PF}_X$ provide the same information as variance-based ones in this simple case of a linear model. However, let us note that dCor and HSIC detect non-influential factors very easily and robustly.

- Loeppky et al. (2013) function

$$\eta_2(X) = 6X_1 + 4X_2 + 5.5X_3 + 3X_1X_2 + 2.2X_1X_3 + 1.4X_2X_3 + X_4 + 0.5X_5 + 0.2X_6 + 0.1X_7$$

  with $X_i \sim \mathcal{U}(0,1)$, $i = 1, \ldots, 10$ (the original function has constraint $\sum_{i=1}^{10} X_i = 1$ but we do not consider it here).

Conclusions are similar for the Loeppky et al. (2013) function, where only the first three inputs have a large impact on the output with very small interactions (total SIs almost equal first-order SIs), see Figure 2. Note that S^{dCorPF}_X and $S^{HSIC_{\mathcal{F},\mathcal{G}}PF}_X$ recover S^1_X since interactions are small. Again, dCor and HSIC clearly identify inputs which are independent of the output.

- Ishigami function (Ishigami and Homma, 1990)

$$\eta_3(X) = \sin(X_1) + 5\sin^2(X_2) + 0.1(X_3)^4 \sin(X_1)$$

  with $X_i \sim \mathcal{U}(-\pi, \pi)$, $i = 1, \ldots, 3$ and constants taken from Borgonovo (2007).

This time we also compute S^{TOT}_X since η_3 encompasses a strong iteration term. We use a sample of size $n = 200$ for computing S^1_X, S^f_X, S^{dCor}_X and S^{dCorPF}_X, but now we also need a sample of size $n \times p = 200 \times 3 = 600$ for S^{TOT}_X with RBD-FAST. Estimates obtained with 100 replications are reported in Figure 3. While first-order SIs indicate that X_3 has a negligible impact, it actually influences the output through an interaction term which is naturally accounted for by the total index. Is is interesting to note that all other indices detect the impact of X_3, as was pointed out by Borgonovo (2007) for the total-variation index. However, one can observe the striking adequacy between S^{TOT}_X and S^{dCor}_X (unlike S^f_X). This clearly shows that distance correlation has the potential to detect any interaction effect since it is specifically designed for nonlinear dependence. An additional appealing property is that its estimation does not depend on the number of inputs, unlike S^{TOT}_X. This is a major advantage for expensive computer codes. Finally, S^{dCorPF}_X tends to bring the same information as S^f_X. But recall that it has the same limitation as S^{TOT}_X concerning computational cost due to the pick-and-freeze technique. The same comments apply to $S^{HSIC_{\mathcal{F},\mathcal{G}}}_X$ and $S^{HSIC_{\mathcal{F},\mathcal{G}}PF}_X$.

We also investigate HSIC on level sets of the Ishigami function to compare our results with Fort et al. (2013). More precisely, we use a categorical kernel and use the change of variable $Z = \mathbb{1}\{\eta_3(X) > 10\}$. Figure 4 shows that we can recover the fact that input factor X_3 is more important than X_1 and X_2 for this level set function, as was observed by Fort et al. (2013).

Screening. We now propose to study the performance of feature selection as an alternative to screening for problems where the number of input variables is large (more than 20). We will deliberately limit the number of samples in order to be as close as possible to a real test case on an expensive code.

- Morris et al. (2006) function

$$\eta_4(X) = \alpha \sum_{i=1}^{k} \left(X_i + \beta \prod_{i<j=2}^{k} X_i X_j \right)$$

  where $\alpha = \sqrt{12} - 6\sqrt{0.1(k-1)}$, $\beta = \sqrt{12}\sqrt{0.1(k-1)}$, $X_i \sim \mathcal{U}(0,1)$, $i = 1, \ldots, 30$ and $1 \leq k \leq 10$ is an integer controlling the number of influential inputs.

We select $n = 50$, $k = 5$ and compute S_X^1, S_X^f, S_X^{dCor} and $S_X^{HSIC_{\mathcal{F},\mathcal{G}}}$ since all other indices are too expensive to compute in this setting (recall that $p = 30$). First remark that first-order SIs identify the influential inputs in mean, but there are many replicates for which they are confounded with non-influential ones. On the contrary, S_X^f completely fails ate detecting them: it will then be excluded from the other tests on screening. Notably, dCor and HSIC perfectly discriminate the first five factors and identify the remaining ones as independent from the output. We also use the HSIC Lasso (25), where for each replicate we use a bootstrap procedure to evaluate the probability of selection of each input factor. Here, HSIC Lasso performs very well since it selects the first 5 inputs factors almost every time.

- Sobol and Levitan (1999) function

$$\eta_5(X) = \exp\left(\sum_{i=1}^{20} b_i X_i \right) - \prod_{i=1}^{p} \frac{\exp(b_i) - 1}{b_i}$$

  with $X_i \sim \mathcal{U}(0,1)$, $i = 1, \ldots, 20$ and b_i are taken from Moon et al. (2012). Only the first eight factors are influential.

Here we also illustrate the feature selection method based on the iterative HSIC scheme detailed in Section 3.4. Results for $n = 50$ are given in Figure 6. As expected, S_X^1 is unable to detect correctly the impact of the inputs. On the other hand, dCor and HSIC accurately estimate higher dependence for the first input factors than for the remaining ones. Similarly, HSIC Lasso never selects the last inputs as influential ones. The iterative feature selection based on HSIC performs well but tends to select more inputs than necessary.

To go further and to compare our results with the ones obtained by Moon et al. (2012), we repeat this experiment with $n = 100$ (approximately the sample size used by Moon et al. (2012)). This time first-order SIs slightly detect the influential inputs, but the more interesting fact is that dCor, HSIC and HSIC Lasso give even better results and almost perfectly identifies them. Finally, the iterative HSIC scheme now almost always discards the non-influential inputs.

