Risk Analysis, Vol. 40, No. 12, 2020 DOI: 10.1111/risa.13571

Fighting the Curse of Sparsity: Probabilistic Sensitivity Measures From Cumulative Distribution Functions

Elmar Plischke [1,*] **and Emanuele Borgonovo** [2]

Quantitative models support investigators in several risk analysis applications. The calculation of sensitivity measures is an integral part of this analysis. However, it becomes a computationally challenging task, especially when the number of model inputs is large and the model output is spread over orders of magnitude. We introduce and test a new method for the estimation of global sensitivity measures. The new method relies on the intuition of exploiting the empirical cumulative distribution function of the simulator output. This choice allows the estimators of global sensitivity measures to be based on numbers between 0 and 1, thus fighting the curse of sparsity. For density-based sensitivity measures, we devise an approach based on moving averages that bypasses kernel-density estimation. We compare the new method to approaches for calculating popular risk analysis global sensitivity measures as well as to approaches for computing dependence measures gathering increasing interest in the machine learning and statistics literature (the Hilbert–Schmidt independence criterion and distance covariance). The comparison involves also the number of operations needed to obtain the estimates, an aspect often neglected in global sensitivity studies. We let the estimators undergo several tests, first with the wing-weight test case, then with a computationally challenging code with up to $k = 30,000$ inputs, and finally with the traditional Level E benchmark code.

KEY WORDS: Given-data estimation; global sensitivity analysis; moment-independent measures; variance-based sensitivity measures

1. INTRODUCTION

In risk-informed decision making, risk analysts and decisionmakers increasingly benefit from the use of quantitative risk assessment models (Apostolakis, 2004). Applications range from the probabilistic risk assessments of nuclear waste disposals (Helton, 1994; Helton, Hansen, & Swift, 2014; Helton & Marietta, 2000; Iman, Helton, & Campbell, 1978), of nuclear power plants (Breeding, Helton, Gorham, & Harper, 1992; Helton & Breeding, 1993; Iman & Hora, 1990; NRC, 1990), to food safety (Patil & Frey, 2004), portfolio insurance (Tsanakas & Millossovich, 2016), flood risk (Koks, Bočkarjova, de Moel, & Aerts, 2015), and occupational exposure (Riedmann, Gasic, & Vernez, 2015) studies.

Often the complexity of the problem requires sophisticated modeling efforts with the model becoming a black box. Analysts then cannot rely on the sole intuition for result interpretation and communication. The literature has developed systematic approaches, collectively named sensitivity analysis, to allow analysts to extract additional insights from risk assessment models, thus increasing transparency and favoring interpretability (see Helton, 1993; Helton & Davis, 2002; Iman, Johnson, & Watson, 2005; Saltelli, 2002b; Saltelli, Tarantola, & Chan, 1998), (EPA, 2009, Appendix D).

[1]TU Clausthal, Clausthal-Zellerfeld, Germany.
[2]Bocconi University, Milan, Italy.
[*]Address correspondence to Elmar Plischke, TU Clausthal, Clausthal-Zellerfeld, Germany; elmar.plischke@tu-clausthal.de.

An important piece of information, often sought in risk assessment, is the importance of the uncertain inputs. This information allows the analyst to identify areas where additional modeling efforts are needed and to prioritize the collection of further data and information. Using the terminology of Saltelli (2002b), we are in a *factor prioritization setting*. An analyst obtains this information using either local, screening, or global sensitivity analysis methods. Local techniques comprise methods such as Tornado Diagrams (Eschenbach, 1992) and partial derivatives (Helton, 1993). Employing a local method, the analyst identifies the key drivers of variability around a point or in a limited region of the parameter space. Screening methods comprise techniques such as the method of Morris (Morris, 1991), sequential bifurcation (Bettonvil & Kleijnen, 1997; Kleijnen, 2017), and permutation importance (we refer to Wei, Lu, & Song, 2015, for greater details). Employing a screening method, the analyst aims to identify the least relevant inputs and to provide a qualitative indication of the most important ones, with a low number of simulator evaluations. Global methods comprise techniques such as variance-based (Saltelli & Tarantola, 2002), moment-independent sensitivity indices (Borgonovo, 2006), and value of information (Felli & Hazen, 1998; Oakley, 2010). Employing a global sensitivity method, the analyst aims at identifying the key drivers of uncertainty quantitatively, while thoroughly exploring the simulator input space.

While best practices recommend the use of global sensitivity methods in the presence of uncertainty (Helton, 1993; Helton & Davis, 2002; Patil & Frey, 2004), the estimation of global sensitivity measures can become a challenging task. Past literature has identified and introduced methods to fight the *curse of dimensionality*, that affects estimation when simulators have a large number of inputs. Sampling-based methods—in the terminology of Helton and Davis (2002)—or one-sample approaches in the terminology of Strong, Oakley, and Chilcott (2012) and Strong and Oakley (2013) are estimation methods that reduce the relevance of dimensionality.

However, the curse of dimensionality might not be the only computational challenge. Numerical values of the output spanning over orders of magnitude may impair convergence at reasonable sample sizes, independently of the dimensionality of the model. We call this effect the curse of sparsity. Although not explicitly isolated from the curse of dimensionality, this issue has been recognized early on in the risk analysis literature. *The scaling problem most often can be overcome by performing uncertainty importance calculations based on a logarithmic scale for the top-event frequencies. The log scale produces a reliable ordering of the uncertainty importance for the events, and expresses the results in terms of log-based risk. However, the log-based uncertainty importance calculations do not readily translate back to a linear scale* (Iman & Hora, 1990, p. 402). These observations evidence two facts. On the one hand, transformations can help reducing the effects of the curse of sparsity. On the other hand, transformations induce interpretation issues, because results are valid on the transformed scale and not on the original scale. The use of transformations has been popular in the risk analysis literature since seminal works such as Iman and Conover (1979) and Saltelli and Sobol' (1995). However, recently it has been noted that the interpretation problems might be overcome if the analyst employs a sensitivity measure which is transformation-invariant (Borgonovo, Tarantola, Plischke, & Morris, 2014). Nonetheless, transformation invariance *per se* is not sufficient to overcome the curse of sparsity. In fact, some transformation-invariant global sensitivity measures (for instance, sensitivity measures based on the Kullback–Leibler entropy or on the L^1-norm between densities) require the estimation of a probability density function (pdf). Density estimation often relies on kernel smoothing whose numerical performance is affected by sparsity. Thus, convergence might be impaired even if the global sensitivity measure under estimation is transformation-invariant.

We propose and evaluate a new method for computing global sensitivity measures that builds on this literature. The new method is based on the intuition of rewriting estimators as function of the empirical (marginal and conditional) distribution function of the model output. In this way, the numerical elaborations are restricted to real numbers in the [0,1] interval, contrasting the curse of sparsity. The estimation is within a given-data (one-sample) framework, thus keeping the estimation cost independent of the number of model inputs.

Before considering the numerical aspects of the new method, we establish the relationship among the L^1-norm between pdfs, the Kolmogorov–Smirnov and the Kuiper metrics. These three metrics are the basis of three global sensitivity measures used in previous risk assessment studies, namely, the measures δ, β^{KS}, and β^{Ku} (Borgonovo, 2006; Wei et al., 2015). We find that if the marginal and all conditional model output distributions are unimodal, the Kuiper

and L^1-norm distances are equivalent. Then, the analyst can bypass density estimation using directly a sensitivity measure based on cumulative distribution functions (cdfs). However, these conditions on the marginal and conditional distributions are not verified in general. We therefore adapt the new method to the estimation of density-based sensitivity measures. The intuition is to employ a moving average to replace kernel smoothing. We study the resulting estimators from two viewpoints. First, we provide a discussion on their consistency in Appendix C. Second, we address their algorithmic properties. In fact, besides the global importance measures discussed above, other dependence measures coming from machine learning are becoming of interest in risk assessment studies—in particular, distance covariance (Székely & Rizzo, 2005) and the Hilbert–Schmidt independence criterion (HSIC) (Da Veiga, 2015; De Lozzo & Marrel, 2016; De Lozzo & Marrel, 2017). These importance measures can be computed from given data and thus, nominally, at the same number of model runs as the method we are studying. The difference is, however, in the number of operations (algorithmic cost) performed once the input–output sample is available. We then study the algorithmic cost of the present method and compare it with the algorithmic cost of the above-mentioned dependence measures.

A series of challenging numerical experiments is carried out. The first experiments involve the simulator developed to study the weight of a wing of a light aircraft and recently studied in Jiménez Rugama and Gilquin (2018). The simulator is smooth, fast to run, and of low-dimensionality: all estimators work well. Then, we consider a simulator in which the inputs and output are Cauchy distributed and the curse of sparsity appears. This synthetic case has the same behavior as the Level E geosphere transport code, the reference code for sensitivity analysis, which is investigated next. Level E features an output spanning orders of magnitude. There is no curse of dimensionality, but the curse of sparsity starts playing a role. The last experiments involve a computationally demanding albeit analytically known model. The goal is to identify the 10 most important inputs out of 30,000 input parameters with a parameterization that makes the output variance close to the numerical range of the floating-point representation (which is $1.79 \cdot 10^{308}$). This test case combines sparsity with dimensionality. For all experiments, we test the new approach also against previous methods as well as against estimators of the above-mentioned dependence measures, performing the analysis with and without output transformation to investigate whether rescaling is essential to reach convergence. In all cases, the results show that the newly proposed estimators achieve convergence at reasonable sample sizes.

2. A CONCISE REVIEW ON GLOBAL SENSITIVITY MEASURES

This section provides a concise review on global sensitivity analysis. The literature is vaster than what can be exposed here. For broad overviews, we refer the reader to the monographs of Saltelli, Ratto, Tarantola, and Campolongo (2012) and Borgonovo (2017) and to the *Handbook of Uncertainty Quantification* (Ghanem, Higdon, & Owhadi, 2017) for comprehensive discussions. Among global methods, in risk analysis regression-based methods have been among the first to be applied for factor prioritization (Helton, 1993; Saltelli & Marivoet, 1990). Reviews are offered in Helton, Johnson, Sallaberry, and Storlie (2006); Storlie and Helton (2008); and Storlie, Swiler, Helton, and Sallaberry (2009). Works such as Helton (1993, 1994, 1999); Helton and Davis (2002, 2003); Frey and Patil (2002); Helton and Sallaberry (2009); and Mohanty et al. (2011) represent outstanding examples of their several applications.

The main sensitivity indicators of nonparametric regression methods are the standardized regression coefficients and Pearson's correlation coefficient. To introduce them, let

$$g : \mathcal{X} \to \mathbb{R}, \qquad \mathbf{X} = (X_1, \ldots, X_k) \mapsto Y = g(\mathbf{X}) \quad \text{and} \quad \mathcal{X} \subseteq \mathbb{R}^k \tag{1}$$

denote the input–output mapping where k is the number of input factors. $\mathbf{X}$ is a random vector on a probability space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mathbb{P})$, where $\mathcal{B}(\mathcal{X})$ is the Borel σ-algebra and $\mathbb{P}$ the probability measure that reflects the analyst's state of knowledge about the factors. The model output becomes a random variable Y whose distribution is induced by $g(\cdot)$. The corresponding probability space is $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), \mathbb{P}_Y)$. We denote by $F_Y(y)$ and $f_Y(y)$ the model output marginal cdf and density, respectively. Conditional distributions, cdfs, and densities are denoted by $\mathbb{P}_{Y|X_i=x_i}$, $F_{Y|X_i=x_i}(y)$, and $f_{Y|X_i=x_i}(y)$, respectively.

If the input–output mapping g can be accurately fitted by a linear regression model, i.e., $g(Y) \approx b_0 + \sum_{i=1}^{k} b_k X_k$ then natural sensitivity measures are the standardized regression coefficients

Table I. Sensitivity Measures Used in This Work

Measure	Symbol	Definition	Estimation
Linear regression based (SRC, PCC)	ϱ	(2)	
Variance based	η	(3)	(11)
Borgonovo, L^1 (pdf)	δ	(7)	(18)
Kolmogorov–Smirnov, L^∞ (cdf)	β^{KS}	(8)	(16)
Kuiper, range (cdf)	β^{Ku}	(9)	(16)
Distance correlation (char. function)	$d\varrho$	(D2)	(D3)
Hilbert–Schmidt independence criterion	HSIC		(D4)

SRC_i (Helton, 1993; Kleijnen & Helton, 1999a, 1999b), $\mathrm{SRC}_i = b_i \frac{\sigma_i}{\sigma_Y}$, where b_i are the linear regression coefficients, σ_i are the standard deviations of X_i, and σ_Y is the standard deviation of the model output Y. Under independence of all random inputs X_i, the standardized regression coefficients coincide with the Pearson product moment correlation coefficients (Pearson, 1901)

$$\varrho_i = \frac{\mathrm{cov}(Y, X_i)}{\sigma_i \sigma_Y} \tag{2}$$

where $\mathrm{cov}(Y, X_i)$ is the covariance between Y and X_i (Table I lists the sensitivity measures used in this work).

The quantities ϱ_i and SRC_i are usually interpreted as measures of the strength of the linear relationship between two random variables. However, when the linear regression fit is poor, the explanatory power of a linear model assumption is weak and confidence in the ranking induced by regression-based techniques diminishes (Campolongo & Saltelli, 1997). Two remedies are present. The first is to resort to rank transformations (Conover & Iman, 1976; Iman & Conover, 1979). The standardized rank regression coefficients and the Spearman regression coefficient (Spearman, 1904) then become natural global sensitivity measures. The second remedy is to apply sensitivity methods that rely less on the regression fit.

