Global Sensitivity Analysis in Aircraft Design Process

Marouane FELLOUSSI

Initiation to research project
Data, Economics and Visualisation team - ENAC

January 20, 2021

Abstract

In this paper, the general framework of sensitivity analysis is first introduced. Two indices are presented, Sobol' and Cramér-Von Mises ones. The first classical Sobol' indices are estimated using the Pick-Freeze method, Polynomial Chaos expansions then Rank Statistics. The second type of indices, based on Cramér-Von Mises distances, are more general in the sense that they take into consideration the whole distribution of the output. They can be approximated using either Pick-Freeze or Rank Statistics. The different indices estimators' asymptotic properties are also given. The paper concludes with an application of the different sensitivity analysis tools on the improvement of aircraft's design process.

Keywords: Global sensitivity analysis, Sobol' indices estimation, Pick-Freeze method, Polynomial Chaos expansions, Cramér-Von Mises distance, Chatterjee's coefficient of correlation, Top-level aircraft requirements, Aircraft design process.

1 Introduction

In the aircraft design process, Top Level Aircraft Requirements (TLARs) summarize the expected performance of future aircraft. Some of these requirements can be modeled as constraints in an optimization problem [1], or as design variables, in order to perform sensitivity analysis [2]. More generally, when it comes to the study of computer code experiments, a very classical problem is the evaluation of the relative influence of the input variables on some numerical result obtained by a computer code. Often, the models are expensive to run in terms of computational time [3]. It is thus crucial to understand, within just a few runs, the global influence of one or several inputs of the system under study. When these inputs are regarded as random elements, this problem is generally referred to as Global Sensitivity Analysis (GSA) [4, 5], as opposed to local sensitivity that investigates effects of variations of the input factors in the vicinity of nominal values through gradients or partial derivatives [6]. Such a topic has been widely studied in the last decades and is still challenging nowadays. A classical tool to perform global sensitivity analysis consists in computing the Sobol' indices, first introduced in [7] then formally defined in [8]. The authors use the Hoeffding decomposition [9] for comparing the conditional variance of the output with respect to some inputs with the total variance of the output. In other literature, we can find many estimation procedures of Sobol' indices, namely Monte-Carlo or quasi Monte-Carlo design experiments [10, 11], and also through polynomial chaos expansions [12]. An efficient estimation can be performed through the Pick-Freeze method [13].
Since Sobol' indices are variance-based, they only quantify the influence of the inputs on the mean behavior of the code. Some authors proposed the use of higher moments to define new indices that take into consideration the whole distribution of the output [14], while others took interest in distances between the measures in question [15].

This paper is organized as follows: after introducing the classical Sobol' indices and the associated general framework in section [2], particularly Hoeffding decomposition, the Pick-Freeze method is introduced in section [3] in order to provide an estimator and its asymptotic properties. In section [4], polynomial chaos expansions are presented as well as the construction of the estimators. Next, Cramér-Von Mises indices are defined alongside their estimators and asymptotic properties in section [5]. The last estimator is given in section [6] based on rank statistics. A short numerical application is provided at the end of each section. The [7]-th and last section concludes the article with an application of the different sensitivity analysis tools introduced, on the aircraft design process.

2 General Framework

We consider $\mathbf{X} := (X_i)_{i=1,\dots,p}$ the input variables defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking its values in a product of measurable spaces $E = E_1 \times E_2 \times \ldots \times E_p$ $(p \in \mathbb{N}^*)$. Let $f : E \to \mathbb{R}^k$ $(k \in \mathbb{N}^*)$ be a measurable function, the output denoted by Y is given by:

$$Y = f(X_1, \dots, X_p) \tag{1}$$

Classically, the X_i's are assumed to be independent random variables and Y to be square-integrable (i.e. $\mathbb{E}[\|\mathrm{Y}\|^2] < \infty$). We also assume that Y's co-variance matrix is positive-definite. Let $\mathbf{u}$ be a sub-set of $\mathbf{I}_p := \{1, \dots, p\}$ and $\tilde{\mathbf{u}}$ its complement in $\mathbf{I}_p$. We denote $X_u = (X_i)_{i\in\mathbf{u}}$ and $E_{\mathbf{u}} = \prod_{i\in\mathbf{u}} E_i$.
We can write f as follows :