4.2 Industrial applications

Acquisition strategy for reservoir characterization. In the petroleum industry, reservoir characterization aims at reducing the uncertainty on some unknown physical parameters of an oil reservoir by using all the data collected on the field, e.g. well logs, seismic images or dynamic data at the wells (pressures, ...). Basically, engineers solve a Bayesian inverse problem where an initial prior distribution assumed on the parameters is updated by incorporating all field observations to produce a posterior distribution. In the end, this posterior distribution is used to predict the expected oil recovery of the reservoir in the future. However, it may be expensive to collect data and usually one wants to gather relevant observations only. This principle is at the core of so-called data acquisition strategies. For example, given the prior distribution, a natural idea is to get data which, when incorporated in the Bayesian procedure, will reduce the most the uncertainty of the obtained posterior distribution. It is easy to see that this idea actually corresponds to performing

a sensitivity analysis of the parameters when data varies, where the difference between the prior and the posterior distribution is given by the measure of uncertainty reduction one chooses, i.e. the dissimilarity function $d(\cdot, \cdot)$ in equation (1). Since the number of both uncertain parameters and observations can be large, we can greatly capitalize on the advantages of dCor and HSIC measures in arbitrary dimension to perform this task.

Our example here makes use of the Punq reservoir test case, which is an oil reservoir model derived from real field data (Manceau et al., 2001). In this simplified model, seven variables which are characteristic of media, rocks, fluids or aquifer activity, are considered as uncertain (permeability multipliers, residual oil saturations, ...) and are assigned a uniform prior distribution. For illustration purposes, we assume that collectable data only consist of gas-oil ratios measurements at a given well. We generate a sample of size $n = 100$ of the prior distribution, and propagate them through a fluid-flow simulator to get a sample of the simulated gas-oil ratios at the well over 10 years. They are given in Figure 8, top. For each day in these 10 years, we compute the dCor measure between the parameters and the simulated ratio, see Figure 8, bottom. This information makes it possible to pick up the days where measurements should be collected in order to reduce as much as possible the uncertainty on the parameters: at the beginning of the reservoir production (before 500 days) or around 3 years after. Obviously, this procedure generalizes to any number of measurements and can be performed sequentially on many observations thanks to the properties of dCor.

Screening for contamination migration in waste storage site. The Marthe test case investigated here concerns prediction of the transport of strontium 90 in a porous water-saturated medium for evaluating the contamination of an aquifer in a temporary storage of radioactive waste (Volkova et al., 2008). Twenty input parameters mainly representative of the geological uncertainty are considered as random, and a set of 300 simulations is available at `http://www.gdr-mascotnum.fr/benchmarks.html`. Accessible outputs are strontium 90 concentrations simulated at ten different wells, as well as the concentrations on a complete 2D map of the area (discretized on $64 \times 64 = 4096$ pixels). We place ourselves in a screening setting where we use only $n = 50$ simulations to identify the influential inputs. To estimate the variability of our results, we pick at random these 50 samples among the 300 available and repeat the procedure 100 times. We use the HSIC dependence measure with a Gaussian kernel first in its standard form by considering the vector of concentrations at the 10 observation wells. But we also take advantage of a kernel designed for the 2D maps as was mentioned in Section 3.3.2. Namely we use the PCA semi-metric (Ferraty and Vieu, 2006) and vary the number of principal components (1, 5 and 20 explaining 50%, 80% and 95% of the total variance, respectively). Results are given in Figure 9. First note that they are coherent with the ones obtained by Volkova et al. (2008) where the authors used the 300 simulations to build a surrogate model. Here, we then get the same detection of influential inputs but with only 50 simulations (parameters i3, kd1, kd2). In addition, the PCA kernel leads to more discriminating indices as soon as the explained variance is sufficient (5 PCs). This clearly illustrates the potential of HSIC for functional data.

5 Conclusion

In this paper, we introduced a new class of sensitivity indices based on dependence measures which overcomes the insufficiencies of variance-based methods in GSA. We demonstrated that when the output distribution is compared with its conditional counterpart through Csiszár f-divergences, sensitivity indices arise as well-known dependence measures between random variables. We then extended these indices by using recent state-of-the-art dependence measures, such as distance correlation and the Hilbert-Schmidt independence criterion. We also emphasized the potential of feature selection techniques relying on such dependence measures as alternatives to screening in high dimension.

Interestingly, these new sensitivity indices are very robust to dimensionality, have low computational cost and can be elegantly extended to functional and categorical output or input variables. This opens the door to new and powerful tools for GSA and factors screening for high dimensional and expensive computer codes.

References

Auder, B. and Iooss, B. (2008), Global sensitivity analysis based on entropy, *in* 'Safety, Reliability and Risk Analysis-Proceedings of the ESREL 2008 Conference', pp. 2107–2115.

Balasubramanian, K., Sriperumbudur, B. and Lebanon, G. (2013), Ultrahigh dimensional feature screening via rkhs embeddings, *in* 'Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics', pp. 126–134.

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

Berlinet, A. and Thomas-Agnan, C. (2004), *Reproducing kernel Hilbert spaces in probability and statistics*, Vol. 3, Kluwer Academic Boston.

Blatman, G. and Sudret, B. (2010), 'Efficient computation of global sensitivity indices using sparse polynomial chaos expansions', *Reliability Engineering & System Safety* **95**(11), 1216 – 1229.

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

Csiszár, I. (1967), 'Information-type measures of difference of probability distributions and indirect observations', *Studia Sci. Math. Hungar.* **2**, 299–318.

Cukier, R. I., Fortuin, C. M., Shuler, K. E., Petschek, A. G. and Schaibly, J. H. (1973), 'Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory', *The Journal of Chemical Physics* **59**, 3873–3878.

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.

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.