Researchers have successfully investigated the use variance-based sensitivity measures (Iman & Hora, 1990; Saltelli, 2002a). One writes

$$\eta_i = \frac{\mathbb{V}\{\mathbb{E}[Y|X_i]\}}{\mathbb{V}[Y]} = 1 - \frac{\mathbb{E}\{\mathbb{V}[Y|X_i]\}}{\mathbb{V}[Y]}. \tag{3}$$

Here $\mathbb{V}[Y]$ is the output variance, $\mathbb{E}[Y|X_i]$ and $\mathbb{V}[Y|X_i]$ are the conditional output expectation and the conditional output variance given the input X_i. The conditional expectation given X_i is the nonparametric regression curve of Y on X_i. The quantity in (3) is the expected reduction in model output variance achieved by learning the true value of X_i. This quantity coincides with the Pearson correlation ratio (Pearson, 1905) and with the first order variance-based sensitivity index (Homma & Saltelli, 1996). Several strategies have been developed over the years to efficiently estimate variance-based sensitivity measures. Among others, we recall the designs based on Fourier Amplitude Sensitivity Test (FAST) (Saltelli, Tarantola, & Chan, 1999) and on Polynomial Chaos Expansion (Le Gratiet, Marelli, & Sudret, 2017; Sudret, 2008).

Recently, Borgonovo, Hazen, and Plischke (2016) show that several global sensitivity measures can be defined through a common rationale. Consider an operator $\zeta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ of the form $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_i=x_i})$, where $\mathcal{P}$ is the set of all distributions on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and $\zeta(\cdot,\cdot)$ is a generalized form of distance (thus, a metric or a divergence) between two distributions in the sense of Glick (1975). $\zeta(\cdot,\cdot)$ is called an inner operator. For consistency, it is assumed that $\zeta(\mathbb{P}, \mathbb{P}) = 0$ for any $\mathbb{P} \in \mathcal{P}$. The value of $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_i=x_i})$ depends on the value attained by X_i. Therefore, the distance $\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_i})$ is a random function of X_i. Taking the expectation with respect to the marginal distribution of X_i, we obtain the quantity

$$\xi_i = \mathbb{E}[\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_i})]. \tag{4}$$

This quantity is the global sensitivity measure of X_i based on inner operator $\zeta(\cdot,\cdot)$. Several probabilistic sensitivity measures used in risk analysis are encompassed by (4). For instance, by setting

$$\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_i}) = \frac{1}{\mathbb{V}[Y]}\left[\int_y y(f_{Y|X_i}(y) - f_Y(y))dy\right]^2, \tag{5}$$

and averaging we obtain the first-order variance-based sensitivity measure η_i of (3). Similarly, setting

$$\zeta(\mathbb{P}_Y, \mathbb{P}_{Y|X_i}) = \frac{1}{2}\int_y \left|f_{Y|X_i}(y) - f_Y(y)\right| dy \tag{6}$$

and averaging, we obtain the δ-importance measure

$$\delta_i = \frac{1}{2}\mathbb{E}\left[\int_{\mathcal{Y}} \left| f_{Y|X_i}(y) - f_Y(y)\right| dy\right]. \quad (7)$$

In addition, the global sensitivity measures

$$\beta_i^{\mathrm{KS}} = \mathbb{E}[\sup_y \left| F_{Y|X_i}(y) - F_Y(y)\right|] \quad (8)$$

and

$$\beta_i^{\mathrm{Ku}} = \mathbb{E}\left[\sup_y \left(F_{Y|X_i}(y) - F_Y(y)\right) - \inf_y \left(F_{Y|X_i}(y) - F_Y(y)\right)\right] \quad (9)$$

are the expected separations between the conditional and unconditional model output cdfs obtained, respectively, using the Kolmogorov–Smirnov and the Kuiper distances (Borgonovo et al., 2014).

The three distance-based sensitivity measures (δ_i, β_i^{KS}, β_i^{Ku}) share the following properties:

(1) Normalization: δ_i, β_i^{KS}, $\beta_i^{\mathrm{Ku}} \in [0, 1]$;
(2) Nullity implies independence: $\delta_i = 0$, $\beta_i^{\mathrm{KS}} = 0$, or $\beta_i^{\mathrm{Ku}} = 0$ imply that X_i and Y are independent;
(3) Monotonic transformation invariance:

$$\delta_i(z(Y)) = \delta_i(Y),\ \beta_i^{\mathrm{KS}}(z(Y)) = \beta_i^{\mathrm{KS}}(Y),$$
$$\beta_i^{\mathrm{Ku}}(z(Y)) = \beta_i^{\mathrm{Ku}}(Y), \quad (10)$$

where $z : \mathcal{Y} \to \mathbb{R}$ is a monotonic function of the model output Y.

Nullity implies independence is also listed as number 5 in Rényi's axioms for measures of statistical dependence (Rényi, 1959). It allows an analyst to conclude that a null value of the sensitivity measure implies that X_i and Y are independent random variables (Da Veiga, 2015). The sensitivity measures SRC_i, ϱ_i, and η_i do not possess the nullity-implies-independence property.

Monotonic transformation invariance helps analysts in fighting the curse of sparsity. As shown in Conover and Iman (1981); Iman and Hora (1990); Saltelli and Sobol' (1995), transformations may improve numerical efficiency in the estimation of global sensitivity measures. However, transformations open the question of reinterpreting results back on the original scale (Iman & Hora, 1990). Moreover, transformations may induce ranking changes consequent to the change in the input–output structure. If rank modifications occur then the analyst would need to make a choice on whether to trust the ranking before or after the transformation. Then, it becomes important to understand whether the transformed or the original output is the quantity of interest for the model user. Conversely, these problems are avoided if the sensitivity measure is transformation-invariant.

Recently, further ways of identifying key drivers of uncertainty have been explored in machine learning (Da Veiga, 2015). A first method relies on distance covariance and distance correlation (Székely, Rizzo, & Bakirov, 2007). We refer to Lyons (2013) and Sejdinovic, Sriperumbudur, Gretton, and Fukumizu (2013) for further readings on theoretical aspects underlying these dependence measures. In Appendix D, we report some additional mathematical details useful to clarify the calculations carried out in the subsequent numerical experiments of our work. Distance covariance quantifies the degree of statistical dependence between Y and X_i via pairwise distances of their realization. In particular, one considers the random variables X_i and Y and their independent replicates X_i', X_i'' and Y', Y''. Then, their distance covariance is calculated from the expression (Lyons, 2013; Sejdinovic et al., 2013):

$$\mathcal{V}^2(Y, X_i) = \mathbb{E}\big[\left|X - X'\right| \cdot \left|Y - Y'\right|\big] + \mathbb{E}\big[\left|X - X'\right|\big] \cdot \mathbb{E}\big[\left|Y - Y''\right|\big] - 2\mathbb{E}\big[\left|X - X'\right| \cdot \left|Y - Y''\right|\big].$$

In this work, we will use the normalized version of distance covariance, which is known as distance correlation—see Appendix D for details. Distance covariance and distance correlation are part of the so-called energy statistics (Székely & Rizzo, 2017), and are a topical research subject.

The second method relies on the HSIC as a sensitivity measure. Lyons (2013); Sejdinovic et al. (2013); Da Veiga (2015) show that this criterion is closely related to distance correlation. In particular, Sejdinovic et al. (2013) prove that when an Euclidean distance is used in HSIC, the square of distance correlation and HSIC are proportional. However, with the traditional use of a Gaussian kernel, HSIC, and distance correlation are not generally equivalent sensitivity measures.

3. ESTIMATION: THE GIVEN-DATA APPROACH

The total cost for estimating a probabilistic sensitivity measure is given by the sum of two components: the cost associated with the generation of the sample (Γ_{Model}) and the cost associated with the calculation of the global sensitivity measure from the sample ($\Gamma_{\mathrm{Estimator}}$). The first component is strictly related to the ability to run the model and is measured in terms of model runs. The second component is

related to the number of operations required by the estimator and is measured in number of computer operations. When the running time of the model is high, Γ_{Model} is usually dominating. However, if two sensitivity measures can be estimated from the same sample, then the lower $\Gamma_{\text{Estimator}}$ the faster is the analysis. In this section, we shall focus on the first component, Γ_{Model}. The algorithmic cost, $\Gamma_{\text{Estimator}}$, is discussed at the end of Section 4.2.

The global sensitivity measures comprised by the common rationale in (4) are associated with a cost $\Gamma_{\text{Model}}^{\text{Brute-Force}} = k \cdot n_{ext} \cdot n_{int}$ where n_{ext} points are sampled from the marginal distribution of X_i, $i = 1, \ldots, k$, and n_{int} points are required to compute the inner statistic conditional to the n_{ext} realizations of X_i.

The required number of model evaluations becomes rapidly prohibitive and several works have addressed ways of reducing computational burden. For variance-based sensitivity measures, the works (Homma & Saltelli, 1996; Saltelli, 2002a; Saltelli et al., 2010; Sobol', 1990) have reduced the computational cost to $\Gamma_{\text{Model}} = (k+2)n$, i.e., $k+2$ evaluations of a basic sample block, for estimating first and total effects.[1] These works use the Jansen–Saltelli–Sobol' design, known also as pick-and-freeze sampling (Gamboa, Janon, Klein, Lagnoux, & Prieur, 2016). Recently, Owen (2013) introduces a specific design that improves the estimation of variance-based sensitivity measures when their values are small. The random balance design (Tarantola, Gatelli, & Mara, 2006), a variant of the FAST method (Cukier, Fortuin, Shuler, Petschek, & Schaibly, 1973), requires $C = n$ evaluations for computing first-order effects. All these approaches are specific and require that the sample is generated from the computer code following a precise scheme.

A *given-data* or one-sample approach, instead, allows the computation global sensitivity measures directly from a sample (X, Y), with $\Gamma_{\text{Model}}^{\text{Given-Data}} = n$, where n is the sample size. The key intuition lies in replacing the point-conditional probability $\mathbb{P}_{Y|X_i=x_i}$ with the class-conditional probability $\mathbb{P}_{Y|X_i\in\mathcal{C}_{m,i}}$, where $\mathcal{C}_{m,i}$ is an element of a partition of the support of X_i. (We recall that, in general, the partition of a set $\mathcal{X}_i$ is a collection of sets such that $\mathcal{X}_i = \bigcup_{m=1}^{M} \mathcal{C}_{m,i}$, $\mathcal{C}_{m,i} \cap \mathcal{C}_{m',i} = \emptyset$, $m \neq m'$, $m = 1, 2, \ldots, M$.) Specifically, in our case, one considers the realizations of

[1]Total effects are given by $\tau_i = 1 - \mathbb{V}[Y]^{-1}\mathbb{E}[\mathbb{V}[Y|X_{\sim i}]]$, and represent the expected reduction in output variance when all factors except X_i are fixed.

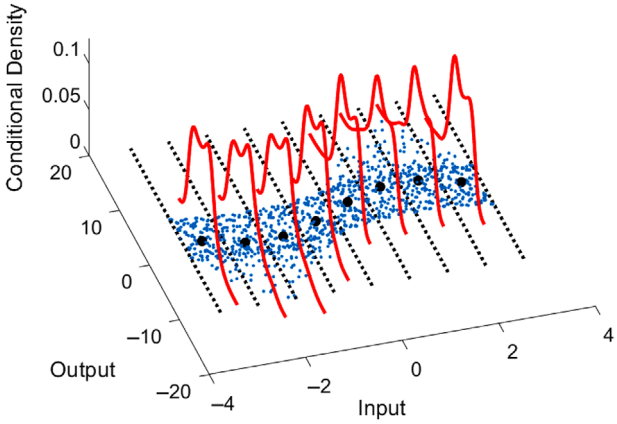

Fig 1. Conditioning the data via binning: Conditional means (black dots) are used for estimating first-order effects, conditional densities for density-based importance measures.

the pair (X_i, Y). Then, one creates the scatterplot of Y against X_i. (Fig. 1 offers a three-dimensional view of the scatterplot of n hypothetical simulator in which X_i and Y are absolutely continuous. The values of X_i are on the x-axis, the realizations of Y on the y-axis and the z-axis plots the empirical conditional densities of Y given values of X_i in a given bin.) By sorting the values of X_i, one then attributes the realizations of X_i to M partition classes $\mathcal{C}_{m,i}$, $m = 1, 2, \ldots, M$. To illustrate, in Fig. 1, the support of X_i is the interval $[-3.14, 3.14]$ and we have $M = 8$, with $\mathcal{C}_{1,i} = [-3.14, -2.36)$, $\mathcal{C}_{2,i} = [-2.36, -1.57), \ldots,$ $\mathcal{C}_{8,i} = [2.36, 3.14]$ (The eight corresponding bins are separated by dotted lines in Fig. 1). A bin contains all realizations of Y that correspond to realizations of X_i in partition $\mathcal{C}_{m,i}$. We denote this set of realizations by $Y_{m,i} = \{y_j\,;\, x_{ji} \in \mathcal{C}_{m,i}\}$. The number of realizations of Y in a bin is $n_{m,i}$. Then, the local mean $\bar{y}_{m,i} = n_{m,i}^{-1} \sum_{y\in Y_{m,i}} y$ is an estimate of the conditional expectation $\mathbb{E}[Y|X_i \in \mathcal{C}_{m,i}]$ (illustrated by black dots at the center of each bin in Fig. 1). The global mean $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ is an estimate of the unconditional mean value of the model output, $\mathbb{E}[Y]$. Then, the ratio

$$\hat{\eta}_i = \frac{\sum_{m=1}^{M} \frac{n_{m,i}}{n} (\bar{y}_{m,i} - \bar{y})^2}{\frac{1}{n-1}\sum_{j=1}^{n} (y_j - \bar{y})^2} \tag{11}$$

is Pearson's 1905 given-data estimator of η_i in Equation (3), where the weights $\frac{n_{m,i}}{n}$ are estimates of the probability of $\mathcal{C}_{m,i}$ under X_i.