$$f(\mathbf{X}) = c + f_{\mathbf{u}}(X_u) + f_{\tilde{\mathbf{u}}}(X_{\tilde{u}}) + f_{\mathbf{u},\tilde{\mathbf{u}}}(X_u, X_{\tilde{u}}), \tag{2}$$

where $c \in \mathbb{R}^k$, $f_{\mathbf{u}} : E_{\mathbf{u}} \to \mathbb{R}^k$, $f_{\tilde{\mathbf{u}}} : E_{\tilde{\mathbf{u}}} \to \mathbb{R}^k$ and $f_{\mathbf{u},\tilde{\mathbf{u}}} : E \to \mathbb{R}^k$ are given by:

$$c = \mathbb{E}[Y],\ f_{\mathbf{u}} = \mathbb{E}[Y|X_u] - c,\ f_{\tilde{\mathbf{u}}} = \mathbb{E}[Y|X_{\tilde{u}}] - c,\ f_{\mathbf{u},\tilde{\mathbf{u}}} = Y - f_{\mathbf{u}} - f_{\tilde{\mathbf{u}}} - c \tag{3}$$

Since the terms appearing in the decomposition are orthogonal in L^2, we can compute the co-variance matrix and obtain:

$$\Sigma = C_{\mathbf{u}} + C_{\tilde{\mathbf{u}}} + C_{\mathbf{u},\tilde{\mathbf{u}}}. \tag{4}$$

Here, Σ, $C_{\mathbf{u}}$, $C_{\tilde{\mathbf{u}}}$ and $C_{\mathbf{u},\tilde{\mathbf{u}}}$ are respectively the co-variance matrixes of Y, $f_{\mathbf{u}}(X_u)$, $f_{\tilde{\mathbf{u}}}(X_{\tilde{u}})$ and $f_{\mathbf{u},\tilde{\mathbf{u}}}(X_u, X_{\tilde{u}})$. This decomposition is called the *Hoeffding* decomposition of f. From now on, it is assumed that $k = 1$ (i.e. $Y \in \mathbb{R}$). Dividing the formula (4) by $\Sigma = \mathrm{Var}(Y)$, we get:

$$1 = \frac{C_{\mathbf{u}}}{\Sigma} + \frac{C_{\tilde{\mathbf{u}}}}{\Sigma} + \frac{C_{\mathbf{u},\tilde{\mathbf{u}}}}{\Sigma}$$

Thus

$$1 = \frac{\mathrm{Var}(\mathbb{E}[Y|X_u])}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}(\mathbb{E}[Y|X_{\tilde{u}}])}{\mathrm{Var}(Y)} + \frac{C_{\mathbf{u},\tilde{\mathbf{u}}}}{\mathrm{Var}(Y)} \tag{5}$$

The following concept has been first introduced by I.Sobol in [8]:

Definition 2.0.1 *When $Y \in \mathbb{R}$, we call the quantity $S^{\mathbf{u}} = \frac{\mathrm{Var}(\mathbb{E}[Y|X_u])}{\mathrm{Var}(Y)}$ the closed Sobol' index with respect to the input $X_u = (X_i)_{i\in\mathbf{u}}$.*

Properties 2.0.1 *The Sobol' indices verify the following properties:*
1. The different contributions sum to 1.
2. They are invariant by translation, by any isometry, and by any non degenerated scaling of the components of Y.

The focus will mainly be on the closed index since its knowledge allows us to recover all indices.

3 Estimating first-order Sobol' indices using the Pick-Freeze Monte Carlo method

3.1 Pick-Freeze estimators

In general, the mathematical calculation of the Sobol' indices is nearly impossible, even if the function f is known. It is thus necessary to be able to estimate them. For applications, it is important to be able to estimate simultaneously several indices. For this purpose, let $\mathbf{u} := (u_1, \dots, u_k)$ be k subsets of $\mathbf{I}_p := \{1, \dots, p\}$. We keep the same notations introduced previously. The vector of closed Sobol' indices is then :

$$S^{\mathbf{u}}_{Cl} := \left(\frac{\mathrm{Var}(\mathbb{E}[Y|X_i, i \in u_1])}{\mathrm{Var}(Y)}, \dots, \frac{\mathrm{Var}(\mathbb{E}[Y|X_i, i \in u_k])}{\mathrm{Var}(Y)} \right). \tag{6}$$