Fan, J., Feng, Y. and Song, R. (2011), 'Nonparametric independence screening in sparse ultra-high-dimensional additive models', *Journal of the American Statistical Association* **106**(494), 544–557.

Fan, J. and Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 849–911.

Fan, J. and Lv, J. (2010), 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica* **20**(1), 101–148.

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

Fort, J.-C., Klein, T. and Rachdi, N. (2013), 'New sensitivity analysis subordinated to a contrast', *arXiv preprint arXiv:1305.2329* .

Fukumizu, K., Bach, F. R. and Gretton, A. (2007), 'Statistical consistency of kernel canonical correlation analysis', *The Journal of Machine Learning Research* **8**, 361–383.

Fukumizu, K., Gretton, A., Sun, X. and Schölkopf, B. (2008), 'Kernel measures of conditional dependence', *Advances in Neural Information Processing Systems* **20**, 489–496.

Gamboa, F., Janon, A., Klein, T. and Lagnoux, A. (2013), 'Sensitivity indices for multivariate outputs', *to appear in Comptes Rendus de l'Académie des Sciences* .

Ginsbourger, D., Rosspopoff, B., Pirot, G., Durrande, N. and Renard, P. (2012), 'Distance-based kriging relying on proxy simulations for inverse conditioning', *Advances in Water Resources* **52**, 275–291.

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

Hyvärinen, A. and Oja, E. (2000), 'Independent component analysis: algorithms and applications', *Neural networks* **13**(4), 411–430.

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

Janon, A., Klein, T., Lagnoux-Renaudie, A., Nodet, M. and Prieur, C. (2013), 'Asymptotic normality and efficiency of two sobol index estimators', *to appear in ESAIM: Probability and Statistics* .

Kanamori, T., Hido, S. and Sugiyama, M. (2009), 'A least-squares approach to direct importance estimation', *The Journal of Machine Learning Research* **10**, 1391–1445.

Kraskov, A., Stögbauer, H. and Grassberger, P. (2004), 'Estimating mutual information', *Physical Review E* **69**(6), 066138.1–066138.16.

Krzykacz-Hausmann, B. (2001), Epistemic sensitivity analysis based on the concept of entropy, *in* 'SAMO 2001', pp. 53–57.

Li, R., Zhong, W. and Zhu, L. (2012), 'Feature screening via distance correlation learning', *Journal of the American Statistical Association* **107**(499), 1129–1139.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D. and Kenny, Q. Y. (2006), 'Variable selection for gaussian process models in computer experiments', *Technometrics* **48**(4), 478–490.

Loeppky, J. L., Williams, B. J. and Moore, L. M. (2013), 'Global sensitivity analysis for mixture experiments', *Technometrics* **55**(1), 68–78.

Lyons, R. (2013), 'Distance covariance in metric spaces', *Annals of Probability* **41**(5), 3284–3305.

Manceau, E., Mezghani, M., Zabalza-Mezghani, I. and Roggero, F. (2001), Combination of experimental design and joint modeling methods for quantifying the risk associated with deterministic and stochastic uncertainties - an integrated test study, *in* '2001 SPE Annual Technical Conference and Exhibition, New Orleans, 30 September-3 October'. paper SPE 71620.

Marrel, A., Iooss, B., Laurent, B. and Roustant, O. (2009), 'Calculations of sobol indices for the gaussian process metamodel', *Reliability Engineering & System Safety* **94**(3), 742–751.

Meilă, M. (2007), 'Comparing clusterings–an information based distance', *Journal of Multivariate Analysis* **98**(5), 873–895.

Moon, H., Dean, A. M. and Santner, T. J. (2012), 'Two-stage sensitivity-based group screening in computer experiments', *Technometrics* **54**(4), 376–387.

Morris, M. D. (1991), 'Factorial sampling plans for preliminary computational experiments', *Technometrics* **33**(2), 161–174.

Morris, M. D., Moore, L. M. and McKay, M. D. (2006), 'Sampling plans based on balanced incomplete block designs for evaluating the importance of computer model inputs', *Journal of Statistical Planning and Inference* **136**(9), 3203–3220.

Müller, A. (1997), 'Integral probability metrics and their generating classes of functions', *Advances in Applied Probability* **29**(2), 429–443.

Oakley, J. E. and O'Hagan, A. (2004), 'Probabilistic sensitivity analysis of complex models : a bayesian approach', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**, 751–769.

Owen, A. B. (2013), 'Better estimation of small sobol' sensitivity indices', *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **23**(2), 11.

Peng, H., Long, F. and Ding, C. (2005), 'Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**(8), 1226–1238.

Pluim, J. P., Maintz, J. A. and Viergever, M. A. (2003), 'Mutual-information-based registration of medical images: a survey', *Medical Imaging, IEEE Transactions on* **22**(8), 986–1004.

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009), 'Sparse additive models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(5), 1009–1030.

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.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013), 'Equivalence of distance-based and rkhs-based statistics in hypothesis testing', *to appear in Annals of Statistics* .

Shannon, C. E. (1948), 'A mathematical theory of communication', *Bell System Technical Journal* **27**(3), 379–423.

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. and Levitan, Y. L. (1999), 'On the use of variance reducing multipliers in monte carlo computations of a global sensitivity index', *Computer Physics Communications* **117**(1), 52–61.

Sobol', I. M. (1993), 'Sensitivity estimates for nonlinear mathematical models', *MMCE* **1**, 407–414.

Sobol, I. M. and Kucherenko, S. (2009), 'Derivative based global sensitivity measures and their link with global sensitivity indices', *Mathematics and Computers in Simulation* **79**(10), 3009–3017.

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.

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., Fukumizu, K., Gretton, A., Schölkopf, B. and Lanckriet, G. R. (2012), 'On the empirical estimation of integral probability metrics', *Electronic Journal of Statistics* **6**, 1550–1599.