More in general, points in a bin follow the conditional distribution $F_{Y|X_i\in\mathcal{C}_{m,i}}$. If Y is absolutely continuous, we have the corresponding conditional densities $f_{Y|X_i\in\mathcal{C}_{m,i}}$. These densities are visualized in

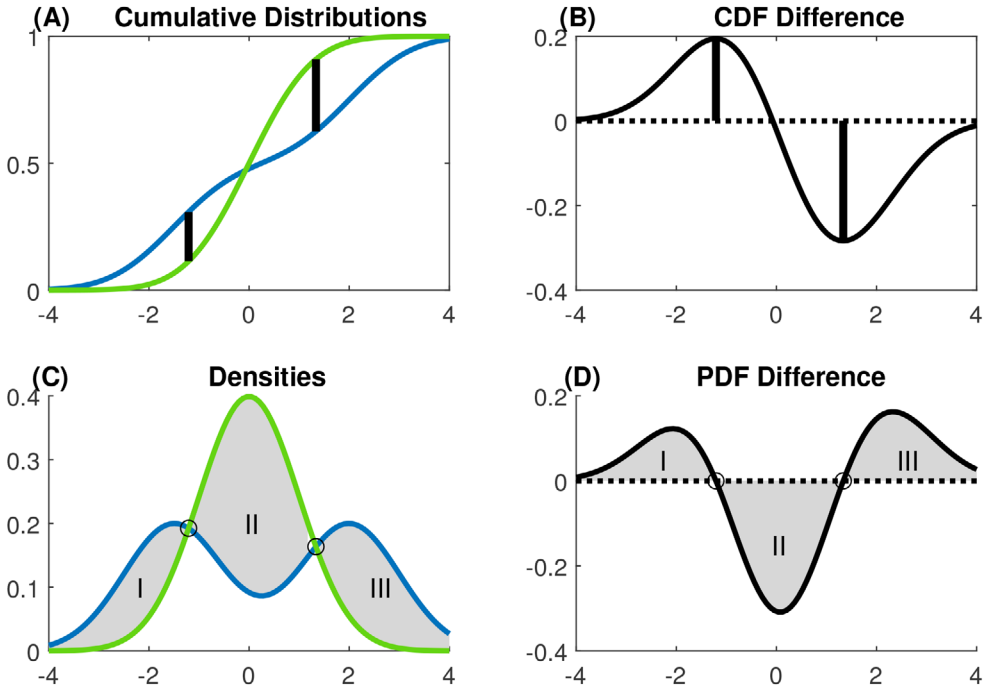

Fig 2. Differences of probability densities and cumulative distributions.

the red lines of Fig. 1. Thus, from the realizations in the bins it is possible to obtain empirical estimates of the cdf $F_{Y|X_i\in\mathcal{C}_{m,i}}$ or of the corresponding density $f_{Y|X_i\in\mathcal{C}_{m,i}}$. This partitioning idea carries over directly to other measures in the common rationale. To illustrate, for the δ-importance in (7) the estimator takes the form

$$\widehat{\delta}_i = \sum_{m=1}^{M} \frac{n_{m,i}}{n} \zeta_i(\widehat{f}_Y, \widehat{f}_{Y|X_i\in\mathcal{C}_{m,i}}), \tag{12}$$

where M is the number of partitions depending on the sample size n and $\widehat{f}_Y$, $\widehat{f}_{Y|X_i\in\mathcal{C}_{m,i}}$ are empirical estimation of the unconditional and conditional simulator output densities.

In Equation (12), a crucial role is played by the estimator of the conditional and unconditional densities. If kernel smoothing is chosen, this choice in itself involves the selection of kernel functions and bandwidths—please see Appendix E for greater details. It is well known (Sheather, 2004) that the values of the kernel-density bandwidths are of critical importance and optimal choices are related to the roughness of the (unknown) pdfs f_Y and $f_{Y|C_{m,i}}$. When the model output is sparse, this part of the estimator may fail in producing reasonable results at limited sample sizes. Then, bypassing density estimation might be advantageous.

4. ESTIMATION USING EMPIRICAL CDFS

This section is divided into two parts. In the first part, we state some results that help us in reformulating estimators of global sensitivity measures based on empirical cdfs. In the second, we introduce the new method.

4.1. A Preliminary Result

In Fig. 2, a bimodal distribution (dark/red) is compared with a unimodal one (light/green). The Kolmogorov–Smirnov measure is determined by the maximum distance between two cdfs. This distance equals the length of the larger of the vertical black bars in graphs (A) or (B). The Kuiper distance is determined by considering the maximum difference in both directions. Hence the sensitivity measure is the sum of the two bars. However, we observe that the separations in both the Kuiper and Kolmogorov–Smirnov distances can be interpreted also in terms of the areas between the corresponding pdfs. This is illustrated in graphs (C) and (D) of Fig. 2. In graph (C), regions I and III are bounded above by the red

pdf and below by the green pdf. Region II is bounded by the green pdf above and the red pdf below. Graph (D) shows the pdf difference, regions I and III are above the 0-line, while region II is below. In this example, the Kolmogorov–Smirnov distance is equal to the area of region III. The Kuiper distance is the area of region II or, alternatively, the combined area of regions I and III. The inner operator of the δ sensitivity measure (L^1-norm, (6)) equals half the sum of the areas of the three regions. It is also worth observing that the locations of the extremes of the difference between the cdfs coincide with the locations of intersections of the pdfs. While the Kolmogorov-Smirnov and Kuiper metrics refer to the global extreme value of the difference between the conditional and unconditional cdfs, the L^1-norm also takes all local extreme values into account (Davies & Kovac, 2004). Extending these insights to the corresponding global sensitivity measures, we then have the following proposition—see Appendix A for the proof.

Proposition 1. *For any random variables X_i and Y the sensitivity measures $\beta_i^{KS} \leq \beta_i^{Ku} \leq 1$, and $\beta_i^{Ku} \leq 2\beta_i^{KS}$. If $g(\mathbf{x})$ is a map depending only on x_i, i.e., there exists a function $g_i(x_i) = g(\mathbf{x})$ then $\beta_i^{Ku} = \delta_i = 1$ and $\beta_i^{KS} \geq 0.5$. If this functional relation is monotonic then $\beta_i^{KS} = 0.75$. Under absolute continuity of Y and $Y|X_i$, $\beta_i^{KS} \leq \beta_i^{Ku} \leq \delta_i \leq 1$. If for almost all values of X_i the difference $(f_{Y|X_i} - f_Y)(y)$ has at most two inner zeros then $\beta_i^{Ku} = \delta_i$. If this difference has exactly one inner zero then $\beta_i^{KS} = \beta_i^{Ku}$.*

Besides formalizing the relationship between global sensitivity measures based on the L^1-norm, on the Kolmogorov–Smirnov metric and on the Kuiper metrics, Proposition 1 shows that sensitivity measures based on the Kuiper distance and on the L^1-norm are equivalent when all the involved model output distributions are unimodal. Thus, if δ_i and β_i^{Ku} are different, the analyst infers that not all conditional and unconditional model output distribution involved in the analysis are unimodal. On the other hand, if the analyst knows that all involved distributions are unimodal, then β_i^{Ku} becomes a substitute for δ_i. Previous numerical experiments carried out in the literature, show that β_i^{Ku} is easier to estimate than δ_i (Borgonovo et al., 2014). Thus, Proposition 1 can be turned into a computational advantage: one estimates β_i^{Ku} instead of δ_i if all distributions are unimodal. However, in general the analyst may not have *a priori* knowledge that conditional and marginal distributions are all unimodal (for instance, this is not the case for the densities in Fig. 1). Moreover, the reason of the computational advantage associated with β_i^{Ku} has not been fully investigated. We argue that the better convergence properties are due to the fact that β_i^{Ku} works directly on the cdf, thus embedding a convenient numerical transformation of the model output data. We investigate these aspects further in the next subsection.

4.2. CDF Estimators

The generic given-data estimator of any global sensitivity measure comprised within the framework of (4) can be expressed as a function of empirical cdfs as follows (Borgonovo et al., 2016):[2]

$$\widehat{\xi_i} = \sum_{m=1}^{M(n)} \frac{n_{m,i}}{n} \zeta(\widehat{F_Y}(y), \widehat{F}_{Y|X_i \in \mathcal{C}_{m,i}}(y)), \tag{13}$$

where the number of partitions $M(n)$ is dependent on the sample size n. The assumptions under which the estimators in Equation (13) are consistent are discussed in Borgonovo et al. (2016) and summarized in theorem 1 therein. Specifically, it is required that: (i) $M(n)$ is monotonically increasing in n and such that $\lim_{n\to\infty} \frac{n}{M(n)} = \infty$, and (ii) the inner statistic $\zeta(\cdot,\cdot)$ is a continuous function of its inputs, and (iii) $\zeta(\cdot,\cdot)$ is Riemann–Stieltjes integrable with respect to the marginal distribution of X_i. In the proof, a key role is played by the fact that $\widehat{F_Y}$ and $\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}$ are consistent, that is, by the fact that $\widehat{F_Y}$ and $\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}$ tend pointwise to F_Y and $F_{Y|X_i\in\mathcal{C}_{m,i}}$ as the sample size n and therefore the size of the partition $M(n)$ increases.

Then, the first step for calculating any estimator in the form of (13) is to obtain consistent cdf estimators. A canonical candidate for this purpose is the empirical cdf of Y. Given n realizations (y_j), $j = 1, 2, \ldots, n$ of the random variable Y, the empirical cdf of Y is defined by counting the number of realizations below y,

$$\widehat{F_Y}(y) = \frac{1}{n}\#\{y_j | y_j \leq y\}, \tag{14}$$

where $\#\mathcal{A}$ denotes the number of elements in the set $\mathcal{A}$. For the conditional random variable $Y|X_i \in \mathcal{C}_{m,i}$, we obtain the subsample $Y_{m,i} = \{y_j | x_{ji} \in \mathcal{C}_{m,i}\}$ containing $n_{m,i}$ realizations of outputs for which the associated input value of interest is contained in the mth

[2]For simplicity, we limit the discussion to continuous random model inputs. However, the result applies to discrete X_i as well (Borgonovo et al., 2016).

class of the partition. Then, the conditional empirical cdf of Y given that $X_i \in \mathcal{C}_{m,i}$ is defined by

$$\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y) = \frac{1}{n_{m,i}}\#\{y_j \in Y_{m,i} | y_j \le y\}. \tag{15}$$

Setting $\mathcal{C}_{m,i} = [F_{X_i}^{-1}(\frac{m-1}{M}), F_{X_i}^{-1}(\frac{m}{M})]$ we have classes or bin intervals that all satisfy $\frac{n_{m,i}}{n} \approx \frac{1}{M(n)}$, i.e., each strip contributes the same weight in (13). With this choice of the partition the estimator in Equation (13) becomes

$$\widehat{\xi_i} = \frac{1}{M(n)}\sum_{m=1}^{M(n)} \zeta\left(\widehat{F}_Y, \widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}\right), \tag{16}$$

which we call cdf-based estimator. As shown in the proof of theorem 1 in Borgonovo et al. (2016), the canonical empirical-cdf estimators are consistent by the law of large numbers and thus the cdf-based estimator in Equation (16) is consistent.

The literature has investigated the problem of obtaining smooth estimates of empirical cdfs intensively. For instance, Berg and Harris (2008) use a diffusion-based approach while Veraverbeke, Gijbels, and Omelka (2014) use local linear regression estimates. One can benefit from these advances to refine the estimators. Recently, Ben Abdellah, L'Ecuyer, Owen, and Puchhammer (2018) show that convergence in the estimation of the distribution occurs not only when the sample is generated through Monte Carlo, but also through Quasi-Monte Carlo.

Let us now come to the numerical implementation. Proposition 1 suggests that a common problem in estimating δ_i, β_i^{KS}, and β_i^{Ku} is to find the critical points (local extrema) for the function $\Delta F_{Y|X_i\in\mathcal{C}_{m,i}}(y) = F_{Y|X_i\in\mathcal{C}_{m,i}}(y) - F_Y(y)$, confirming the intuition in Liu and Homma (2009). For simplicity in the following discussion, we assume that the output sample is without ties and reordered according to $y_j < y_{j+1}$, $j = 1, \ldots, n-1$, with the associated inputs being rearranged accordingly. The computation is simplified by the fact that in this case $\widehat{F}_Y(y_j)$ is already given by $\frac{j}{n}$. It is also useful to introduce the notion of *subsample run*.

Definition 1. *A* subsample run *is a maximal sequence of adjacent output values*

$$\{y_j, y_{j+1}, \ldots, y_{j+r-1}\} \subset Y_m$$

for which the associated input values $x_{\cdot i}$ are all contained in $\mathcal{C}_{m,i}$.

We can detect whether y_j belongs to a subsample run by considering the difference between $\Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y) = \widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y) - \widehat{F}_Y(y)$ at two consecutive realizations of y. In fact, it holds:

$$\begin{aligned}
&\Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y_j) - \Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y_{j-1}) = \\
&\left(\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y_j) - \widehat{F}_Y(y_j)\right) - \left(\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y_{j-1}) - \widehat{F}_Y(y_{j-1})\right) = \\
&\left(\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y_j) - \widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y_{j-1})\right) - \left(\widehat{F}_Y(y_j) - \widehat{F}_Y(y_{j-1})\right) \\
&= \begin{cases} \frac{1}{n_{m,i}} - \frac{1}{n}, & j : x_{ji} \in \mathcal{C}_{m,i}, \\ -\frac{1}{n}, & \text{otherwise.} \end{cases}
\end{aligned} \tag{17}$$

Hence, the difference in (17) is increasing in jumps of $\frac{1}{n_{m,i}} - \frac{1}{n} > 0$ within each subsample, while it is decreasing outside the subsample in smaller steps of $-\frac{1}{n}$. If the partition bins are chosen to be equally likely then $\frac{1}{n_{m,i}} - \frac{1}{n} \approx \frac{1}{M(n)}$, which shows how the partition size impacts the cdf differences. Regarding β_i^{KS} and β_i^{Ku}, a straightforward application of the cdf-based estimator (16) yields consistent estimators as the conditions of theorem 1 in Borgonovo et al. (2016) are satisfied. The insight added by (17) is that local extrema of cdf differences are located at the beginning and at the end of subsample runs. This fact can be exploited to speed up the search for the global extrema required by the Kolmogorov–Smirnov and the Kuiper metrics.