The desired quantity to estimate is a fraction whose denominator is fairly easy to estimate. The problem lies within the estimation of the numerator. The idea behind estimating a conditional mathematical expectation is by using co-variance instead [16]. For that, we introduce a new Pick-Freeze variable Y^u, defined as follows: let $\mathbf{X} = (X_1, \dots, X_p)$ and $\mathbf{X}^u = (X^u_1, \dots, X^u_p)$, where $X^u_j = X_j$ if $j \in \mathbf{u}$ and else, X^u_j is an independent copy of X_j. We denote then:

$$Y^u = f(\mathbf{X}^u)$$

Lemma 3.0.1 *Assuming that the variable Y is square-integrable. We have:*

$$\mathrm{Var}(\mathbb{E}[Y|\mathbf{X}^u]) = \mathrm{Cov}(Y, Y^u) \tag{7}$$

In particular

$$S^{\mathbf{u}} = \frac{\mathrm{Cov}(Y, Y^u)}{\mathrm{Var}(Y)}. \tag{8}$$

Let $(X_i)_{i=1,\dots,N}$ an N-sample of $\mathbf{X}$ and $(X^u_i)_{i=1,\dots,N}$ an N-sample of $\mathbf{X}^u$, we set $Y_i = f(X_i)$ and $Y^u_i = f(X^u_i)$. It is then possible to estimate S^u using:

$$S^u_N = \frac{\frac{1}{N}\sum Y_i Y^u_i - (\frac{1}{N}\sum Y_i)(\frac{1}{N}\sum Y^u_i)}{\frac{1}{N}\sum Y_i^2 - (\frac{1}{N}\sum Y_i)^2} \tag{9}$$

We can find this estimator in [16], where it has been shown to have good practical behaviour. The attentive reader will notice however that we have at our disposal two N-samples of the same law, but only used one of the two to estimate $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$. Intuitively, using all of the observations would allow us to come up with a better estimator. For that purpose, the following estimator has been introduced in [17]:

$$T^u_N = \frac{\frac{1}{N}\sum Y_i Y^u_i - (\frac{1}{N}\sum[\frac{Y_i + Y^u_i}{2}])^2}{\frac{1}{N}\sum[\frac{Y_i^2 + (Y^u_i)^2}{2}] - (\frac{1}{N}\sum[\frac{Y_i + Y^u_i}{2}])^2} \tag{10}$$

3.2 Asymptotic properties

Theorem 3.1 (Consistency) *If $\mathbb{E}[Y^2] \leqslant +\infty$ then*

$$S^X_N \text{ and } T^X_N \text{ both converge almost surely to } S^X_N \text{ as } N \text{ goes to infinity.}$$

Theorem 3.2 (Central limit theorem) *If $\mathbb{E}[Y^4] \leqslant +\infty$ then*

$$\sqrt{N}(S^X_N - S^X) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0, \sigma^2_S) \tag{11}$$

where

$$\sigma^2_S = \frac{\mathrm{Var}\left((Y - \mathbb{E}[Y])\left[(Y^X - \mathbb{E}[Y]) - S^X(Y - \mathbb{E}[Y])\right]\right)}{\mathrm{Var}(Y^2)}$$

4 Sensitivity Analysis using Polynomial Chaos expansions

4.1 Introduction

We have seen how Sobol' indices are well-tailored to the case of scalar outputs. Since they are usually computed by Monte Carlo simulation, they are practically not applicable to CPU-demanding models such as finite element models. The formalism of Polynomial Chaos expansions allows obtaining a complete representation of the random response of the model. Their great advantage is that the full randomness of the response is contained in the set of the expansion coefficients. In the following sections, we first will prove that the Sobol' indices of a PC expansion can be computed analytically from the expansion coefficients. To achieve this, we shall recall Sobol' composition first then provide the derivation of Sobol' indices from the polynomial chaos expansion of a model. A regression approach is followed in order to estimate the PC coefficients, that is by minimizing the mean square error of the response approximation in the mean square sense.