Sugiyama, M., Suzuki, T. and Kanamori, T. (2012), *Density ratio estimation in machine learning*, Cambridge University Press.

Suzuki, T., Sugiyama, M., Kanamori, T. and Sese, J. (2009), 'Mutual information estimation reveals global associations between stimuli and biological processes', *BMC bioinformatics* **10**(Suppl 1), S52.

Suzuki, T., Sugiyama, M., Sese, J. and Kanamori, T. (2008), 'Approximating mutual information by maximum likelihood density ratio estimation', *Journal of Machine Learning Research-Proceedings Track* **4**, 5–20.

Székely, G. J. and Rizzo, M. L. (2013*a*), 'The distance correlation t-test of independence in high dimension', *Journal of Multivariate Analysis* **117**, 193–213.

Székely, G. J. and Rizzo, M. L. (2013*b*), 'Energy statistics: A class of statistics based on distances', *Journal of Statistical Planning and Inference* **143**(8), 1249–1272.

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.

Tarantola, S., Gatelli, D. and Mara, T. A. (2006), 'Random balance designs for the estimation of first order global sensitivity indices', *Reliability Engineering & System Safety* **91**(6), 717–727.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Tissot, J.-Y. and Prieur, C. (2012), 'Bias correction for the estimation of sensitivity indices based on random balance designs', *Reliability Engineering & System Safety* **107**, 205–213.

Touzani, S. and Busby, D. (2012), 'Smoothing spline analysis of variance approach for global sensitivity analysis of computer codes', *Reliability Engineering & System Safety* **112**, 67–81.

Volkova, E., Iooss, B. and Van Dorpe, F. (2008), 'Global sensitivity analysis for a numerical model of radionuclide migration from the rrc kurchatov institute radwaste disposal site', *Stochastic Environmental Research and Risk Assessment* **22**(1), 17–31.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M. (2013), 'High-dimensional feature selection by feature-wise non-linear lasso', *to appear in Neural Computation* .

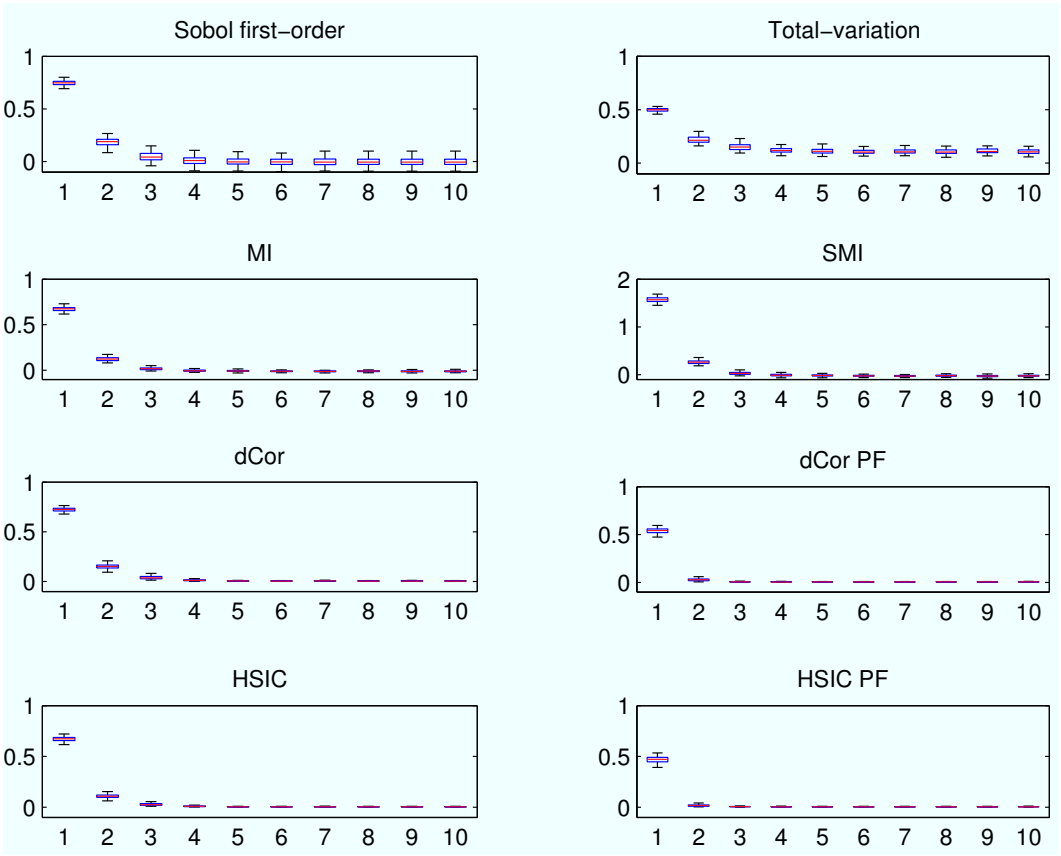

Figure 1: First-order SI, S_X^f, S_X^{dCor}, S_X^{dCorPF}, $S_X^{HSIC_{\mathcal{F},\mathcal{G}}}$ and $S_X^{HSIC_{\mathcal{F},\mathcal{G}}PF}$ for function η_1, $n = 500$, 100 replicates.