Regarding δ_i, the construction of estimators based on Equation (16) is more elaborate. In the remainder of this section, we summarize the main steps and intuition. Appendix B reports the technical details and Appendix C discusses the consistency of the estimators. The first step is to link the estimation of a density to the estimation of a cdf via Scheffé's theorem (Borgonovo, 2006; Devroye & Györfi, 1985; Scheffé, 1947). This theorem implies that it is equivalent to count (and sum) all the areas between the conditional and unconditional cdfs, or to count twice only the areas where the conditional cdf is greater (smaller) than the unconditional cdf. So, the problem boils down to determine the subset where $f_{Y|X_i\in\mathcal{C}_{m,i}}(y)$ is greater than $f_Y(y)$. Second, the δ importance measure requires that Y is an absolutely continuous random variable. In this case, one observes that $\Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}$ is not a good approximation of $\Delta F_{Y|X_i\in\mathcal{C}_{m,i}}$. The left graph in Fig. 3 offers a visual illustration, using the same data of Fig. 1. Note the jiggling in the left graph. The reason is that $\Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}$ is discontinuous, as Equation (17) suggests, while $\Delta F_{Y|X_i\in\mathcal{C}_{m,i}}$ is continuous because of the absolute continuity of Y. Here, the literature offers two main alternatives: kernel smoothing or moving

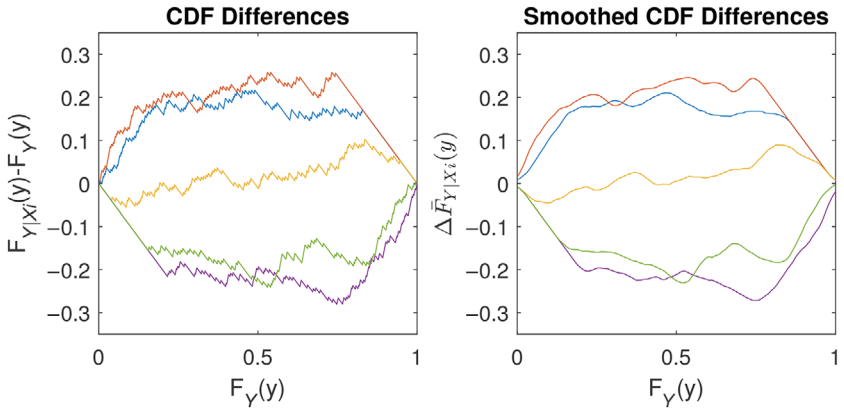

Fig 3. CDF differences without (left) and with (right) a moving average.

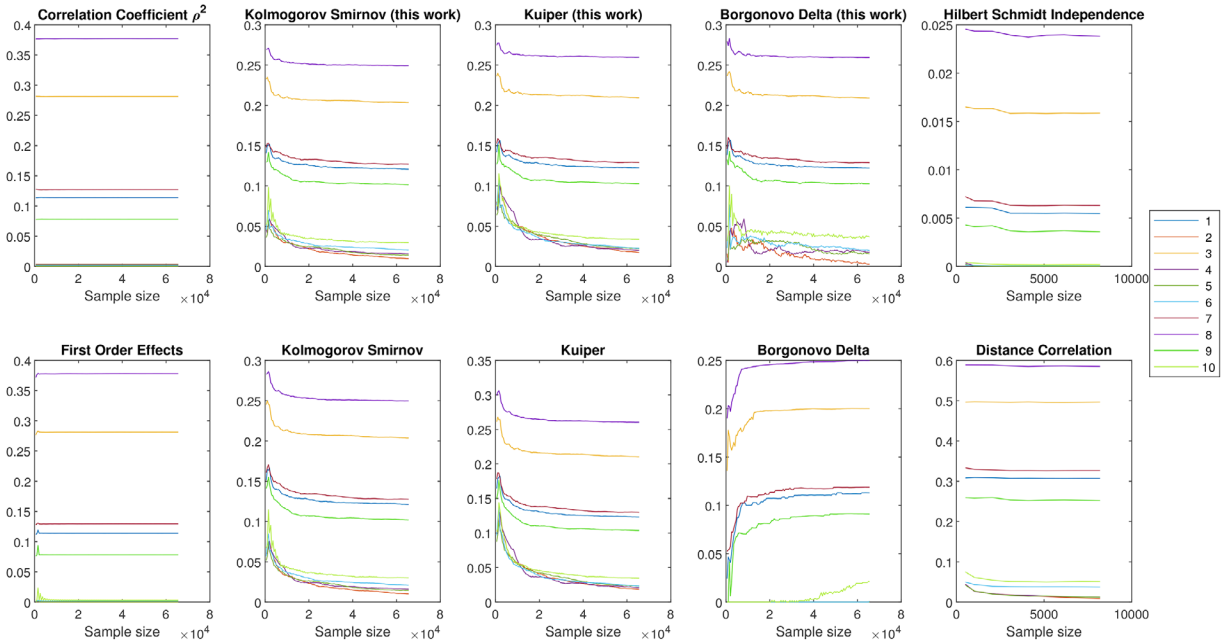

Fig 4. Numerical experiments for the wing-weight simulator. Convergence of the estimates of ρ, β^{KS}, β^{Ku}, δ. A sequence of quasi-random sample sizes from $n = 512$ up to $n = 65,536$ is used.

averages. To avoid kernel smoothing, we then apply a moving-average approach. The right graph in Fig. 3 displays the effect of processing $\Delta \widehat{F}_{Y|X_i \in \mathcal{C}_{m,i}}$ with a moving-average operator. The operation considerably reduces the jiggling, while the local and global extrema are still present. In general, smoothing has the advantage of reducing variance, but at the cost of introducing bias. At any finite sample size, despite the smoothing, the determination of the extreme values is still an error-prone process. Therefore, we need to take additional provisions, defining positive and negative subsample runs. These are technical aspects that we discuss in Appendix B.

All in all, one obtains the estimator

$$\hat{\delta}_i = \sum_{m=1}^{M} \frac{n_{m,i}}{n} \left(\sum_{t=1}^{T} \Delta \overline{F}_{Y|X_i \in \mathcal{C}_{m,i}}(\hat{b}_t^m) - \Delta \overline{F}_{Y|X_i \in \mathcal{C}_{m,i}}(\hat{a}_t^m) \right), \quad (18)$$

where $\hat{b}_t^m$ and $\hat{a}_t^m$ are properly selected values of Y in each partition, and $\overline{F}_{\{\cdot\}}$ denotes the smoothed cdfs.

The pseudo-code of Algorithm 1 summarizes the given-data estimators discussed thus far. It is based on an efficient way to estimate conditional cdfs and computes the sensitivity measures β^{KS}, β^{Ku}, δ by supplying two vectors, x and y, of length n.

Implementationwise, the selection of the conditional subsample with respect to input values

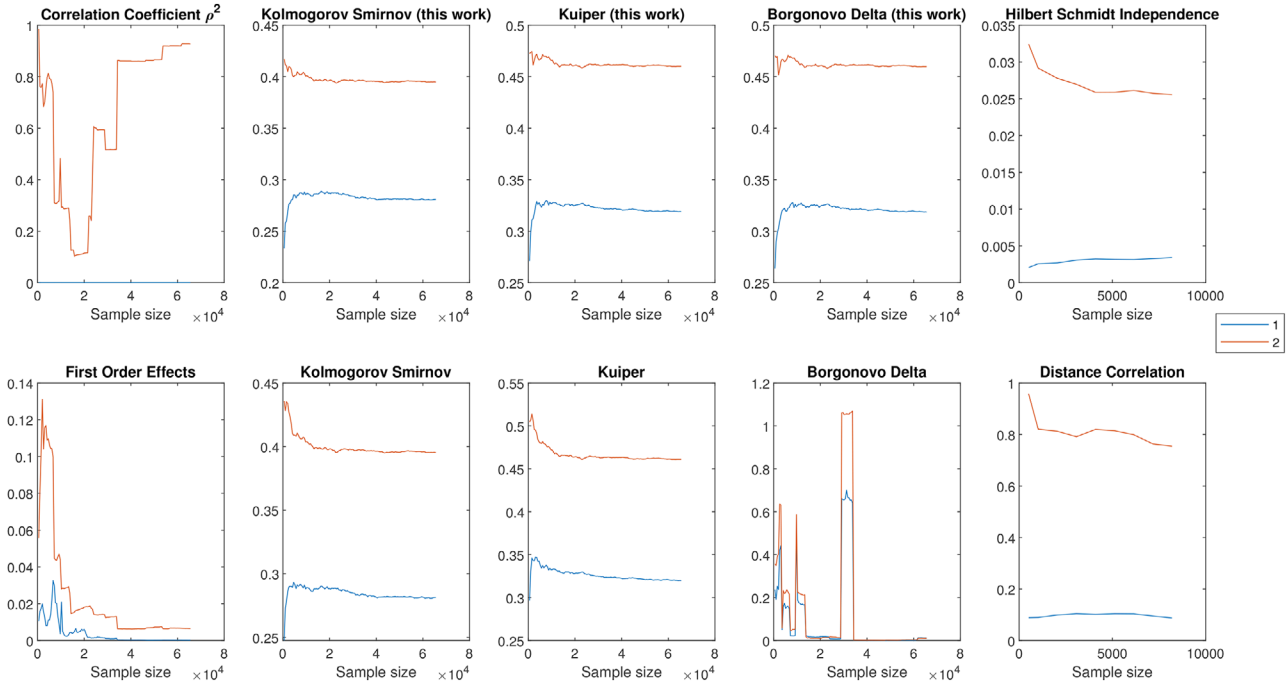

Fig 5. Numerical experiments for the two-variable Cauchy example. The horizontal axis reports a sequence of sample sizes from $n = 512$ up to $n = 65,536$ (crude Monte Carlo). The vertical axis reports the corresponding estimates of ρ, β^{KS}, β^{Ku}, δ, HSIC, and distance correlation.

Algorithm 1 Sensitivity indicators for input $x \in \mathbb{R}^n$ and output $y \in \mathbb{R}^n$ using M partitions

$z \leftarrow$ Sort x using y as key
for $m \leftarrow 1$ to M **do**
  $s \leftarrow$ Select elements of z between quantiles $\frac{m-1}{M}$ and $\frac{m}{M}$
    ▷ s is a Boolean false/true vector converted to integers 0 and 1
  $\Delta F \leftarrow$ Cumulative sum of s divided by sum of s − linear ramp from 1 to n divided by n
  $\Delta \bar{F} \leftarrow$ Apply a moving-average filter to ΔF (window size $\pm 3M$)
  $\mathrm{KS}_m \leftarrow \max |\Delta F|$
  $\mathrm{Ku}_m \leftarrow \max \Delta F - \min \Delta F$
  $\mathrm{Bo}_m \leftarrow$ Sum of maxima from positive runs − sum of minima from negative runs of $\Delta \bar{F}$
end for
return Means: $\hat{\beta}^{\mathrm{KS}} \leftarrow \frac{1}{M}\sum_{m=1}^{M} \mathrm{KS}_m$, $\hat{\beta}^{\mathrm{Ku}} \leftarrow \frac{1}{M}\sum_{m=1}^{M} \mathrm{Ku}_m$, $\hat{\delta} \leftarrow \frac{1}{M}\sum_{m=1}^{M} \mathrm{Bo}_m$

between the $\frac{m-1}{M}$ quantile and the $\frac{m}{M}$ quantile is implemented using the "is member of the subsample" information which is coded into a true/false vector that is interpreted as integers 1 and 0. Because the vector z is a presorted copy of x, the conditional cdf is given by a scaled version of the cumulative sum of the membership vector, while the unconditional cdf is a linear ramp. Thus, the operations to be performed are differences of cumulative sums that are used to determine the points of local extrema. Finally, the algorithm returns the estimates of the sensitivity measures by averaging over all partition bins.

We now analyze the algorithmic cost of the new estimation method, $\Gamma_{\mathrm{Estimator}}$. The run-time of Algorithm 1 is essentially driven by sorting the available data with respect to the output and also with respect to all inputs, which yields $k+1$ sorting operations of data of size n. Each sort can be performed with $O(n\log(n))$ operations (Knuth, 1997) so that $O((k+1)n\log n)$ operations are needed. Thus, the memory requirements of $\Gamma^{\text{New Method}}_{\text{Estimator}}$ are approximately linear in the sample size. We note that for kernel-based approaches like DCov or HSIC as presented in Appendix D, matrices of distances

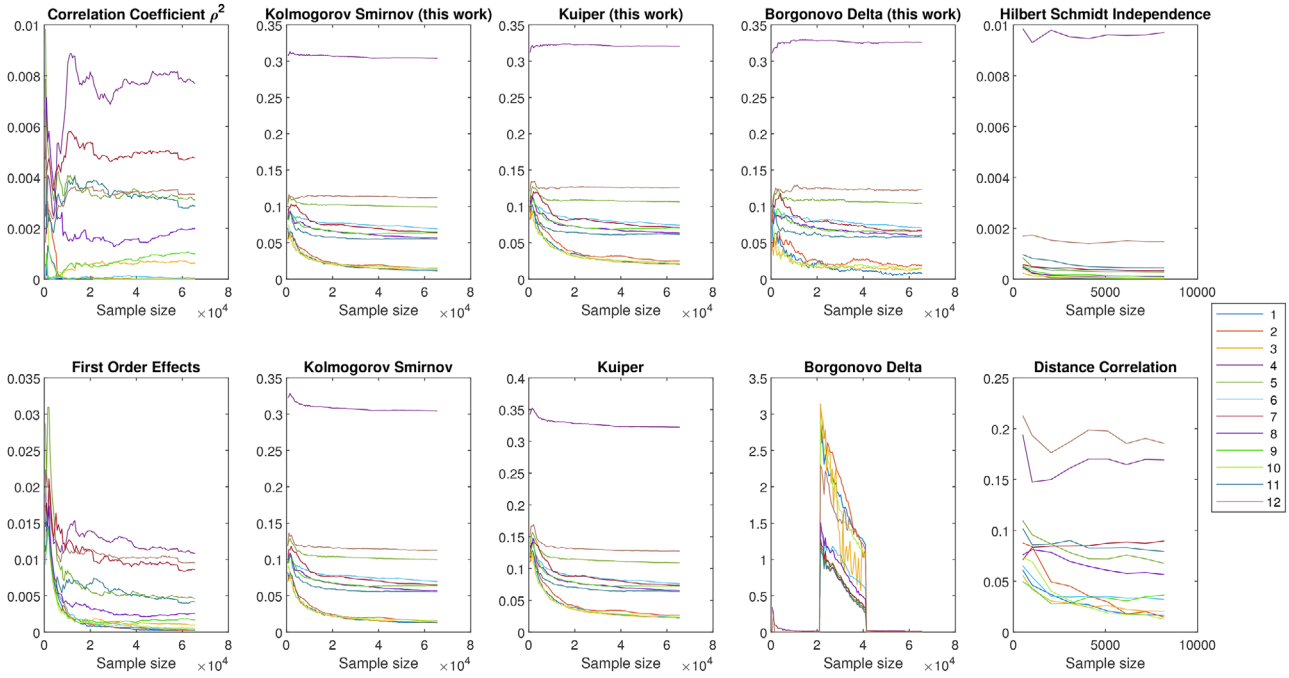

Fig 6. Numerical experiments for the Level E transport model. The horizontal axis reports a sequence of quasi-random sample sizes from $n = 512$ up to $n = 65,536$. The vertical axis reports the corresponding estimates of ρ, β^{KS}, β^{Ku}, δ, HSIC, and distance correlation.

between all data pairs have to be formed, both for the inputs and for the outputs. Hence the implementation implies, in principle, quadratic memory requirements $O(2n^2)$ when evaluating the kernel on all pairs of data. Furthermore, for HSIC with Gaussian kernels, one needs bandwidth estimation, which adds further complexity.

