# Estimating Global Sensitivity Statistics from Given Data

**Elmar Plischke**[a*], **Emanuele Borgonovo**[b], **Curtis L. Smith**[c]
[a]Institute of Disposal Research, Clausthal University of Technology, Germany
[b]Department of Decision Sciences, Bocconi University, Milan, Italy
[c]Idaho National Laboratory, Idaho Falls, USA

**Abstract:** International agencies explicitly recommend the use of global sensitivity statistics as part of best practice in scientific codes audit and validation. However, estimation of these statistics is a computationally intensive task, tempting analysts to resort to less informative but numerically cheaper methods. This paper introduces a method for the estimation of global sensitivity indices from given data, i.e., at the minimum computational cost. We address the problem with a statistic based on the $L^1$-norm. A formal definition of the estimator is provided and corresponding consistency theorems are provided. The determination of confidence intervals through a bias-reducing bootstrap estimator is investigated. The strategy is applied in the identification of the key-uncertainty drivers of the complex computer code developed at the National Aeronautics and Space Administration (NASA) for the risk assessment of lunar space missions.

**Keywords:** Global Sensitivity Analysis, Density-Based Sensitivity Statistics

## 1. INTRODUCTION

Scientific models as implemented in software packages support analysts and decision-makers in *virtually all areas of science and engineering* [25]. Agencies such as the US Environmental Protection Agency [37], the National Aeronautics and Space Administration [3], the Florida Commission Hurricane Loss Projection Methodology [14] and the European Commission [26, 27] recommend the utilization of global sensitivity methods for best practices of model validation and audit. Identifying which factors *are the most influential, in some sense, in inducing the output uncertainty* [18] becomes a crucial part of uncertainty management when there is a large number of model inputs. In this case, it is of particular relevance for analysts to identify those areas where to focus data collection and/or further modelling efforts. However, a large number of factors is indeed one of the main obstacles to the use of global methods, often termed *the curse of model dimensionality* [23, 24].

In this article, we introduce a design for estimating global sensitivity measures from given samples making the estimation cost independent of the number of factors. In a model input-output framework, we aim at post-processing the dataset generated by a Monte Carlo simulation, without additional model runs.

We focus on the estimation of a recently introduced $L^1$-norm sensitivity measure $\delta$ [1]. First, distribution-based statistics are receiving increasing attention as they are able to overcome limitations of variance-based statistics. Second, well-estimating them is a numerically challenging task and, to date, no strategies for their estimation from given data has been suggested.

The material of this paper is largely taken from [22], submitted, where all mathematical proofs and further discussions can be found.

## 2. ESTIMATING GLOBAL SENSITIVITY STATISTICS

This section offers a short review of numerical aspects in global sensitivity analysis (SA). We start with the global SA frame. Consider $\mathbf{x} = [x_1, x_2, \ldots, x_k] \in \mathcal{X} \subseteq \mathbb{R}^k$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$ related through the

function

$$g : \mathcal{X} \to \mathcal{Y}, \qquad \mathbf{x} \mapsto y. \tag{1}$$

The function $g(\mathbf{x})$ is not necessarily known analytically and is generally the output of a computer code performing a numerical simulation. If an analyst is interested in studying the response of the numerical code at a reference point $\mathbf{x}^0 \in \mathcal{X}$, then she utilizes a local sensitivity method. Conversely, if she is interested in apportioning uncertainty in $y$ to its sources then she needs to explore the response of $y$ as $\mathbf{x}$ spans the entire input space. In a global SA, $\mathbf{x}$ is considered as a random vector $\mathbf{X} = [X_1, X_2, \ldots, X_k]$ on measurable space $(\mathcal{X}, \mathcal{A})$, with $X_i$ on space $(\mathcal{X}^i, \mathcal{A}^i)$. The model output $y$ becomes a random variable $Y$ on $(\mathcal{Y}, \mathcal{B})$. The probability distribution of $X_i$ is denoted by $P_{X_i}(A) = \mathbb{P}(X_i \in A)$, $A \in \mathcal{A}^i$, its distribution function by $F_{X_i}(x) = \mathbb{P}(X_i < x)$, $x \in \mathcal{X}^i$ and its density by $f_{X_i}(\cdot)$. For $Y$, similar notations apply.

Before discussing global sensitivity methods, let us recall the concept of a sensitivity analysis setting [31]. A setting is a *way of framing the sensitivity analysis quest in such a way that the answer can be confidently entrusted to a given method* [30]. The two main settings are *factor prioritization* and *factor fixing*. They correspond to the identification of the most and least relevant factors, respectively.

Several global methods have been developed since the 90's to address these two settings, among them are screening methods [16], non-parametric or regression-based [29, 12], variance-based methods [35, 18], density-based [19, 5, 1, 15] and expected-value-of-information based ones [17]. The common feature of the last three classes of methods is that they are, on the one hand, the most informative in terms of uncertainty appraisal and, on the other hand, the most computationally intensive since these methods use averages over inner statistics.