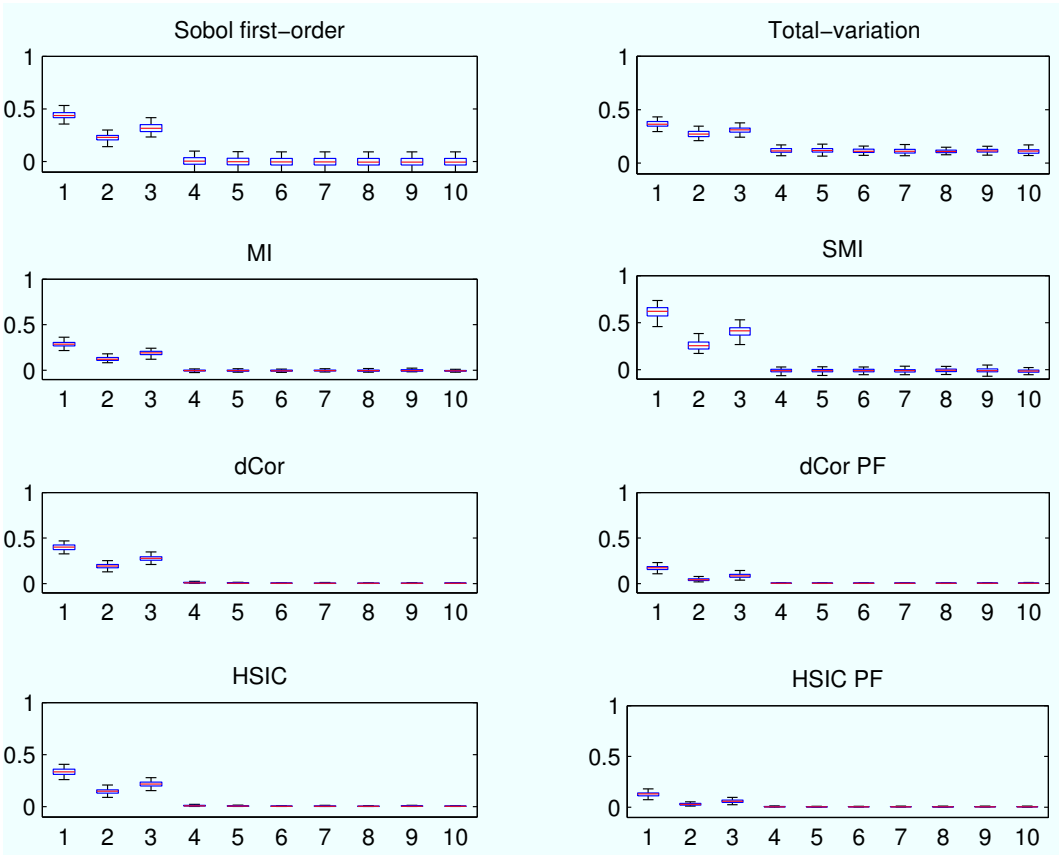

Figure 2: First-order SI, S_X^f, S_X^{dCor}, S_X^{dCorPF}, $S_X^{HSIC_{\mathcal{F},\mathcal{G}}}$ and $S_X^{HSIC_{\mathcal{F},\mathcal{G}}PF}$ for function η_2, $n = 500$, 100 replicates.

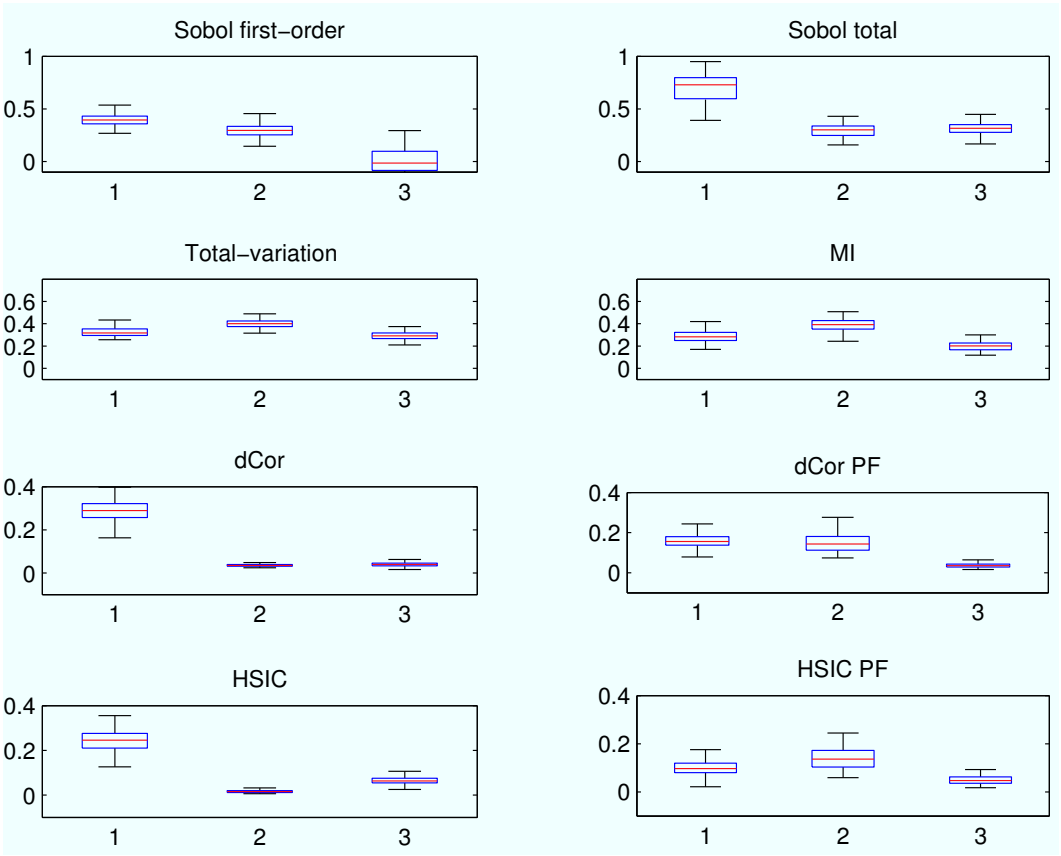

Figure 3: First-order SI, total SI, S_X^f, S_X^{dCor}, S_X^{dCorPF}, $S_X^{HSIC_{\mathcal{F},\mathcal{G}}}$ and $S_X^{HSIC_{\mathcal{F},\mathcal{G}}PF}$ for function η_3, $n = 200$, 100 replicates.