With respect to previous works, we also note that Algorithm 1 sorts the output, in contrast to the scatterplot partitioning idea of Section 3 and the algorithms discussed in Plischke (2012); Plischke, Borgonovo, and Smith (2013) where the sorting is on the realizations of X_i, with the output realizations that are then reordered accordingly.

Finally, we recall that counting the number of runs from the conditional subsample in the ordered output sample is at the basis of the Wald–Wolfowitz run test statistic (Wald & Wolfowitz, 1940). Hence, taking the largest cdf difference within each run combines ideas from the Kolmogorov–Smirnov test and the Wald–Wolfowitz run test.

5. NUMERICAL EXPERIMENTS

We start with a premise on partition selection. As mentioned in Strong and Oakley (2013); Borgonovo, Hazen, and Plischke (2016), there is a trade-off between the number of realizations to allocate to a partition and the number of partitions. In fact, the lower the number of partitions the more accurately we can estimate the inner statistic, but a low number of partitions may lead to failure in capturing the behavior of $\zeta_i(\mathbb{P}_Y, \mathbb{P}_{Y|X_i})$ as a function of X_i. The trade-off is well detailed in Strong and Oakley (2013) and Borgonovo et al. (2016). While for large sample sizes estimates become insensitive to the partition size, for sample sizes below $n = 2,000$ the partition selection strategy becomes relevant. The partition size M in the following experiments is linked to the sample size n via $M = \min\{\frac{n}{64}, 32\}$ guaranteeing a minimal subsample size of 64 realizations when one selects equally likely partition bins.

5.1. When Everything Works

We consider first the wing-weight simulator, a numerical code recently studied in the context of the estimation of Sobol' sensitivity indices in Jiménez Rugama and Gilquin (2018). The model simulates the weight of a light aircraft wing depending on 10 design parameters. The MATLAB file and implementation details are available from the library of

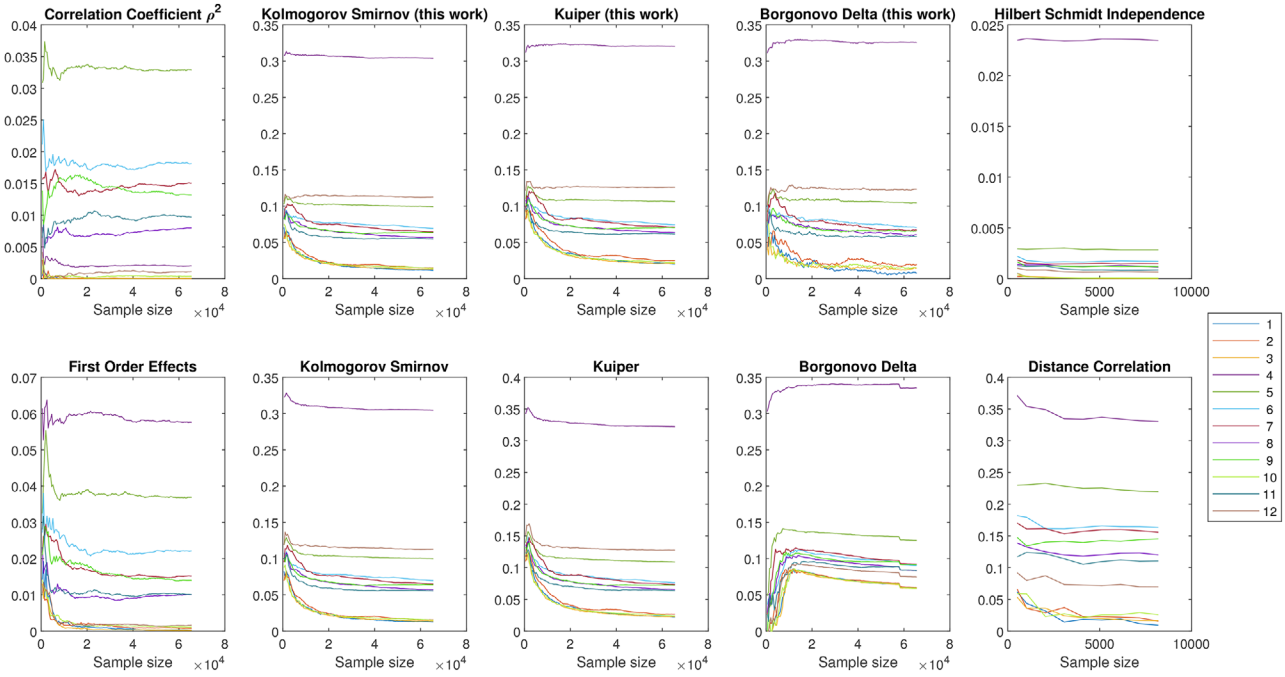

Fig 7. Level E Transport Model at time $t = 300,000$. Convergence study on log-transformed output, 65,536 QMC max.

Surjanovic and Bingham (2019). The code has 10 uncertain input parameters. Following the distributional assignment in Jiménez Rugama and Gilquin (2018, table 9, p. 735), we generate a sequence of samples of increasing size, up to $n = 65,536$, for the estimation of variance-based sensitivity measures (the size is the same as in Jiménez Rugama and Gilquin, 2018). We utilize the same samples and distributional assignments to estimate, besides first-order sensitivity measures, correlation coefficients, the β^{KS}, β^{Ku}, δ sensitivity measures, as well as HSIC and distance correlation. For the β^{KS}, β^{Ku}, δ, we compare the estimators proposed in this work against the estimators of Plischke et al. (2013) and Borgonovo et al. (2014).

The graphs in Fig. 4 report the sample sizes on the horizontal axis, and the corresponding estimates on the vertical axis. The upper left panel reports the squared correlation coefficients, the lower left variance-based first-order indices. The upper three centered panels are obtained by Algorithm 1 applied directly to the sample data, while the lower panels use the estimators in Plischke et al. (2013). The rightmost panels display the results for the HSIC (upper graph) and distance correlation (lower graph).

Fig. 4 shows that all estimates rapidly converge and the ranking of the most important inputs is correctly reported by all estimators already at the smallest tested sample size ($n = 512$), with convergence in the estimates obtained for $n \geq 1,024$. For this simulator, a linear regression surface just with additive terms would fit well, capturing over 90% of the output variance, signaling a mild behavior of the input–output mapping. Thus, in this case, the newly introduced estimators do not display an advantage over estimators previously introduced. (Note that the horizontal axis of distance-correlation and of HSIC stops at about $n = 7,000$. The large memory requirements would make it not possible to obtain numerical estimates for larger sample sizes. However, all relevant inputs are already identified by these sensitivity measures at smaller sample sizes.)

5.2. When Sparsity Jumps In

Let us now consider what happens when the output is sparse (ranges over orders of magnitude), a situation often encountered in risk analysis applications (see Iman & Hora, 1990). To show that sparsity issues can emerge independently of the simulator dimensionality, we start with a low-dimensional test case. Consider $Y = X_1^{-1} + X_2$, with X_1 Cauchy(0,3)-distributed and X_2 independently Cauchy(1,0.5)-distributed. Then, Y is

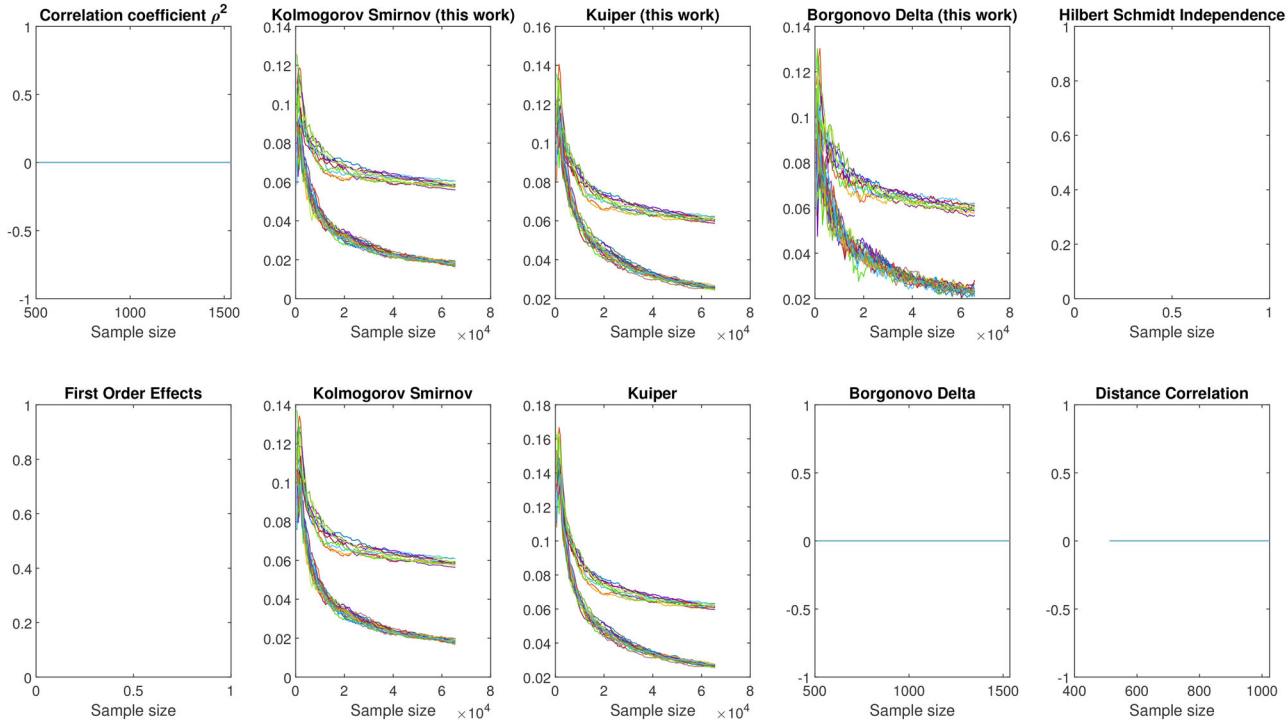

Fig 8. Product of lognormals, input dimension $d = 30,000$. Convergence study, 65,536 MC max. Shown are the 30 first parameters.

Cauchy$(1, \frac{5}{6})$-distributed, and also all the conditional distributions are Cauchy. Therefore, it is possible to obtain the values of moment-independent sensitivity measures analytically. Specifically, we have $\delta_1 = 0.31$ and $\delta_2 = 0.46$. Because the distributions are all unimodal, by Proposition 1 we expect $\delta_i = \beta_i^{\mathrm{Ku}}$. Fig. 5 shows the estimates of the sensitivity measures addressed thus far. For variance-based sensitivity measures, we observe that the larger the sample size, the more numerical issues appear. Furthermore, the kernel-density–based estimator of δ of Plischke et al. (2013) fails (lower row) to produce consistent estimates. Then, from the estimators in the lower row of Fig. 5 one would not recover the theoretically expected identity between δ_i and β_i^{Ku}. This identity is, instead, recovered by the estimators in the upper row.

The previous test case is based on Cauchy-distributed random variables, which have mode and median, but no mean. Thus, techniques that rely on the first moment do not converge. The test case may seem artificial, but it captures a problem that occurs with one of the most widely used test cases in sensitivity analysis, Level E.

The geosphere transport model Level E has been widely studied in sensitivity analysis since Saltelli et al. (1999) and Saltelli and Tarantola (2002). The simulator computes the annual radiation dose to humans as a result of leakage from a hypothetical underground disposal site for nuclear waste spanning a time horizon going from 2×10^4 to 2×10^9 years into the future. We analyze the output at timestep $t = 300,000$. The simulator features 12 uncertain input parameters, whose distributions have been assigned based on expert opinions in OECD (1989) and have been used consistently in all subsequent studies. We refer to Saltelli and Tarantola (2002) for the description. Issues with this model can be spotted from the slow convergence of the first-order effects. Moreover, the ranking of the first two parameters is not consistent over all measures. A neat separation of the sensitivity measures of minor contributors is only possible for large sample sizes, $n \geq 8,192$ (Fig. 6).