An example of a function studied in the global SA literature is the Ishigami function [32],

$$Y = \sin X_1 + 7 \sin^2 X_2 + 0.1 X_3^4 \sin X_1 = (1 + 0.1 X_3^4) \sin X_1 + 7 \sin^2 X_2 \tag{2}$$

with $\mathbf{X} = [X_1, X_2, X_3, X_4]$, $\mathcal{X} = (-\pi, \pi)^4$ and $X_i \sim U(-\pi, \pi)$ (independently uniformly distributed). $X_4$ is an additional dummy factor. The first four graphs in Figure 1 show the density $f_Y(y)$ (fat)

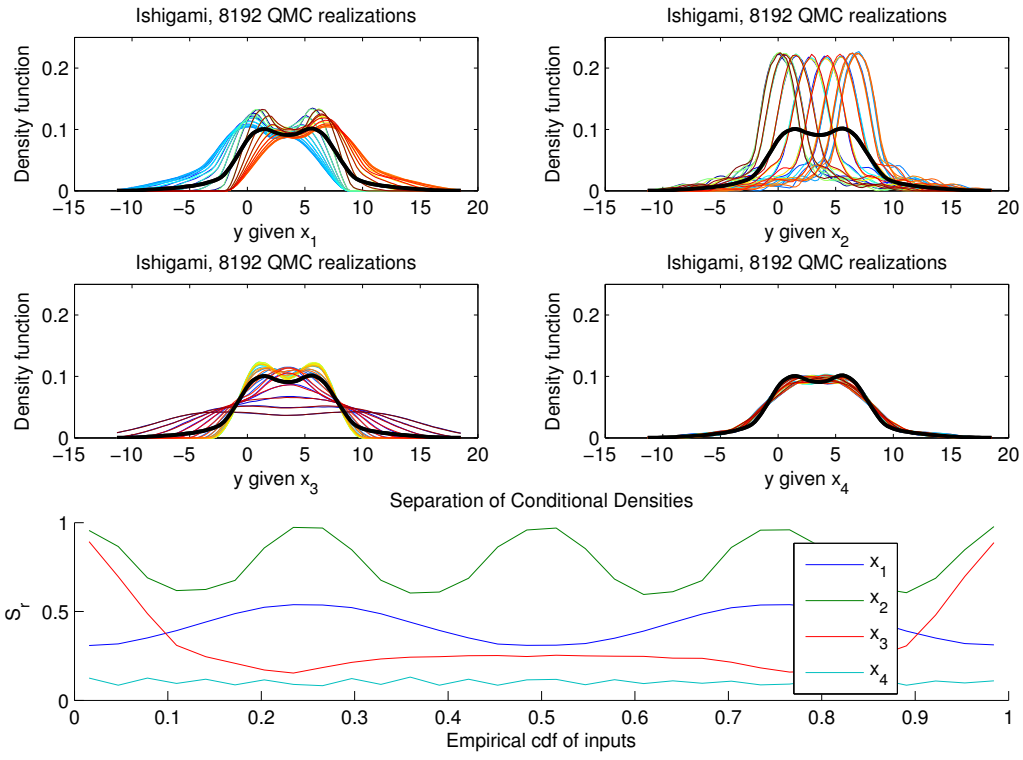


Figure 1: Conditional and unconditional densities of the Ishigami test function.

and the conditional densities $f_{Y|X_i=x_i}(y)$ for $i = 1, 2, 3, 4$. One notes that knowlegde of $X_1$, $X_2$ or $X_3$ leads to evident modifications in the shape of the density of $Y$. However, if an analyst quantifies the contribution of $X_3$ to uncertainty by individual contribution to variance, she would consider $X_3$ as non-influential because the conditional mean $\mathbb{E}[Y|X_3]$ is constant. However, as Figure 1 shows, knowing $X_3 = x_3$ modifies $f_Y(y)$.

## 3. DENSITY-BASED SENSITIVITY METHODS

Let $X : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{A})$ and $Y : (\Omega, \mathcal{F}) \to (\mathcal{Y}, \mathcal{B})$ be two random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $X = x$, then the conditional distribution $F_{Y|X=x}(y)$ represents the decision-maker's new degree-of-belief about $Y$. Measuring the separation between $F_Y(y)$ and $F_{Y|X=x}(y)$ or between $f_Y(y)$ and $f_{Y|X=x}(y)$ is a way to quantify the effect of fixing $X$ at $x$ on the decision-maker's degree-of-belief. We use the $L^1$-norm between densities. The separation is written as

$$s_i(x) = \int_{\mathcal{Y}} \left| f_Y(x) - f_{Y|X=x}(y) \right| \, \mathrm{d}y. \tag{3}$$

By Scheffé's theorem [33, 7] it holds that

$$s_i(x) = \int_{\mathcal{Y}} \left| f_Y(x) - f_{Y|X=x}(y) \right| \, \mathrm{d}y = 2 \sup_{B \in \mathcal{B}} \left| P_Y(B) - P_{Y|X}(B) \right| \tag{4}$$

where the sup operation is extended to all sets $B$ in the algebra $\mathcal{B}$ of $\mathcal{Y}$. Eq. (4) implies that instead of measuring the separation of two distributions utilizing the $L^1$-norm (left hand side) one can equivalently use the variational distance. More specifically, the variational distance in the right-hand side of (4) is a generalization of the Kolmogorov-Smirnov distance $d_{\mathrm{KS}}$ and the discrepancy metric $d_D$ [9],

$$\begin{aligned} d_{\mathrm{KS}} = \sup_{y \in \mathcal{Y}} \left| F_Y(y) - F_{Y|X}(y) \right| \leq d_D = \sup_{\text{all closed balls } A} & \left| P_Y(A) - P_{Y|X}(A) \right| \\ & \leq d_\delta = \sup_{B \in \mathcal{B}} \left| P_Y(B) - P_{Y|X}(B) \right|. \end{aligned} \tag{5}$$

The Kolmogorov-Smirnov distance inspects discrepancy over all half-rays in $\mathcal{Y}$, the discrepancy metric over all closed balls in $\mathcal{B}$, while the variational distance in $d_\delta$ considers all (measurable) sets in $\mathcal{B}$.