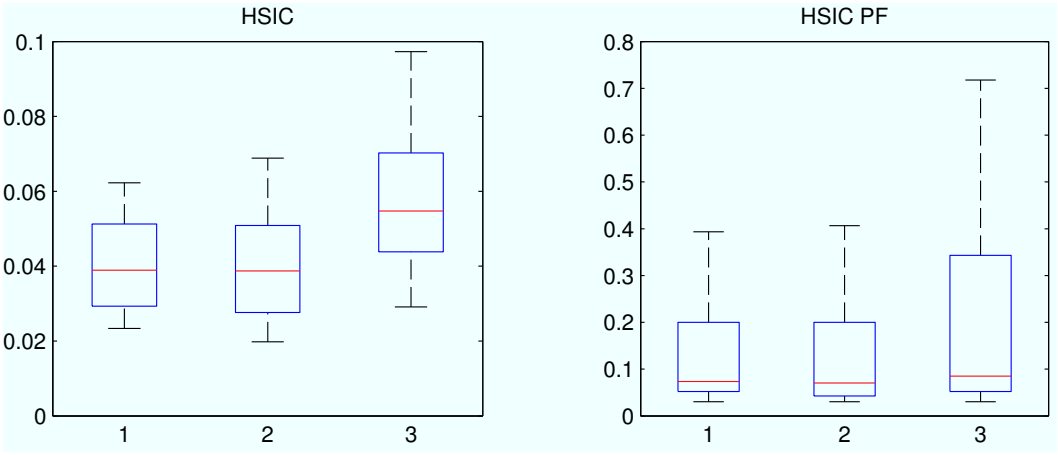

Figure 4: $S_X^{HSIC_{\mathcal{F},\mathcal{G}}}$ and $S_X^{HSIC_{\mathcal{F},\mathcal{G}}PF}$ for function $\mathbb{1}\{\eta_3 > 10\}$, $n = 200$, 100 replicates.

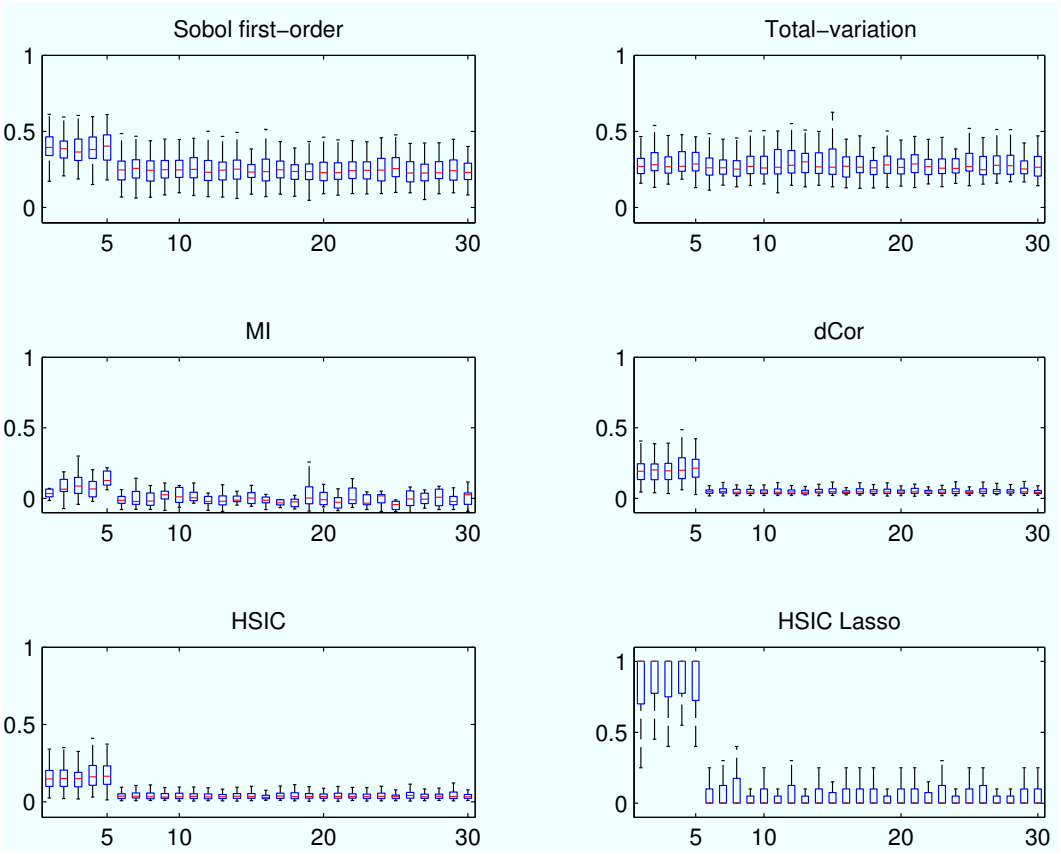

Figure 5: First-order SI, S_X^f, S_X^{dCor}, $S_X^{HSIC_{\mathcal{F},\mathcal{G}}}$ and HSIC Lasso for function η_4, $n = 50$, 100 replicates.