As the output spans orders of magnitudes, the analyst usually employs an output transformation. We then apply a logarithmic transformation of the output. Zero values and stray negative values are mapped to the smallest positive number representable in floating-point arithmetic before applying the transformation. As Fig. 7 shows, the numerical convergence issues are now resolved, but

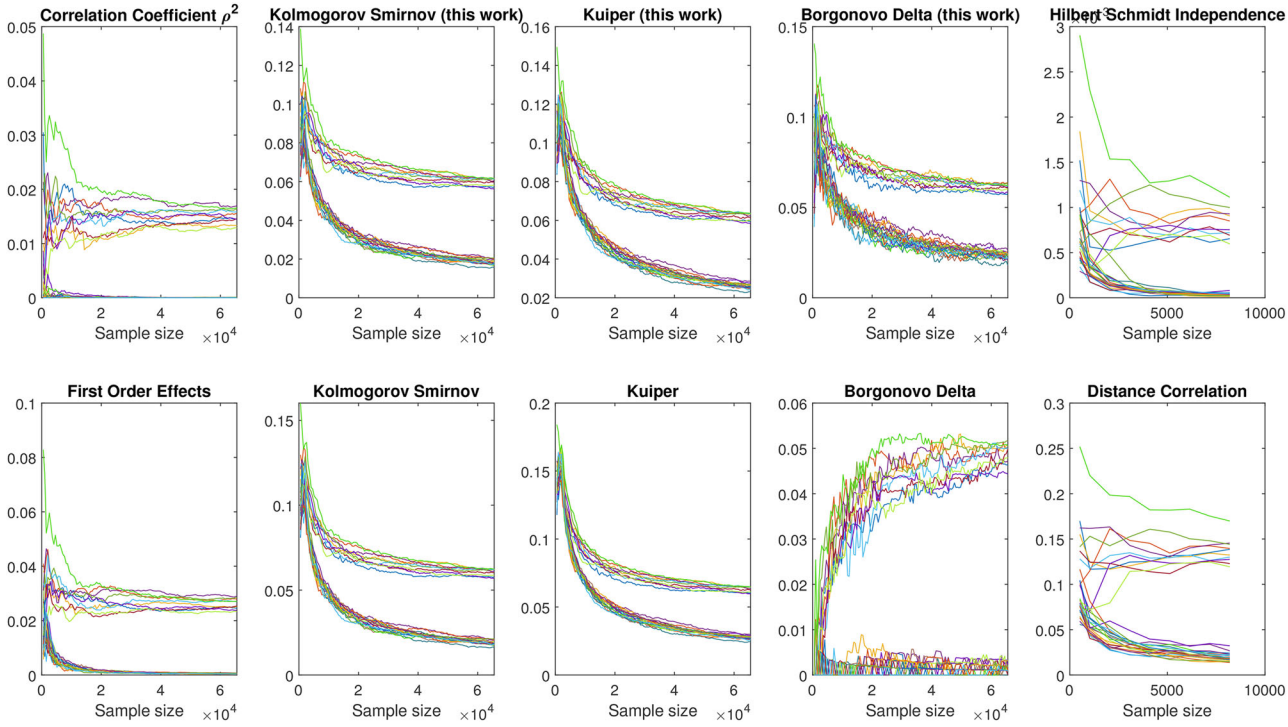

Fig 9. Product of lognormals, input dimension $d = 30,000$. Convergence study on log-transformed output, 65,536 MC max. Shown are the 30 first parameters.

the transformation has changed the ranking of the inputs.

Overall, the analysis shows that for the dose at $t = 300,000$ years, the estimation of global sensitivity measures is challenging. The analysis with and without transformations, however, permits to identify two key drivers of uncertainty. The steam flow rate X_{12} is a stochastic variable whose properties cannot be influenced by technical design considerations. Thus, a risk manager knows that she/he cannot intervene on this parameter by changing the design.

The fact that this parameter is a key driver of uncertainty then means that additional information on the parameter will, in fact, reduce uncertainty about the output but will not be informative about possible ameliorations of the performance of the waste repository. The second important parameter X_4, the water velocity in the first geosphere layer, refers to characteristics of the host rock formation. Thus, getting more specific on this aspect of the repository design has the potential to reduce the problem uncertainty.

5.3. When Dimensionality Adds To Sparsity

Now, if sparsity of output is paired with a large number of input parameters, then classical screening methods for which the sample design depends on the number of inputs are not applicable. Given-data techniques, however, may still extract information from a sample when its size of is smaller than the number of dimensions. We consider the product of standard lognormals, $Y = \prod_{i=1}^{10} X_i^{32} \prod_{i=11}^{d} X_i$, $X_i \sim \log N(0,1)$ i.i.d, with $d = 30,000$. Additional tests were performed with $d = 30$, $d = 300$, and $d = 3,000$, but these offer qualitatively the same results and are not reported here. In order to deal with over- and underflow, infinite values in the simulator output have been replaced by the largest representable positive floating point number and zeros by the smallest representable positive floating point number. Fig. 8 shows the results of computing different sensitivity measures using alternative algorithms. Estimation methods not derived from cdf estimations are not working reliable on this data set. This is a consequence of the sparsity of the output. The log-transformed output is studied in Fig. 9. Note that

the simulator output becomes linear after the transformation. As the sparsity and therefore also the variance of the model output is now notably reduced, all methods produce results.

Throughout we have considered a sample of size n, with the intuition that this is the maximum sample size that the time/resource budget allows. However, Figs. 4–9 suggest an iterative version of the estimation procedure. One starts with a trial budget n_0 of model runs, and then increases it to $n_1, n_2, \ldots$ monitoring the convergence in the estimates. Once the difference in two subsequent estimates is small enough the analyst can stop the process. The application of an automated sequential approach is part of future research.

6. CONCLUSIONS

The computation of global importance measures is an important part of quantitative risk assessment. However, such computation can be extremely challenging. This work contributes to improving the efficiency and lowering the computational cost associated with the estimation of global sensitivity measures.

First, the investigation contributes in raising the awareness that estimation challenges do not come solely from dimensionality, but also from sparsity. Thus, our finding add to the risk analysis literature on the use of transformations originated by the work of Iman and Hora (1990). We have proposed and analyzed a new computation method that relies on rewriting the estimators of global sensitivity measures using the cdf of the output. It is based on the given-data principle, and thus it reduces the effect of the curse of dimensionality. Moreover, the new approach allows the manipulation of numbers on the [0,1] scale. As such, it works seemlessly for the estimation of any global sensitivity measure based on the cdf of the output such as β^{Ks} and β^{Ku}. For density-based sensitivity measures, we have proposed a moving-average approach that bypasses kernel-density estimation. We have studied the resulting estimator proving its convergence. We have also discussed the algorithmic cost of the new implementation and compared it to the algorithmic cost of dependence measures such as distance covariance and HSIC.

Because the new method allows the manipulation of numbers on the [0,1] scale, it becomes a potential remedy to the curse of sparsity. We have tested this assertion through a series of experiments of increasing complexity, from the computationally friendly wing-weight simulator to a 30, 000 input case study in which the model output variance reaches the numerical range of the floating-point representation of numbers in the computer. For each experiment, we have performed tests with and without transformations. Results show that the estimates to converge at reasonable sample sizes for all the examples even without the use of transformations.

There are some key recommendations for a risk analyst emerging from our investigation. First, sparsity may affect the performance of global sensitivity estimators as much as dimensionality and, in any case, sparsity acts independently of dimensionality. The use of transformations may not directly solve the curse of sparsity. Not only do transformations introduce interpretation issues, but a transformation may not be numerically effective, especially if the estimators rely on kernel smoothing. Moreover, sparsity does not impact estimators of alternative global sensitivity measures in the same way; a recommendation is, then, that the analyst employs an ensemble of global sensitivity measures. Relying on a single indicator may lead to unreliable conclusions due to numerical estimation issues. In this respect, the case studies have shown that employing the new method simultaneously with other global sensitivity measures allows the analyst to obtain solid insights on the key drivers of uncertainty while keeping computational burden under control.

Avenues for future research are the comparison, in high-dimensional settings, of the global estimators discussed in this work with screening techniques such as the method of Morris or sequential bifurcation, as well as the implementation of a sequential version of the method. Moreover, the addition of further cdf-based distances like Cramer/von Mises to the global sensitivity measure portfolio is also pursued.

ACKNOWLEDGMENTS

We wish to thank the editors Prof. Tony Cox and Prof. Seth Guikema for their editorial efforts. We are also really grateful to the three anonymous reviewers for several perceptive and constructive suggestions.

Open access funding enabled and organized by Projekt DEAL.

APPENDIX A: PROOFS

Proposition 1. *Using the notations* $\Delta F_{Y|X_i}(y) = F_{Y|X_i}(y) - F_Y(y)$ *and* $\Delta f_{Y|X_i}(y) = f_{Y|X_i}(y) - f_Y(y)$ *the inequality* $\beta_i^{KS} \le \beta_i^{Ku} \le 2\beta_i^{KS}$ *follows immediately from*

$$\begin{aligned}&\sup\left|\Delta F_{Y|X_i}(y)\right| = \sup\{-\Delta F_{Y|X_i}(y), \Delta F_{Y|X_i}(y)\}\\&\le \sup\{-\Delta F_{Y|X_i}(y)\} + \sup\{\Delta F_{Y|X_i}(y)\} \le 2\sup\left|\Delta F_{Y|X_i}(y)\right|.\end{aligned} \tag{A1}$$

Suppose that there are two functions $y_0, y_1 : \mathcal{X}_i \to \mathcal{Y}$, $y_0(\cdot) \neq y_1(\cdot)$ *such that*

$$\begin{aligned}&\Delta f_{Y|X_i}(y_0(x_i)) = 0 = \Delta f_{Y|X_i}(y_1(x_i))\\&\text{and}\quad \Delta f_{Y|X_i}(y) \neq 0 \quad \text{for all other } y.\end{aligned}$$

These zeros correspond to the minimum (≤ 0) *and maximum* (≥ 0) *of* $\Delta F_{Y|X_i}$. *Hence,*

$$\beta_i^{Ku} = \int \left(\left|\Delta F_{Y|X_i}(y_0(x_i))\right| + \left|\Delta F_{Y|X_i}(y_1(x_i))\right|\right) dx_i.$$

Now recall Scheffé's theorem (Devroye & Györfi, 1985; Scheffé, 1947) which states that given two probability density functions (pdfs) f_1 *and* f_2 *one has*

$$\int_{\mathcal{Y}} \left|f_1(y) - f_2(y)\right| dy = 2\sup_{B\in\mathcal{B}}\left|\int_B f_1(y)\,dy - \int_B f_2(y)\,dy\right|. \tag{A2}$$

Hence for the intervals $B(x_i) = [\min\{y_0(x_i), y_1(x_i)\}, \max\{y_0(x_i), y_1(x_i)\}]$, *by (A2)* δ_i *satisfies*

$$\begin{aligned}&\delta_i = \int\left|\int_{B(x_i)} (f_y - f_{Y|X_i})\,dy\right| dx_i =\\&\int\left|\Delta F_{Y|X_i}(y_0(x_i)) - \Delta F_{Y|X_i}(y_1(x_i))\right| dx_i = \beta_i^{Ku}.\end{aligned}$$

If there is only one nonvanishing extremum for all x_i *then* $\Delta F_{Y|X_i}$ *contains no sign change and therefore* $\beta_i^{Ku} = \beta_i^{KS}$.

If $g(\cdot)$ *is a function depending on its sole scalar parameter* x_i *then we have for* $y_1 < g(x_i) < y_2$ *that* $Ku(x_i) \ge |1 - F_Y(y_2) - (0 - F_Y(y_1))| \to 1$ *as* $y_1 \to y_2$. *If* $g(\cdot)$ *is monotonically increasing then by transformation invariance we have* $\beta^{KS} = \int_0^1 \max\{u, 1-u\}\,du = \frac{3}{4}$.