In a global SA context, $s_i(x)$ is conditional on $X_i = x$. The lower graph in Figure 1 shows estimates of $s_i(x)$ $(i = 1, \ldots, 4)$ for the Ishigami model. When averaging over the possible values of $s_i$ attained by $X_i$ we obtain the following definition.

**Definition 1** *Given two random variables $X$ and $Y$ on measurable spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, we define the importance of $X$ on $Y$ as*

$$\delta(Y, X) = \tfrac{1}{2}\mathbb{E}[s_i(X)] = \int_{\mathcal{X}} f_X(x) \cdot \sup_{B \in \mathcal{B}} \left| P_Y(B) - P_{Y|X=x}(B) \right| \, \mathrm{d}x. \tag{6}$$

Note that, by Scheffé's theorem, it is

$$\delta(Y, X) = \frac{1}{2} \int_{\mathcal{X}} f_X(x) \int_{\mathcal{Y}} \left| f_Y(y) - f_{Y|X=x}(y) \right| \, \mathrm{d}y \, \mathrm{d}x. \tag{7}$$

Using the definition of a conditional density function, the following symmetry relationship follows easily,

$$\delta(Y, X) = \frac{1}{2} \int_{\mathcal{X} \times \mathcal{Y}} |f_X(x) f_Y(y) - f_{XY}(x, y)| \, \mathrm{d}y \, \mathrm{d}x = \delta(X, Y). \tag{8}$$

Equation (8) shows that $\delta(Y,X)$ provides a way of judging whether $Y$ is dependent on $X$. If $Y$ is independent of $X$ then $\delta = 0$. In fact, by definition of independence between random variables, $f_{XY}(x,y) = f_X(x)f_Y(y)$. Conversely, if some dependence is present, $f_{XY}(x,y) \neq f_X(x)f_Y(y)$, which implies $\delta(Y,X) \neq 0$, regardless of what moment of $Y$ the input $X$ is contributing to. Thus, by $\delta(Y,X)$, one avoids type I errors, i.e. deeming an input factor as unimportant when it is indeed influencing the output.

In the remainder, we explicitly consider $Y$ as the model output in (1) and $\mathbf{X}$ as the random vector of input factors. We fix one input variable $X_i$ of interest. Then, the importance of factor $X_i$ on $Y$ is given by

$$\delta_i = \delta(Y,X_i) = \tfrac{1}{2}\,\mathbb{E}[s_i(X_i)] = \frac{1}{2}\int_{\mathcal{X}^i} f_{X_i}(x)\int_{\mathcal{Y}} \left|f_Y(y) - f_{Y|X_i=x}(y)\right|\,\mathrm{d}y\,\mathrm{d}x. \tag{9}$$

$\delta_i$ possesses additional properties [1, 2], one being normalization: $0 \leq \delta_i \leq 1$, with $\delta_i = 0$ if and only if $Y$ is independent of $X_i$. A second property is scale invariance: Suppose that $z(y)$ is a monotonic function of $Y$. Then, it has been proven that $\delta(Y,X_i) = \delta(z(Y),X_i)$ [2]. By (8), we even have $\delta(Y,X) = \delta(z_1(Y),z_2(X))$ for monotonic maps $z_1, z_2$. A third property is the following [1],

$$\delta_{1,2,\ldots,k} = \frac{1}{2}\mathbb{E}_{\mathbf{X}}\left[\int_{\mathcal{Y}} \left|f_Y(y) - f_{Y|X_1=x_1,X_2=x_2,\ldots,X_k=x_k}(y)\right|\,\mathrm{d}y\right] \tag{10}$$

where $f_{Y|X_1=x_1,X_2=x_2,\ldots,X_k=x_k}(y)$ is a Dirac $\delta-$density centered at $(x_1,x_2,\ldots,x_k)$. By (10), $\delta_{1,2,\ldots,k}$ is the distance between the present state-of-knowledge and the state in which uncertainty in $Y$ is completely resolved. Such distance is unity independently of the point $(x_1,x_2,\ldots,x_k)$ at which the factors are fixed. Then, we gain an additional insight about $\delta_i$: It is the distance towards certainty that we expect to travel by getting to know factor $X_i$.

## 4. ESTIMATORS

Our first step is to set forth a formal definition of the estimators. We rely on the notion of class-conditional densities. Here a class is a sub-sample stemming from a suitable partition of the dataset.

Having fixed a generic factor $i$, we partition $\mathcal{X}^i$ into $M$ classes. The key-intuition of our approach is the following. We replace the density conditional to $X_i = x$ by the conditional density generated by $X_i$ belonging to the class-interval of a suitably chosen partition of $\mathcal{X}^i$. Formally, let $\mathcal{P} = \{\mathcal{C}_m | m = 1,\ldots,M\}$ with $\bigcup_{m=1}^{M}\mathcal{C}_m = \mathcal{X}^i$ and $\mathcal{C}_m \cap \mathcal{C}_{m'} = \emptyset$, $m \neq m'$ denote a partition of $\mathcal{X}^i$ into $M$ classes. The probability of $X_i$ belonging to class $\mathcal{C}_m$ is given by $P_{X_i}(\mathcal{C}_m) = \int_{\mathcal{C}_m} f_{X_i}(x)\,\mathrm{d}x$. By the total probability theorem, the class-conditional density of $Y$ given $\mathcal{C}_m \subset \mathcal{X}^i$ is

$$f_{Y|\mathcal{C}_m}(y) = \frac{\int_{\mathcal{C}_m} f_{Y|X_i=x}(y)f_{X_i}(x)\,\mathrm{d}x}{\int_{\mathcal{C}_m} f_{X_i}(x)\,\mathrm{d}x} = \frac{1}{P_{X_i}(\mathcal{C}_m)}\int_{\mathcal{C}_m} f_{X_iY}(x,y)\,\mathrm{d}x. \tag{11}$$