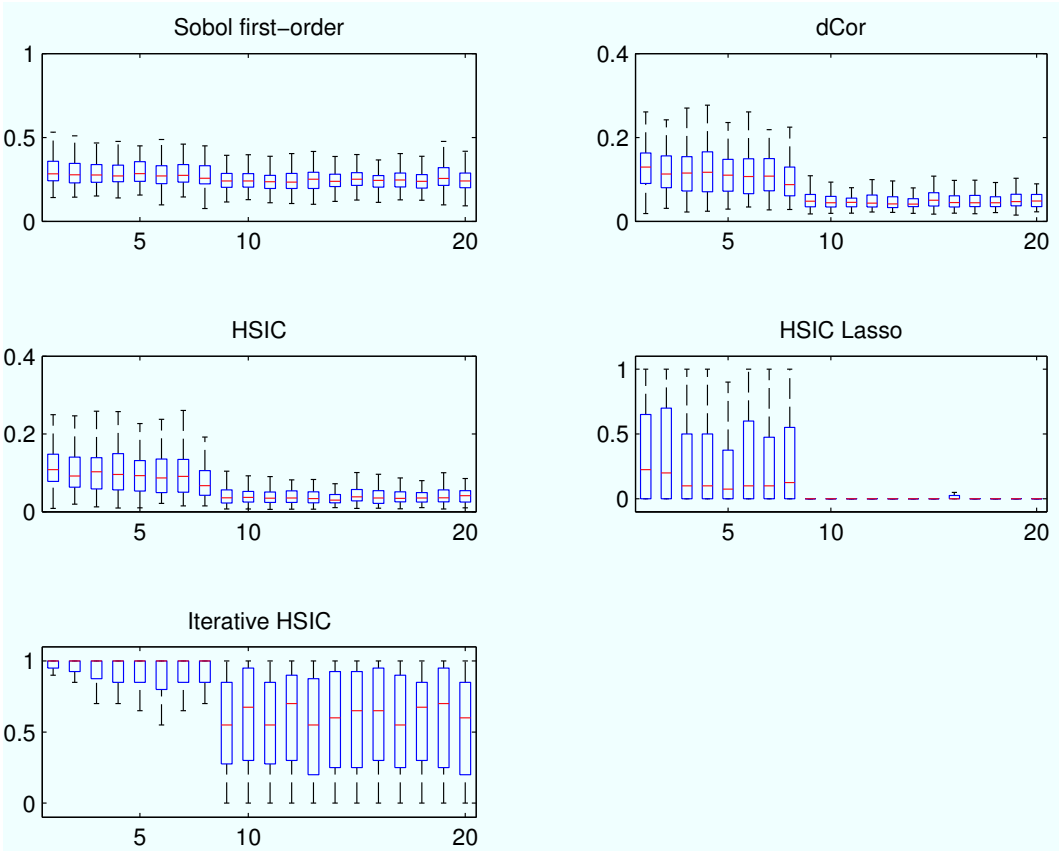

Figure 6: First-order SI, S_X^f, S_X^{dCor}, $S_X^{HSIC_{\mathcal{F},\mathcal{G}}}$ and HSIC Lasso for function η_5, $n = 50$, 100 replicates.

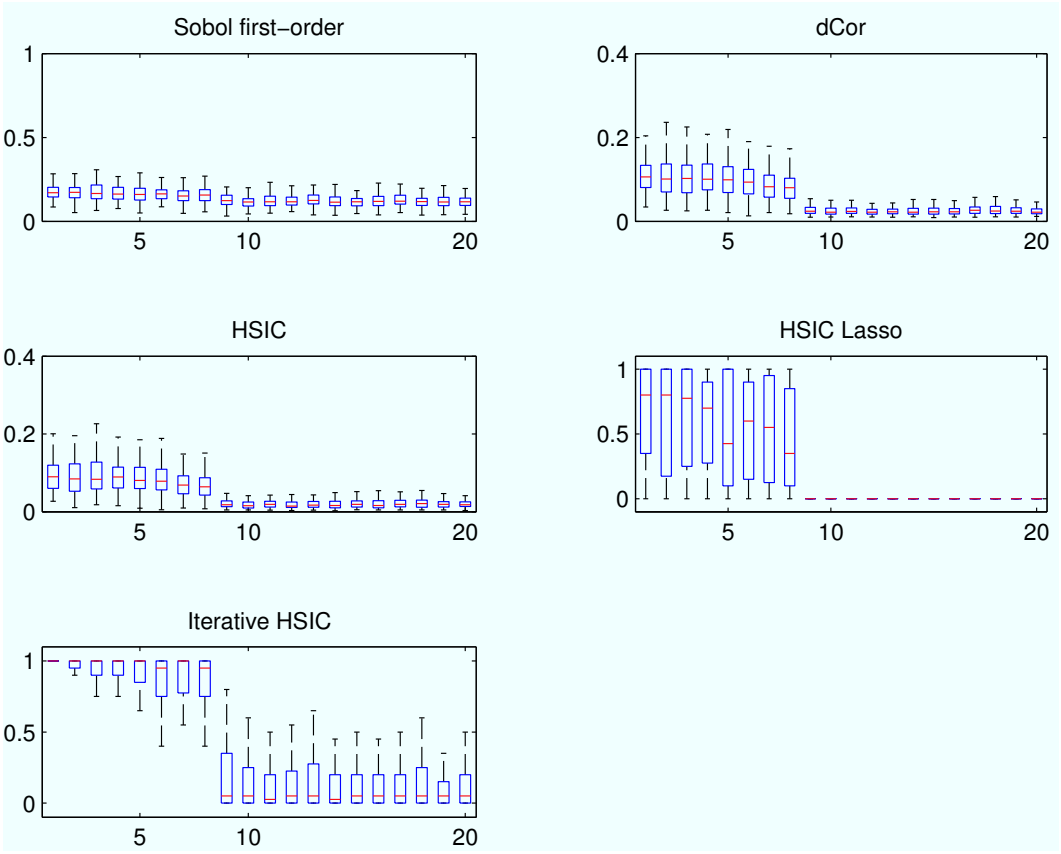

Figure 7: First-order SI, S_X^f, S_X^{dCor}, $S_X^{HSIC_{\mathcal{F},\mathcal{G}}}$ and HSIC Lasso for function η_5, $n = 100$, 100 replicates.