APPENDIX B: ESTIMATION OF BORGONOVO'S δ FROM CDFS: TECHNICAL DETAILS

Scheffé's theorem (Devroye & Györfi, 1985; Scheffé, 1947) allows us to write

$$\begin{aligned}&\int_{\mathcal{Y}} \left|f_{Y|X_i\in\mathcal{C}_{m,i}}(y) - f_Y(y)\right| dy\\&= 2\int_{B_+^m} \left(f_{Y|X_i\in\mathcal{C}_{m,i}}(y) - f_Y(y)\right) dy,\end{aligned} \tag{B1}$$

where $B_+^m = \{y : f_{Y|X_i\in\mathcal{C}_{m,i}}(y) \ge f_Y(y)\}$, i.e., B_+^m is the subset in $\mathcal{Y}$ where $f_{Y|X_i\in\mathcal{C}_{m,i}}(y)$ is above $f_Y(y)$. With a slight notation abuse, (B1) can also be written in terms of cdfs

$$\begin{aligned}&\int_{\mathcal{Y}} \left|f_{Y|X_i\in\mathcal{C}_{m,i}}(y) - f_Y(y)\right| dy\\&= 2\left(F_{Y|X_i\in\mathcal{C}_{m,i}}(B_+^m) - F_Y(B_+^m)\right).\end{aligned} \tag{B2}$$

Equation (B2) has a geometric interpretation which can be used for the estimation of δ. Returning to Fig. 2, observe that, in association with (B2), it is equivalent to count (and sum) all the areas between the conditional and unconditional cdfs, or to count twice only the areas where the conditional cdf is greater (smaller) than the unconditional cdf. So, the problem boils down to determining the subset B_+^m (Liu & Homma, 2009). This set is a union of intervals of which the endpoints are critical points (local extrema) for the function $\Delta F_{Y|X_i\in\mathcal{C}_{m,i}}(y) = F_{Y|X_i\in\mathcal{C}_{m,i}}(y) - F_Y(y)$ (Borgonovo, Castaings, & Tarantola, 2011; Liu & Homma, 2009). Now, the subset B_+^m has to be determined from the sample and the conditional subsample. But for this task, $\Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}$ is not a good approximation of $\Delta F_{Y|X_i\in\mathcal{C}_{m,i}}$, as the former is discontinuous (see Equation (17)), while the latter is continuous because of the absolute continuity of Y. We therefore suggest the following simple form of data-smoothing: The averaged version $\Delta\overline{F}_{Y|X_i\in\mathcal{C}_{m,i}}$ of $\Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}$ is obtained from

$$\Delta\overline{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y_j) = \frac{1}{6M+1}\sum_{\ell=\max(j-3M,1)}^{\min(3M,n)} \Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y_\ell). \tag{B3}$$

For the size of the moving-average window, note that the chance of selecting an output realization which is contained in the subsample is, on average, $\frac{1}{M}$ (where M is the number of partitions, as above). Hence, one out of M realizations is from the subsample and therefore responsible for a jump in $\Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}$. Therefore, when applying a moving average to $\Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m,i}}$

the minimal window size is equal to the partition size as then a single jump contribution is balanced out. In our experience, a larger ($\pm 3M$) moving window is better suited to capture the variability. For finite sample sizes, despite the smoothing, the determination of the extreme values to approximate the set B_m^+ is still an error-prone process as local extreme values appear and vanish depending on the smoothing parameters and it is not clear if they are numerical artifacts or ground truth. Therefore, we consider the union of intervals $\hat{B}_+ = \bigcup_{t=1}^{T}[\hat{a}_t, \hat{b}_t]$ as an approximation of B_+^m where one identifies $\hat{b}_t^m$ as arguments maximizing $\Delta\overline{F}_{Y|X_i\in\mathcal{C}_{m,i}}$ in each of the positive runs, and $\hat{a}_t^m$ as minimizers in each of the negative runs (i.e., one value per run). Here, we call positive run a maximal sequence of adjacent output values where the smoothed cdf difference is positive,

$$\left\{ y_j, \ldots, y_{j+r-1} : \begin{cases} \Delta\overline{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y) > 0 \text{ for all } y = y_j \ldots, y_{j+r-1}, \\ \Delta\overline{F}_{Y|X_i\in\mathcal{C}_{m,i}}(y) \le 0 \text{ for } y = y_{j-1}, y_{j+r} \end{cases} \right\}.$$

A negative run is defined in a similar way. The maximum distances between the cdfs within each of the run have to be considered for forming the estimate of the δ measure, lead to Equation (18).

APPENDIX C: MOVING-AVERAGE ESTIMATOR CONSISTENCY

In this section, we provide greater details on the convergence properties of the cdf-based estimators proposed here. Consider first that X_i is an absolutely continuous random variable in $\mathcal{X}_i \subset \mathbb{R}$. Then, for every point $x_i^0 \in \mathcal{X}_i$ there exists a series of intervals in the partitions such that $x_i^0 = \bigcap_n \mathcal{C}_{m(n),i}(n)$, $m \le M(n)$. The notation $M(n)$ makes explicit the dependence of partition classes on the sample size and we assume that the function $M(n)$ satisfies the assumptions of theorem 1 in Borgonovo et al., (2016). We have the following inequality:

$$\left|\Delta\widehat{F}_{Y|X_i\in\mathcal{C}_{m(n),i}(n)}(y) - \Delta F_{Y|X_i=x_i^0}(y)\right| \le \left|\widehat{F}_{Y|X_i\in\mathcal{C}_{m(n),i}(n)}(y)\right)$$
$$\left(-\widehat{F}_{Y|X_i\in\mathcal{C}_{m(n),i}(n)}(y)\right| + \left|\widehat{F}_Y(y) - \widehat{F}_Y(y)\right|.$$

For the first term, absolute continuity of X_i allows us to write

$$\hat{F}_{Y|X_i\in\mathcal{C}_{m(n),i}(n)}(y) = \left(\int_{\mathcal{C}_{m(n),i}(n)} f_Y(y)dy\right)^{-1}$$
$$\int_{\mathcal{C}_{m(n),i}(n)} \hat{F}_{Y|X_i=\xi} f_i(\xi)d\xi \le \max_{\xi\in\mathcal{C}_{m(n),i}(n)} \hat{F}_{Y|X_i=\xi},$$

which converges to $\hat{F}_{Y|X_i=x_i^0}$ by partition refinement. For the second term, $\hat{F}_Y(y) \to F_Y(y)$, by the law of large numbers, as also given in the proof of theorem 1 of Borgonovo et al. (2016). Therefore, the estimators are asymptotically consistent.

Now, consider the averaged version of the distance between the conditional and the unconditional empirical cdf. For this estimator, we need the assumption that also Y is absolutely continuous. We have $\Delta\bar{F}_{Y|X_i}(y) = \frac{1}{6M(n)+1}\sum_{j=-3M(n)}^{3M(n)} \Delta\hat{F}_{Y|X_i}(\hat{F}_Y^{-1}(\hat{F}_Y(y) + \frac{j}{n}))$. $\Delta\bar{F}$ is therefore sandwiched between the minimal and maximal values of $\Delta\hat{F}$ over this range,

$$\min_{\substack{j=-3M(n),\ldots,3M(n)\\ \hat{F}_Y(y)+\frac{j}{n}\in[0,1]}} \Delta\hat{F}_{Y|X_i}\left(\hat{F}_Y^{-1}\left(\hat{F}_Y(y) + \frac{j}{n}\right)\right) \le$$
$$\Delta\bar{F}_{Y|X_i}(y) \le$$
$$\max_{\substack{j=-3M(n),\ldots,3M(n)\\ \hat{F}_Y(y)+\frac{j}{n}\in[0,1]}} \Delta\hat{F}_{Y|X_i}\left(\hat{F}_Y^{-1}\left(\hat{F}_Y(y) + \frac{j}{n}\right)\right).$$

We have seen above that $\hat{F}_Y(y) \to \hat{F}_Y(y)$ and $\Delta\hat{F}_{Y|X_i} \to \Delta F_{Y|X_i}$ as n increases. We also have that $|\frac{j}{n}| \le \frac{3M(n)}{n} \to 0$ as $n \to \infty$. Then, we need to be reassured that $\hat{F}_Y^{-1}(\hat{F}_Y(y) + \frac{j}{n}) \to y$ as n increases, where $\hat{F}_Y^{-1}$ is the generalized inverse, i.e., $\hat{F}_Y^{-1}(u) = \inf\{y \in \mathcal{Y} : \hat{F}_Y(y) \ge u\}$. For the term $\hat{F}_Y(y) + \frac{j}{n}$ we register $\hat{F}_Y(y) + \frac{j}{n} \to F_Y(y)$ as n tends to infinity. Then, we need that $\hat{F}_Y^{-1} \to F_Y^{-1}$. For this, absolute continuity of Y comes into play. In fact, if Y is absolutely continuous, then F_Y is differentiable and strictly increasing. Therefore, we always have a nonzero derivative at any point and the quantile function is defined for any value of y. Under these conditions, the empirical quantile function tends to the true value (see van der Vaart, 2000, chapter 21).

APPENDIX D: DETAILS ON DISTANCE CORRELATION AND HILBERT-SCHMIDT INDEPENDENCE CRITERION (HSIC)

If Z is an ℓ-dimensional random vector and Y is a random variable, the distance covariance is defined by taking a weighted difference between (possibly multidimensional) characteristic functions,

$$\text{dcov}\,(Z, Y)^2 = C_\ell \int_{\mathbb{R}}\int_{\mathbb{R}^\ell} ||s||^{-1-\ell}|t|^{-2}\left|\mathbb{E}\left[e^{is^TZ+itY}\right]\right)$$
$$\left(-\mathbb{E}\left[e^{is^TZ}\right]\mathbb{E}\left[e^{itY}\right]\right|^2 dsdt, \qquad \text{(D1)}$$

where C_ℓ is a normalizing constant related to the volume of a unit hyperball. The weights are chosen in such way that the measure is invariant with respect to rotations of the ℓ-dimensional space of Z. Analogously to the standard correlation of (2), the distance correlation is given by

$$\mathrm{d}_\varrho(Z,Y) = \frac{\mathrm{dcov}(Z,Y)}{\sqrt{\mathrm{dvar}(Z)dvar(Y)}}$$
$$\text{where} \quad \mathrm{dvar}(Z) = \mathrm{dcov}(Z,Z). \tag{D2}$$

Under the condition that the first moments of Z and Y are finite, $\mathrm{dcov}(Z,Y) = 0$ if and only if Z and Y are independent, i.e., it satisfies the nullity-implies-independence property. The distance correlation d_ϱ is also invariant under affine linear transformations of Z and Y. Estimators that avoid the use of characteristic functions are working on the trace product of difference matrices $A_{i,j} = (||z_{i,\cdot} - z_{j,\cdot}||)_{i,j=1,\ldots,n}$ and $B_{i,j} = (|y_i - y_j|)_{i,j=1,\ldots,n}$ for realizations $z_{i\cdot}, z_{j,\cdot} \in \mathbb{R}^\ell$ and $y_i, y_j \in \mathbb{R}$,

$$\mathrm{dcov}^2(\mathrm{Z},\mathrm{Y}) = \frac{1}{(n-1)^2}$$
$$\mathrm{trace}\left(A(I - n^{-1}\mathbf{1}\mathbf{1}^T)B(I - n^{-1}\mathbf{1}\mathbf{1}^T)\right), \tag{D3}$$

where I is the $n \times n$ identity matrix, $\mathbf{1}$ is a vector of ones, and n is the sample size. In Székely et al. (2007), the entries of the distance matrices A and B are centered with respect to column, row, and overall averages, however, due to the properties of the trace product, (D3) needs only the row averages, $n^{-1}A\mathbf{1}$ and $n^{-1}B\mathbf{1}$. The HSIC replaces the difference matrices in (D3) with appropriate kernels from reproducing kernel Hilbert spaces, i.e., $A^{\mathrm{HSIC}}_{i,j} = K_Z(z_{i,\cdot}, z_{j,\cdot})$ and $B^{\mathrm{HSIC}}_{i,j} = K_Y(y_i, y_j)$ (see Gretton, Bousquet, Smola, & Schölkopf, 2005, for further details). Then, analogously to (D3), HSIC can be computed via

$$\mathrm{HSIC} = \frac{1}{(n-1)^2}\,\mathrm{trace}\left(A^{\mathrm{HSIC}}(I - n^{-1}\mathbf{1}\mathbf{1}^T)\right)$$
$$\cdot\left(B^{\mathrm{HSIC}}(I - n^{-1}\mathbf{1}\mathbf{1}^T)\right). \tag{D4}$$

For using distance correlation or HSIC as sensitivity measures, Z is given by a single input factor X_i or a group of factors of interest, $(X_{i_1}, \ldots, X_{i_\ell})$.

APPENDIX E: DETAILS ON KERNEL-DENSITY ESTIMATION

Let $\{(x_{ji}, y_j) \mid j = 1, \ldots, n\}$ be a given sample of realizations of $(\mathbf{X}, Y)$. The estimate $\hat{f}_Y(\cdot)$ is obtained from a kernel-density estimation of all realizations $\{y_j | j = 1, \ldots, n\}$ while $\hat{f}_{Y|X_i \in \mathcal{C}_{m,i}}(\cdot)$ is obtained from a kernel-density estimation of the subset $Y_{m,i} = \{y_j | x_{ji} \in \mathcal{C}_{m,i}\}$. For a given kernel $K(\cdot)$ and $m = 1, \ldots, M$, the kernel-density estimates are

$$\hat{f}_Y(y) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\alpha}K\left(\frac{y - y_j}{\alpha}\right),$$
$$\hat{f}_{Y|X_i \in \mathcal{C}_{m,i}}(y) = \frac{1}{n_{m,i}}\sum_{x_{ji} \in \mathcal{C}_{m,i}}\frac{1}{\alpha_m}K\left(\frac{y - y_j}{\alpha_m}\right). \tag{E1}$$

Here, $n_{m,i} = \sum_{x_{ji} \in \mathcal{C}_{m,i}} 1$ is the number of realizations in class $\mathcal{C}_{m,i}$ of the partition of $\mathcal{X}_i$. Let us recall a vector-valued formulation of (E1).

Proposition E1 (Bowman & Azzalini, 2003). *Given a sample $z = (z_i)$ of the random variable Z, quadrature points $\zeta = (\zeta_j)$, and a kernel function K with bandwidth h, construct the weight matrix*

$$W_{ij}(z,\zeta) = \frac{1}{h}K\left(\frac{z_i - \zeta_j}{h}\right). \tag{E2}$$

Then $\hat{f}_Z(\zeta_j) = \frac{1}{n}\sum_i W_{ij}$.

This proposition is directly applicable to the realizations y and the quadrature points v to obtain $\hat{f}_Y$. If the bandwidths satisfy $\alpha = \alpha_m$, $m = 1, \ldots, M$ then estimating the conditional pdfs amounts to a subset selection of the rows of the weight matrix W analogous to the calculation of $\bar{y}_j$ preceding (11). In particular, we have $\hat{f}_Y(v_q) = n^{-1}\sum_{j=1}^{n} W_{jq}(y,v)$ and $\hat{f}_{Y|\mathcal{C}_{m,i}}(v_q) = n_{m,i}^{-1}\sum_{j:x_{ji} \in \mathcal{C}_{m,i}} W_{jq}(y,v)$. One of the remaining issues is the choice of the kernel function K and an associated bandwidth h. We use a rule-of-thumb bandwidth (Sheather, 2004) and keep it constant when passing over to conditional pdfs. This avoids also the paradoxical situation that the estimates of the conditional pdfs have a larger support than the estimate of the unconditional pdf, as the less data are available the larger an automatically selected bandwidth is chosen.

REFERENCES

Apostolakis, G. (2004). How useful is quantitative risk assessment? How useful is quantitative risk assessment? *Risk Analysis*, *24*(3), 515–520.

Ben Abdellah, A., L'Ecuyer, P., Owen, A. B., & Puchhammer, F. (2018). Density estimation by randomized Quasi-Monte Carlo. http://www.arxiv.org/abs/1807.06133.