Then, we call

$$S_m = S(\mathcal{C}_m) = \int_{\mathcal{Y}} \left|f_Y(y) - f_{Y|\mathcal{C}_m}(y)\right|\,\mathrm{d}y \tag{12}$$

the class separation induced by $\mathcal{C}_m \subset \mathcal{X}^i$. Correspondingly, we define an approximation of the distributional-importance of $X_i$ for partition $\mathcal{P}$ as

$$\delta_i^{\mathcal{P}} = \frac{1}{2}\sum_{\mathcal{C}\in\mathcal{P}} S(\mathcal{C})P_{X_i}(\mathcal{C}) = \frac{1}{2}\sum_{m=1}^{M} S_m P_{X_i}(\mathcal{C}_m). \tag{13}$$

We have the following result.

**Theorem 2** *Suppose that $X_i$ has a continuous density on $\mathcal{X}^i$. Consider a series of partitions $\mathcal{P}_j = \{\mathcal{C}_1^j, \ldots, \mathcal{C}_{2^j}^j\}$ of $\mathcal{X}^i$ with $\mathcal{C}_1^0 = \mathcal{X}$, $\mathcal{C}_{2m-1}^j \cup \mathcal{C}_{2m}^j = \mathcal{C}_m^{j-1}$ for $j > 0$ which is finely grained such that $\lim_{i\to\infty} \max_{m=1,\ldots,M} P_{X_i}(\mathcal{C}_m^j) = 0$ and which has positive mass in each class such that for all $j$ and $m$, $P_{X_i}(\mathcal{C}_m^j) > 0$. Then $\lim_{j\to\infty} \delta_i^{\mathcal{P}_j} = \delta_i$.*

Theorem 2 ensures that, as the number of partitions grows, the approximation $\delta_i^{\mathcal{P}}$ tends to $\delta_i$. From Theorem 2 and (12), we are faced with the problem of estimating $f_Y$ and $f_{Y|\mathcal{C}_m}$ for which we use kernel-density [20, 7]. Assume that $\{(x_j, y_j)|j = 1, \ldots, n\}$ is a sample of $n$ pairs of realizations of $X_i$ and $Y$ where we suppress the dependency on factor $i$ in $x$. The estimate $\hat{f}_Y(\cdot)$ is obtained from a kernel-density estimation of all realizations $\{y_j|j = 1, \ldots, n\}$ while $\hat{f}_{Y|\mathcal{C}_m}(\cdot)$ is obtained from a kernel-density estimation of the subset $\{y_j|x_j \in \mathcal{C}_m\}$. For a given kernel $K(\cdot)$ and $m = 1, \ldots, M$ these kernel-density estimates are

$$\begin{aligned} \hat{f}_Y(y) &= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\alpha} K\left(\frac{y - y_j}{\alpha}\right), \\ \hat{f}_{Y|\mathcal{C}_m}(y) &= \frac{1}{n_m}\sum_{x_j \in \mathcal{C}_m} \frac{1}{\alpha_m} K\left(\frac{y - y_j}{\alpha_m}\right). \end{aligned} \tag{14}$$

Here, $n_m = \sum_{x_j \in \mathcal{C}_m} 1$ is the number of realizations in class $\mathcal{C}_m$ of $\mathcal{P}$. With a given set of $\ell$ quadrature points $\{\tilde{y}_j|j = 1, \ldots, \ell\}$, (14) we define the point-wise separation of the estimated densities by

$$s_{m,j} = \hat{f}_Y(\tilde{y}_j) - \hat{f}_{Y|\mathcal{C}_m}(\tilde{y}_j), \qquad j = 1, \ldots, \ell, \quad m = 1, \ldots, M. \tag{15}$$

The numerical integration of these estimates may be performed using the trapezoidal rule, yielding

$$\hat{S}_m = \frac{1}{2}\sum_{j=1}^{\ell-1} \left(|s_{m,j+1}| + |s_{m,j}|\right)\left(\tilde{y}_{j+1} - \tilde{y}_j\right), \qquad m = 1, \ldots, M. \tag{16}$$

**Definition 3** *An estimator of $\delta_i$ on the partition $\mathcal{P} = \{\mathcal{C}_m|m = 1, \ldots, M\}$ with quadrature points $\{\tilde{y}_j|j = 1, \ldots, \ell\}$ is denoted by*

$$\hat{\delta}_i = \frac{1}{2n}\sum_{m=1}^{M} n_m \hat{S}_m. \tag{17}$$

Fixing a factor $i$, we suggest the following program for estimating $\delta_i = \delta(Y, X_i)$.

1. *a)* Setting the dependent variable among the variables of a given dataset, if one is not in a global sensitivity analysis framework;
   or
   *b)* performing a traditional uncertainty analysis, if we are investigating the output of a computer code.
2. Partitioning the dataset to form the classes $\mathcal{C}_m$, $m = 1, 2, \ldots, M$.
3. Approximating the densities conditional to these classes via kernel smoothing, (14).
4. Estimating $\delta_i$ using (16) and (17).