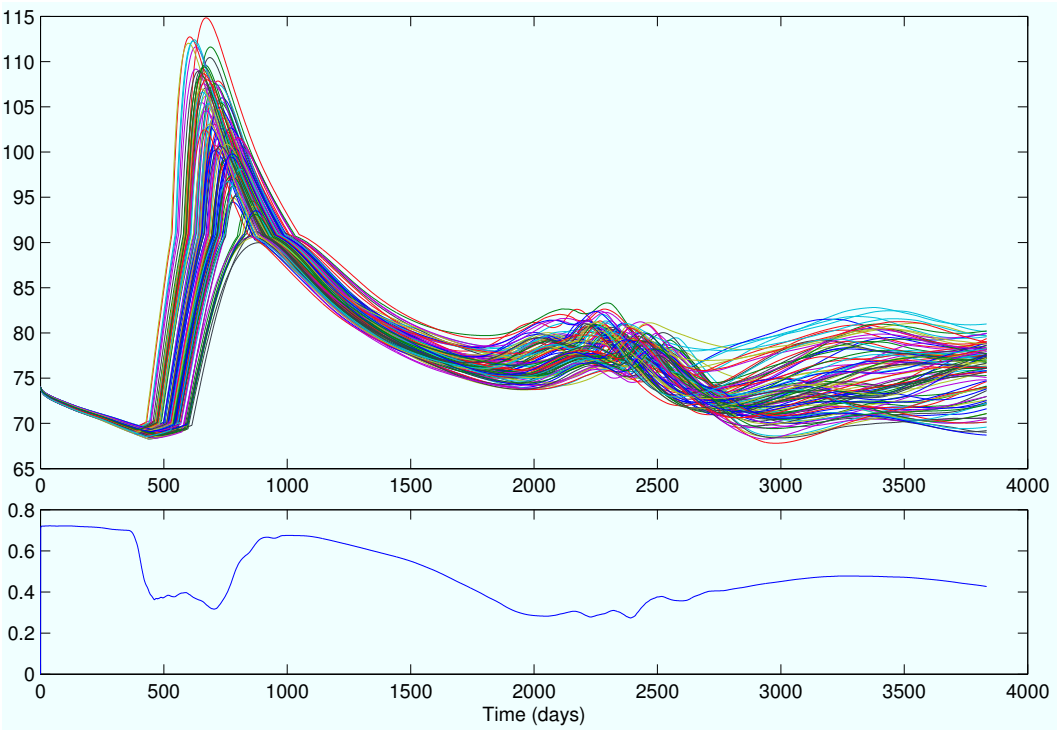

Figure 8: Sample of size $n = 100$ of the collectable data (top) and dCor measure between the parameters and the data at each time step (bottom) for the Punq test case.

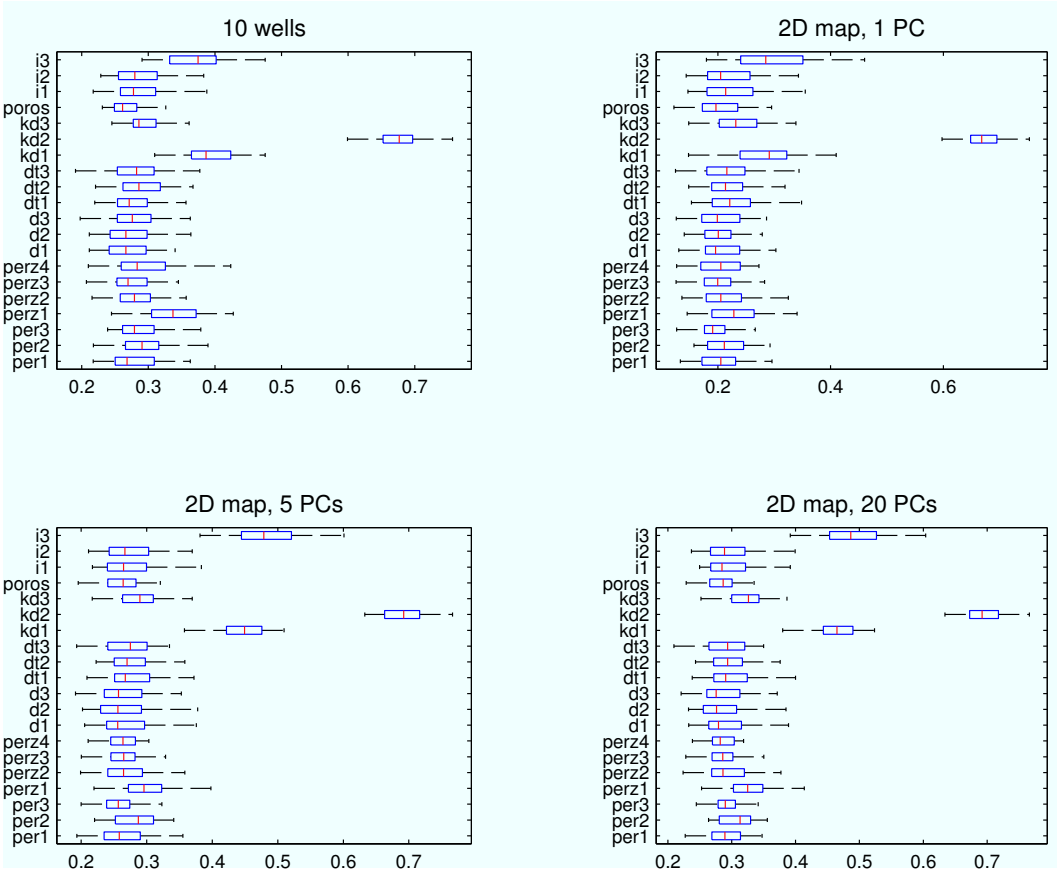

Figure 9: HSIC measure between the parameters and the data for the Marthe test case (10 observation wells and 2D maps with a PCA kernel with 1, 5 and 20 principal components).