Berg, B. A., & Harris, R. C. (2008). From data to probability densities without histograms? *Computer Physics Communications*, *179*, 443–448.

Bettonvil, B., & Kleijnen, J. P. C. (1997). Searching for important factors in simulation models with many factors: Sequential

bifurcation. *European Journal of Operational Research*, *96*(1), 180–194.
Borgonovo, E. (2006). Measuring uncertainty importance: Investigation and comparison of alternative approaches. *Risk Analysis*, *26*(5), 1349–1361.
Borgonovo, E. (2017). *Sensitivity analysis. An introduction for the management scientist*. Cham: Springer Verlag.
Borgonovo, E., Castaings, W., & Tarantola, S. (2011). Moment independent importance measures: New results and analytical test cases. *Risk Analysis*, *31*(3), 404–428.
Borgonovo, E., Hazen, G. B., & Plischke, E. (2016). A common rationale for global sensitivity measures and their estimation. *Risk Analysis*, *36*(10), 1871–1895.
Borgonovo, E., Tarantola, S., Plischke, E., & Morris, M. D. (2014). Transformations and invariance in the sensitivity analysis of computer experiments. *Journal of the Royal Statistical Society, Series B*, *76*, 925–947.
Bowman, A. W., & Azzalini, A. (2003). Computational aspects of nonparametric smoothing with illustrations from the sm library. *Computational Statistics & Data Analysis*, *42*, 545–560.
Breeding, R. J., Helton, J. C., Gorham, E. D., & Harper, F. T. (1992). Summary description of the methods used in the probabilistic risk assessments for NUREG-1150. *Nuclear Engineering and Design*, *135*, 1–27.
Campolongo, F., & Saltelli, A. (1997). Sensitivity analysis of an environmental model: An application of different analysis methods. *Reliability Engineering & System Safety*, *57*(1), 49–69.
Conover, W. J., & Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. *Communications in Statistics - Theory and Methods*, *5*(14), 1349–1368.
Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician*, *35*, 124–133.
Cukier, R. I., Fortuin, C. M., Shuler, K. E., Petschek, A. G., & Schaibly, J. H. (1973). Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I. Theory. *Journal of Chemical Physics*, *59*, 3873–3878.
Da Veiga, S. (2015). Global sensitivity analysis with dependence measures. *Journal of Statistical Computation and Simulation*, *85*(7), 1283–1305.
Davies, P. L., & Kovac, A. (2004). Densities, spectral densities and modality. *Annals of Statistics*, *32*(3), 1093–1136.
De Lozzo, M., & Marrel, A. (2016). New improvements in the use of dependence measures for sensitivity analysis and screening. *Journal of Statistical Computation and Simulation*, *86*(15), 3038–3058.
De Lozzo, M., & Marrel, A. (2017). Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulators. *Stochastic Environmental Research and Risk Assessment*, *31*, 1437–1453.
Devroye, L., & Györfi, L. (1985). *Nonparametric density estimation: The L^1 view*. New York, NY: Wiley.
Eschenbach, T. G. (1992). Spiderplots versus tornado diagrams for sensitivity analysis. *Interfaces*, *22*(6), 40–46.
Felli, J. C., & Hazen, G. B. (1998). Sensitivity analysis and the expected value of perfect information. *Medical Decision Making*, *18*, 95–109.
Frey, H., & Patil, S. (2002). Identification and review of sensitivity analysis methods. *Risk Analysis*, *22*(3), 553–578.
Gamboa, F., Janon, A., Klein, T., Lagnoux, A., & Prieur, C. (2016). Statistical inference for Sobol pick-freeze Monte Carlo method. *Statistics*, *50*(4), 881–902.
Ghanem, R., Higdon, D., & Owhadi, H. (Eds.) (2017). *Handbook of uncertainty quantification*. Cham: Springer Verlag.
Glick, N. (1975). Measurements of separation among probability densities or random variables. *Canadian Journal of Statistics*, *3*(2), 267–276.
Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Proceedings of 16th International Conference on Algorithmic Learning Theory, ALT* (pp. 63–77). Berlin: Springer Verlag.
Helton, J. C. (1993). Uncertainty and sensitivity analyses techniques for use in performance assessment for radioactive waste disposal. *Reliability Engineering & System Safety*, *42*(2–3), 327–367.
Helton, J. C. (1994). Treatment of uncertainty in performance assessments for complex systems. *Risk Analysis*, *14*(4), 483–511.
Helton, J. C. (1999). Uncertainty and sensitivity analysis in performance assessment for the waste isolation pilot plant. *Computer Physics Communications*, *117*(1–2), 156–180.
Helton, J. C., & Breeding, R. J. (1993). Calculation of reactor accident safety goals. *Reliability Engineering and System Safety*, *39*, 129–158.
Helton, J. C., & Davis, F. J. (2002). Illustration of sampling-based methods for uncertainty and sensitivity analysis. *Risk Analysis*, *22*(3), 591–622.
Helton, J. C., & Davis, F. J. (2003). Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety*, *81*(1), 23–69.
Helton, J. C., Hansen, C. W., & Swift, P. N. (2014). Performance assessment for the proposed high-level radioactive waste repository at Yucca Mountain, Nevada. *Reliability Engineering & System Safety*, *122*, 1–6.
Helton, J. C., Johnson, J. D., Sallaberry, C. J., & Storlie, C. B. (2006). Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering & System Safety*, *91*(10–11), 1175–1209.
Helton, J. C., & Marietta, M. G. (2000). The 1996 performance assessment for the Waste Isolation Pilot Plant. *Reliability Engineering & System Safety*, *63*(1–3), 1–3.
Helton, J. C., & Sallaberry, C. J. (2009). Computational implementation of sampling-based approaches to the calculation of expected dose in performance assessments for the proposed high-level radioactive waste repository at Yucca Mountain, Nevada. *Reliability Engineering & System Safety*, *94*(3), 699–721.
Homma, T., & Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, *52*(1), 1–17.
Iman, R. L., & Conover, W. J. (1979). The use of the rank transform in regression. *Technometrics*, *21*(4), 499–509.
Iman, R. L., Helton, J. C., & Campbell, J. E. (1978). *Risk methodology for geologic disposal of radioactive waste: Sensitivity analysis techniques* (SAND-78-0912). Albuquerque (NM): Sandia Labs.
Iman, R. L., & Hora, S. C. (1990). A robust measure of uncertainty importance for use in fault tree system analysis. *Risk Analysis*, *10*(3), 401–406.
Iman, R. L., Johnson, M. E., & Watson, C. C., Jr. (2005). Uncertainty analysis for computer model projections of hurricane losses. *Risk Analysis*, *25*(5), 1299–1312.
Jiménez Rugama, L. A., & Gilquin, L. (2018). Reliable error estimation for Sobol' indices. *Statistics and Computing*, *28*(4), 725–738.
Kleijnen, J. P. C. (2017). Design and analysis of simulation experiments: Tutorial design and analysis of simulation experiments: Tutorial. In *Advances in Modeling and Simulation: Seminal Research from 50 Years of Winter Simulation Conferences* (pp. 135–158). Cham: Springer Verlag.
Kleijnen, J. P. C., & Helton, J. C. (1999a). Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: Review and comparison of techniques. *Reliability Engineering & System Safety*, *65*(2), 147–185.

Kleijnen, J. P. C., & Helton, J. C. (1999b). Statistical analyses of scatterplots to identify important factors in large-scale simulations, 2: Robustness of techniques. *Reliability Engineering & System Safety*, *65*(2), 187–197.

Knuth, D. E. (1997). *The art of computer programming. Vol. 3: Sorting and searching* (2nd ed.). Boston, MA: Addison-Wesley.

Koks, E., Bočkarjova, M., de Moel, H., & Aerts, J. (2015). Integrated direct and indirect flood risk modeling: Development and sensitivity analysis. *Risk Analysis*, *35*(5), 882–900.

Le Gratiet, L., Marelli, S., & Sudret, B. (2017). Metamodel-based sensitivity analysis: Polynomial chaos expansions and Gaussian processes. In *Handbook of uncertainty quantification* (pp. 1289–1325). Cham: Springer Verlag.

Liu, Q., & Homma, T. (2009). A new computational method of a moment-independent uncertainty importance measure. *Reliability Engineering & System Safety*, *94*(7), 1205–1211.

Lyons, R. (2013). Distance covariance in metric spaces. *Annals of Probability*, *41*(5), 3284–3305. Errata: 2018, 46(4), 2400–2405.

Mohanty, S., Codell, R., Wu, Y. T., Pensado, O., Osidele, O., & Esh, D. (2011). *History and value of uncertainty and sensitivity analyses at the nuclear regulatory commission and center for nuclear waste regulatory analyses*. San Antonio, TX: Center for Nuclear Waste Regulatory Analyses.

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

Oakley, J. E. (2010). *Eliciting univariate probability distributions. Rethinking risk measurement and reporting* (Vol. I). London: Risk Books.

OECD (1989). *PSACOIN level E intercomparison. An international code intercomparison exercise on a hypothetical safety assessment case study for radioactive waste disposal systems*. Paris: Nuclear Energy Agency (NEA) of the Organisation for Economic Co-operation and Development (OECD).

Owen, A. B. (2013). Variance components and generalized Sobol' indices. *SIAM/ASA Journal on Uncertainty Quantification*, *1*(1), 19–41.

Patil, S. R., & Frey, H. (2004). Comparison of sensitivity analysis methods based on applications to a food safety risk assessment model. *Risk Analysis*, *24*(3), 573–585.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*, 559–572.

Pearson, K. (1905). *On the general theory of skew correlation and non-linear regression* (Vol. XIV). London: Dulau & Co.

Plischke, E. (2012). How to compute variance-based sensitivity indicators with your spreadsheet software. *Environmental Modelling & Software*, *35*, 188–191.

Plischke, E., Borgonovo, E., & Smith, C. L. (2013). Global sensitivity measures from given data. *European Journal of Operational Research*, *226*(3), 536–550.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, *10*, 441–451.

Riedmann, R. A., Gasic, B., & Vernez, D. (2015). Sensitivity analysis, dominant factors, and robustness of the ECETOC TRA v3, Stoffenmanager 4.5, and ART 1.5 occupational exposure models. *Risk Analysis*, *35*(2), 211–225.

Saltelli, A. (2002a). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297.

Saltelli, A. (2002b). Sensitivity analysis for importance assessment. *Risk Analysis*, *22*(3), 579–590.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & 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., & Marivoet, J. (1990). Non-parametric statistics in sensitivity analysis for model output: A comparison of selected techniques. *Reliability Engineering & System Safety*, *28*(2), 229–253.

Saltelli, A., Ratto, M., Tarantola, S., & Campolongo, F. (2012). Update 1 of: Sensitivity analysis for chemical models. *Chemical Reviews*, *112*, PR1–PR21. https://doi.org/10.1021/cr200301u.

Saltelli, A., & Sobol', I. M. (1995). About the use of rank transformation in the sensitivity analysis of model output. *Reliability Engineering & System Safety*, *50*(3), 225–239.

Saltelli, A., & Tarantola, S. (2002). On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal. *Journal of the American Statistical Association*, *97*(459), 702–709.

Saltelli, A., Tarantola, S., & Chan, K. (1998). Presenting results from model based studies to decision-makers: Can sensitivity analysis be a defogging agent? *Risk Analysis*, *18*(6), 799–803.

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

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, *18*(3), 434–438.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., & Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, *41*(5), 2263–2291.

Sheather, S. J. (2004). Density estimation. *Statistical Science*, *19*(4), 588–597.

Sobol', I. M. (1990). On sensitivity estimation for nonlinear mathematical models (In Russian). *Matematicheskoe Modelirovanie*, *2*(1), 112–118.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

Storlie, C. B., & Helton, J. C. (2008). Multiple predictor smoothing methods for sensitivity analysis: Description of techniques. *Reliability Engineering & System Safety*, *93*, 28–54.

Storlie, C. B., Swiler, L. P., Helton, J. C., & Sallaberry, C. J. (2009). Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliability Engineering & System Safety*, *94*(11), 1735–1763.

Strong, M., & Oakley, J. E. (2013). An efficient method for computing single-parameter partial expected value of perfect information. *Medical Decision Making*, *33*(6), 755–766.

Strong, M., Oakley, J. E., & Chilcott, J. (2012). Managing structural uncertainty in health economic decision models: A discrepancy approach. *Journal of the Royal Statistical Society, Series C*, *61*(1), 25–45.

Sudret, B. (2008). Global sensitivity analysis using polynomial chaos expansion. *Reliability Engineering & System Safety*, *93*, 964–979.

Surjanovic, S., & Bingham, D. (2019). Virtual library of simulation experiments: Test functions and datasets. Retrieved from http://www.sfu.ca/~ssurjano

Székely, G. J., & Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification*, *22*, 151–183.

Székely, G. J., & Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, *4*, 447–479.

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

Tarantola, S., Gatelli, D., & Mara, T. A. (2006). Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering & System Safety*, *91*, 717–727.

Tsanakas, A., & Millossovich, P. (2016). Sensitivity analysis using risk measures. *Risk Analysis*, *36*(1), 30–48.

U.S. Environment Protection Agency. (2009). *Guidance on the development, evaluation, and application of environmental models* (EPA/100/K-09/003). Council for Regulatory Environmental Modeling.

U.S. Nuclear Regulatory Commission. (1990). *Severe accident risks: An assessment for five U.S. nuclear power plants* (Final Summary Report NUREG-1150). Washington, DC: Office of Nuclear Regulatory Research, Division of Systems Research. In 3 Vols.

van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. *3*). Cambridge, MA: Cambridge University Press.

Veraverbeke, N., Gijbels, I., & Omelka, M. (2014). Preadjusted non-parametric estimation of a conditional distribution function. *Journal of the Royal Statistical Society, Series B*, *76*(2), 399–438.

Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, *11*(2), 147–162.

Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering & System Safety*, *142*, 399–432.