We observe about step 1 that no restrictions apply on the random number generator (simple random, quasi Monte-Carlo or Latin hypercube sampling) used for obtaining the realizations from the random vector $\mathbf{X}$; about step 2 that several partition strategies are available [21]. One way which has been proven effective by the authors is partitioning the data by factor ranks, forming nearly equipopulated classes. Numerical experiments have shown that increasing equipopulated partitions beyond 50 classes has negligible effect on the estimation accuracy. We observe about step 3 that from the knowledge of

the conditional distributions additional information can be extracted from the data. In fact, plotting the unconditional model output density against the conditional densities provides a direct way for assessing whether and how fixing a factor modifies the model output density.

It is a necessity to show that $\hat{\delta}_i$ is a consistent estimator of $\delta_i$ for two reasons. First, if the model allows the adoption of an estimation strategy based on increasing sample sizes, then we need to be ensured that $\hat{\delta}_i \to \delta_i$. Second, if the model allows only a fixed budget of model runs, then we need to assess confidence intervals around $\hat{\delta}_i$ at finite or small sample sizes. When using bootstrap we need to prove that $\hat{\delta}_i$ is consistent for ensuring that the bootstrap estimator is also consistent.

**Theorem 4** *$\hat{\delta}_i$ is a consistent estimator of $\delta_i$, i.e., $\lim_{n,M,\ell\to\infty} \hat{\delta}_i = \delta_i$.*

We note the complementarity of Theorems 2 and 4. Theorem 2 states that a strategy based on partitions leads to a consistent estimator of $\delta_i$, provided that class densities are consistently estimated. Theorem 4 ensures that this is the case if one combines the trapezoidal rule and kernel-density, under the assumptions of convergence of the kernel-density estimators. The latter can be found, for instance, in [7, 11].

## 5. BIAS REDUCTION AND CONFIDENCE BOUNDS

To control undesired numerical influences for a point estimation strategy we propose the following approach. The rationale is to profit from information about the conditional and unconditional distributions of $Y$, yielded by Step 3 of Section 4 and to utilize a statistical test to check whether the difference in $F_Y(y)$ and $F_{Y|\mathcal{C}_m}(y)$ is significant. The estimate $\hat{S}_m$ is ignored, if the contribution of class $\mathcal{C}_m$ is deemed insignificant by the test; otherwise, it is summed in eq. (17). Because its statistic can be related to $\hat{\delta}_i$, the (asymptotic) two-sample Kolmogorov-Smirnov (KS) test [6] is a natural choice for this approach. Let $\hat{F}_Y(y)$ denote the empirical distribution functions of $Y$ and $\hat{F}_{Y|\mathcal{C}_m}(y)$ the class-based empirical conditional distribution function. Then, the contribution of class $\mathcal{C}_m$ to $\hat{\delta}_i$ is insignificant at niveau $\alpha$ if

$$\max_{y\in\mathcal{Y}} \left| \hat{F}_Y(y) - \hat{F}_{Y|\mathcal{C}_m}(y) \right| \leq K_\alpha \sqrt{\tfrac{1}{n} + \tfrac{1}{n_m}} \tag{18}$$

where $K_\alpha$ is the upper $\alpha$-quantile of the Kolmogorov distribution, $n$ is the overall sample size and $n_m$ the subsample size of class $\mathcal{C}_m$. Employing (18) to estimate the KS test statistics adds one additional calculation in Step 4. Using the trapezoidal quadrature rule, (18) is then replaced by

$$\max_{\kappa=1,\ldots,\ell-1} \left| \sum_{j=1}^{\kappa} \left(s_{m,j} + s_{m,j+1}\right)\left(\tilde{y}_{j+1} - \tilde{y}_j\right) \right| \leq 2K_\alpha \sqrt{\tfrac{1}{n} + \tfrac{1}{n_m}}. \tag{19}$$

To set a rejection level $\alpha$ in (18), we can utilize a dummy variable and exploit our knowledge of the fact that it is uninfluential.

For maintaining the confidence assessment in the estimators, the literature offers us two main methods: the bootstrap and the jackknife [34, 10]. To our purposes, the bias-reducing bootstrap estimator of [8] arose as the most efficient one and it is used in the example in Section 6. Let $\widehat{\text{bias}}(\hat{\delta}) = \bar{\delta}^* - \hat{\delta}$, where $\bar{\delta}^*$ is the average of the moment-independent measure estimates derived from bootstrap replicates of the given observations (i.e., drawing a sample of $n$ realizations from the $n$ available observations, with replacement). Then, one obtains the bias-reducing bootstrap estimate of $\delta$:

$$\hat{\hat{\delta}} = \hat{\delta} - \widehat{\text{bias}}(\hat{\delta}) = 2\hat{\delta} - \bar{\delta}^*. \tag{20}$$

By the theory of bootstrap, one knows that $\hat{\hat{\delta}}$ is a consistent estimator of $\delta$, provided that $\hat{\delta}$ is which is ensured by Theorem 4. We can therefore utilize the distribution of $2\hat{\delta} - \bar{\delta}^*$ for assessing confidence in the estimates. This is particularly relevant at small sample sizes.

## 6. UNCERTAINTY MANAGEMENT IN THE DESIGN PHASE OF A LUNAR SPACE MISSION

We now challenge the presented approach by application to the output of a complex computational code for a realistic decision-support model. The code is a simulation model utilized by NASA and the US Idaho National Laboratory for safety assessment in the design phase of the next generation of lunar space missions [3]. This model has been developed by a team of NASA's and Idaho National Laboratories risk experts to corroborate the risk assessment of lunar space missions in accordance with NASA's Risk Assessment Procedures [36]. The model is computationally intensive, with 872 uncertain input factors.

The mission is modelled as an 8-phase process, from launch to orbit around the moon, to astronauts activity on lunar soil to return to earth. For a detailed description of the phases, we refer to [3]. To our purposes, let us evidence that the model is a black-box processing $k = 872$ uncertain input factors. This high number of factors makes it crucial to determine which factors analysts need to focus resources in data collection and further modelling efforts (i.e. identifying areas where to intervene when). We investigate whether this information can be gathered from a dataset generated by quasi-Monte Carlo uncertainty propagation through the model. The sample of the dataset is of size $65536 \times 873$, where $n = 65535$ is the number of realizations and 873 is split into $k = 872$ input factors plus the model output.

Note that, for this model, the cost for estimating $\delta_i$ through a double-design is $872 \cdot n_{int} \cdot n_{ext} + n$ [4]. Even if a low value of $n_{ext}$ were used (say 4 as in [4]), the cost would rapidly become prohibitive. As a reference, at $n_{int} = 65536$, $C \approx 230,000,000$ model runs are required. Similarly, for first order effects $\eta_i^2$, if one assumes independence and utilizes the result in [28] one obtains $C \approx 57,000,000$ model runs at $n_{int} = 65536$. Such high $C$, which is determined by the high number of factors, would impair the identification of key-uncertainty drivers, because of the long computational time and of memory limitations. Conversely, by the proposed approach it is $C = 65536$. The total time required to process the dataset is around 600secs on a personal computer. Figure 2 displays the results of Step 3 of the algorithm proposed in Section 4. It reports the distributions of the model output obtained by conditioning on $X_{748}$, $X_{152}$, $X_{143}$, $X_{713}$, $X_7$ and $X_{88}$ (this is a subset of the 872 factors).

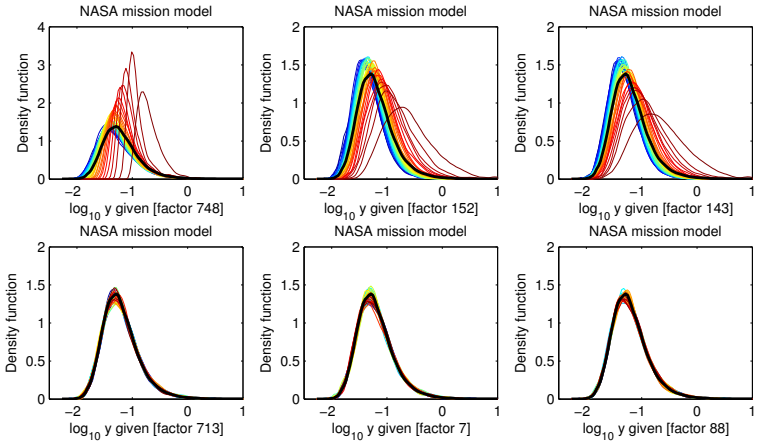


Figure 2: $f_Y(y)$ and $f_{Y|X_{748}}(y)$,$f_{Y|X_{152}}(y)$,$f_{Y|X_{143}}(y)$,$f_{Y|X_{713}}(y)$,$f_{Y|X_7}(y)$,$f_{Y|X_{88}}(y)$ for the output of the NASA space mission model.

Figure 2 allows us to visually appreciate that $Y$ is statistically influenced by factors $X_{748}, X_{152}$ and $X_{143}$ in a much stronger fashion than by factors $X_{713}$, $X_7$, and $X_{88}$.

In Step 4, (see Section 4), $\hat{\delta}_i$ and $\hat{\eta}_i^2$ are determined. We discuss these results in conjunction with the assessment of confidence intervals through the bootstrap estimator (20). Figure 3 shows that at $n = 65536$ the confidence intervals are non overlapping for both the most important and least

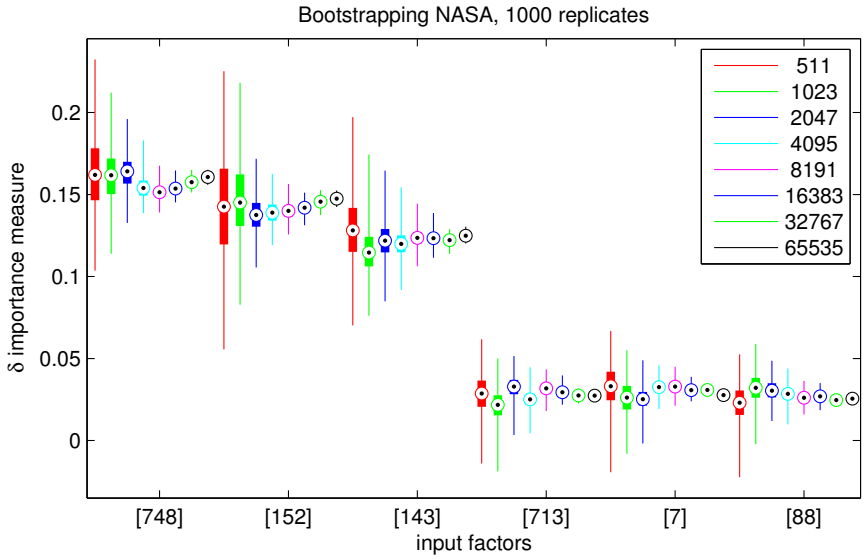


Figure 3: Bootstrapping results for $\hat{\hat{\delta}}$ of $X_{748}$, $X_{152}$, $X_{143}$, $X_{713}$, $X_7$ and $X_{88}$

relevant factors. Thus, information on the key-uncertainty drivers is reliable. At the lowest sample size of $n = 512$ these factors are still identified as the most important ones, although there is a slight overlapping among the distribution of $\hat{\hat{\delta}}$ of $X_{152}$ and both $X_{713}$, $X_7$. However, at $n = 1023$ there is no overlapping. Thus, in a factor fixing setting, already at $n = 1023$ one can conclude that factors $X_{713}$, $X_7$ and $X_{88}$ do not deserve priority when compared to factors $X_{748}$, $X_{152}$, $X_{143}$. The simultaneous estimation of $\delta_i$ and $\eta_i^2$, $i = 1, 2, \cdots, 872$ provides analysts with additional insights. Figure 4 displays $\hat{\delta}_i$ and $\hat{\eta}_i^2$, $i = 1, 2, \ldots, 872$

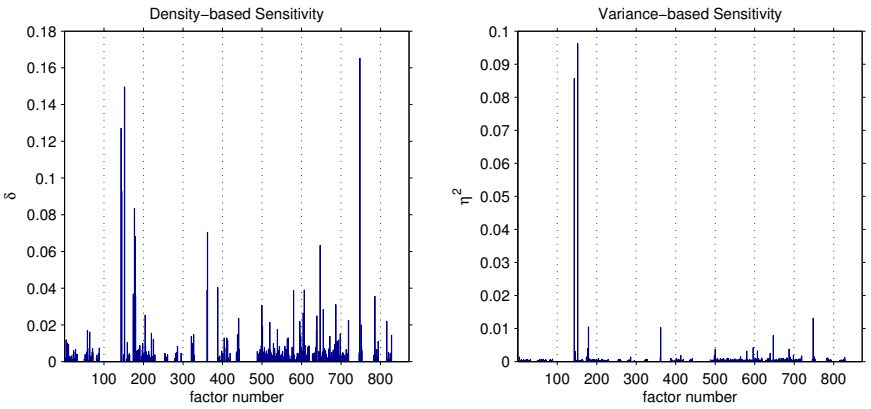


Figure 4: Results of density and variance based sensitivity measures for the NASA space mission model.

Overall, the agreement about the key-drivers of uncertainty between is high, although not perfectly coincident. In particular, the value of the correlation coefficient on ranks is $\rho_{\delta,\eta}^{\text{Rank}} = 0.86$, while the value on the corresponding Savage scores [13] is $\rho_{\delta,\eta}^{\text{SS}} = 0.89$. By construction, these values indicate that the disagreement concentrates mostly on the non-relevant factors. Also, results reveal that 479 variables are associated with null values of both $\hat{\delta}_i$ and $\hat{\eta}_i^2$. To further corroborate this finding, one has available (and can examine) the results of the KS-test filter for all partitions and all factors. The KS-test systematically shows that these factors have no influence on the model output.

Figure 4 shows that factors $X_{143}$, $X_{152}$ and $X_{748}$ are associated with global sensitivity statistics outstanding over the remaining ones both according to $\delta_i$ and $\eta_i^2$. However, $X_{748}$ ranks third with variance-based sensitivity measures, while ranking first with $\delta_i$. The reason is functional dependence and, namely, the presence of interactions. By computing $\sum_{i=1}^{n} \hat{\eta}_i^2$ one understands whether interac-

tions matter in the model response. In our case, it is $\sum_{i=1}^{872} \hat{\eta}_i^2 \approx 0.42$. Thus, individual effects account for around 42% of the model output variance. This difference highlights the active role of interactions in determining the model behaviour. We know that interaction effects are not captured by $\eta_i^2$. $\eta_i^2$ does not account for the importance of $X_{782}$ associated with its interactions with the remaining factors. This finding is in agreement with the analysis of interactions performed by [3] for the same model employing a deterministic design (finite change sensitivity indices). Such design delivers useful information on maintenance and inspection policies, but does not aim at producing information on uncertainty drivers. Indeed, the very low value (0.08) of Savage score correlation between the ranking induced by $\delta_i$ and finite change sensitivity indices confirms the intuition that deterministic methods ought not to be utilized as surrogates of global methods for uncertainty analysis purposes. However, factor $X_{152}$ represents a notable exception. It is ranked among the three most important factors by all methods ($\hat{\delta}$, $\hat{\eta}_i^2$ and the finite change sensitivity indices). This fact suggests that $X_{152}$ indeed deserves priority in further data collection and modelling efforts.

## 7. CONCLUSIONS

This work has presented a new strategy for estimating global sensitivity measures from given data. We have defined new estimators for density-based statistics. Numerical aspects have been analysed, with the introduction of a bias-reduction strategy as well as the determination of confidence bounds through bootstrapping. The method has the following advantages. It allows a notable reduction in computational burden, making the estimation cost independent of the number of factors. Thus, it is appropriate in the factor prioritization and factor fixing settings for models with a high number of inputs.

## REFERENCES


[1] Borgonovo, E. A new uncertainty importance measure. *Reliab. Eng. Syst. Saf. 92*, 6 (2007), 771–784.

[2] Borgonovo, E., Castaings, W., and Tarantola, S. Moment independent importance measures: New results and analytical test cases. *Risk Analysis 31*, 3 (2011), 404–428.

[3] Borgonovo, E., and Smith, C. A study of interactions in the risk assessment of complex engineering systems: An application to space PSA. *Operations Research* (2011). Forthcoming.

[4] Castaings, W., Borgonovo, E., Morris, M., and Tarantola, S. Sampling strategies in density-based sensitivity analysis. *Environmental Modelling&Software* (2012). Submitted.

[5] Chun, M.-H., Han, S.-J., and Tak, N.-I. An uncertainty importance measure using a distance metric for the change in a cumulative distribution function. *Reliab. Eng. Syst. Saf. 70*, 3 (2000), 313–321.

[6] Conover, W. J. *Practical Nonparametric Statistics.* John Wiley&Sons, New York, 1971.

[7] Devroye, L., and Györfi, L. *Nonparametric Density Estimation: The $L^1$ View.* John Wiley&Sons, New York, NY, 1985.

[8] Efron, B., and Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician 37*, 1 (1983), 36–48.

[9] Gibbs, A. L., and Su, F. E. On choosing and bounding probability metrics. *International Statistical Review 70*, 3 (2002), 419–435.

[10] Hall, P. *The bootstrap and Edgeworth expansion.* Springer-Verlag, New York, NY, 1992.

[11] Härdle, W. *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, 1990.

[12] Helton, J. Uncertainty and sensitivity analyses techniques for use in performance assessment for radioactive waste disposal. *Reliab. Eng. Syst. Saf. 42*, 2–3 (1993), 327–367.

[13] Iman, R., and Conover, W. A measure of top-down correlation. *Technometrics 29*, 3 (1987), 351–357.

[14] Iman, R., Johnson, M., and Watson, C.C., J. Sensitivity analysis for computer model projections of hurricane losses. *Risk Analysis 25*, 5 (2005), 1277–1297.

[15] Liu, Q., and Homma, T. A new computational method of a moment-independent importance measure. *Reliab. Eng. Syst. Saf. 94*, 7 (2009), 1205–1211.

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

[17] Oakley, J., Brennan, A., Tappenden, P., and Chilcott, J. Simulation sample sizes for Monte Carlo partial EVPI calculations. *Journal of Health Economics 29*, 3 (2010), 468–477.

[18] Oakley, J. E., and O'Hagan, A. Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J. R. Statist. Soc. B 66*, 3 (2004), 751–769.

[19] Park, C., and Ahn, K. A new approach for measuring uncertainty importance and distributional sensitivity in probabilistic safety assessment. *Reliab. Eng. Syst. Saf. 46*, 3 (1994), 253–261.

[20] Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Statist. 33*, 3 (1962), 1065–1076.

[21] Plischke, E. An adaptive correlation ratio method using the cumulative sum of the reordered output. *Reliab. Eng. Syst. Saf.* (2012). In press. DOI information: 10.1016/j.ress.2011.12.007.

[22] Plischke, E., Borgononovo, E., and Smith, C. L. Estimating global sensitivity statistics from given data: Fighting the curse of dimensionality and an application to a lunar space mission code. Submitted.

[23] Rabitz, H. Systems analysis at the molecular scale. *Science 246* (1989), 221–226.

[24] Rabitz, H., and Alış, Ö. F. General foundations of high-dimensional model representations. *J. Math. Chem. 25*, 2–3 (1999), 197–233.

[25] Sacks, J., Schiller, R., and Welch, W. Designs for computer experiments. *Technometrics 31*, 1 (1989), 41–47.

[26] Saisana, M., Saltelli, A., and Tarantola, S. Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *J. R. Statist. Soc. A 168* (2005), 307–323.

[27] Saltelli, A. Editorial - Special Issue on Sensitivity Analysis. *Reliab. Eng. Syst. Saf. 94*, 7 (2009), 1133–1134.

[28] Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput. Phys. Commun. 181*, 2 (2010), 259–270.

[29] Saltelli, A., and Marivoet, J. Non-parametric statistics in sensitivity analysis for model output: A comparison of selected techniques. *Reliab. Eng. Syst. Saf. 28*, 2 (1990), 229–253.

[30] Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. *Global Sensitivity Analysis – The Primer.* John Wiley&Sons, Chichester, 2008.

[31] Saltelli, A., and Tarantola, S. On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal. *J. Amer. Statist. Assoc. 97*, 459 (2002), 702–709.

[32] Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. *Sensitivity Analysis in Practise – A Guide to Assessing Scientific Models.* John Wiley&Sons, Chichester, 2004.

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

[34] Shao, J., and Tu, D. *The Jackknife and Bootstrap.* Springer-Verlag, New York, NY, 1995.

[35] Sobol', I. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling & Computational Experiments 1* (1993), 407–414.

[36] Stamatelatos, M., Apostolakis, G., Dezfuli, H., Everline, C., Guarro, S., Moieni, P., Mosleh, A., Paulos, T., and Youngblood, R. Probabilistic risk assessent procedures guide for NASA managers and practitioners. Tech. rep., Office of Safety and Mission Assurance, NASA Headquarters, Washington, DC 20546, 2002. http://www.hq.nasa.gov/office/codeq/doctree/praguide.pdf.

[37] US EPA. Guidance on the development, evaluation, and application of environmental models, March 2009. http://www.epa.gov/crem.