4.2 PCE-based Sobol' indices

Let us denote by $\mathcal{M}$ the mathematical model describing the behavior of a system. Let $\mathbf{X} = \{X_1, \ldots, X_M\}$ denote the M-dimensional random, considered real-valued, input vector with joint PDF $f_{\mathbf{X}}$. The output is considered a scalar and is given by:

$$Y = \mathcal{M}(\mathbf{X})$$

The Sobol' decomposition of $\mathcal{M}(\mathbf{X})$ into summands of increasing dimension reads:

$$\mathcal{M}(\mathbf{X}) = \mathcal{M}_0 + \sum_{i=1}^{M} \mathcal{M}_i(X_i) + \sum_{1 \leqslant i < j \leqslant M} \mathcal{M}_{i,j}(X_i, X_j) + \ldots + \mathcal{M}_{1,\ldots,M}(\mathbf{X}) \tag{12}$$

or equivalently, by:

$$\mathcal{M}(\mathbf{X}) = \mathcal{M}_0 + \sum_{\mathbf{u} \neq \emptyset} \mathcal{M}_{\mathbf{u}}(X_{\mathbf{u}}),$$

where $\mathcal{M}_0$ is the mean value of Y, $\mathbf{u} = \{i_1, \ldots, i_s\} \subset \{1, \ldots, M\}$ are index subsets and $X_{\mathbf{u}}$ denotes a subvector of $\mathbf{X}$ containing only those components of which the indices belong to $\mathbf{u}$. The number of summands on the above equation is:

$$\sum_{i=1}^{M} \binom{M}{i} = 2^M - 1$$

Let $D_{\mathbf{X}}$ be the support of the vector $\mathbf{X}$ and f_{X_k} the marginal PDF of random variable X_k. The Sobol' decomposition is unique whenever $\mathcal{M}(\mathbf{X})$ is integrable over the M-dimensional unit cube K^M by choosing summands satsifying the following properties

$$\mathcal{M}_0 = \int_{\mathcal{D}_{\mathbf{X}}} \mathcal{M}(t) f_{\mathbf{X}}(t) dt \tag{13}$$

and

$$\int_{\mathcal{D}_{X_k}} \mathcal{M}_{i_1,\ldots,i_s}(t_{i_1,\ldots,i_s}) f_{X_k}(t_k) dt_k = 0 \tag{14}$$

where

$$K^M = \{X : 0 \leqslant X_i \leqslant 1, i = 1, \ldots, M\}$$

Leading to the orthogonal property

$$\mathbb{E}\left[\mathcal{M}_{\mathbf{u}}(X_{\mathbf{u}}) \mathcal{M}_{\mathbf{v}}(X_{\mathbf{v}})\right] = 0 \; if \; \mathbf{u} \neq \mathbf{v}. \tag{15}$$

The uniqueness and orthogonality properties allow decomposition of the variance D of Y as

$$D = \mathrm{Var}\left[\mathcal{M}(\mathbf{X})\right] = \sum_{\mathbf{u}\neq\emptyset} D_{\mathbf{u}}, \tag{16}$$

where $D_{\mathbf{u}}$ denotes the partial variance

$$D_{\mathbf{u}} = \mathrm{Var}\left[\mathcal{M}_{\mathbf{u}}(X_{\mathbf{u}})\right] = \mathbf{E}\left[\mathcal{M}^2_{\mathbf{u}}(X_{\mathbf{u}})\right] \tag{17}$$

The Sobol' index $S_{\mathbf{u}}$ is defined as

$$S_{\mathbf{u}} := D_{\mathbf{u}}/D, \tag{18}$$

By definition, $\sum_{\mathbf{u}\neq\emptyset} \mathcal{S}_{\mathbf{u}} = 1$. The total sensitivity indices S_i^{Tot} is per usual given by:

$$S_i^{\mathrm{Tot}} = \sum_{\mathcal{I}_i} D_{\mathbf{u}}/D, \quad \mathcal{I}_i = \{\mathbf{u} \supset i\}$$

Evaluation of Sobol' indices by Monte Carlo simulation is based on a recursive relationship which requires computing 2^M Monte Carlo integrals involving $\mathcal{M}(\mathbf{X})$. This is clearly not affordable when the computational model is a time-consuming algorithmic sequence. On the other hand, when PCE of the quantity of interest are available, Sobol' indices can be obtained analytically at almost no additional computational cost.

4.3 Computation of Polynomial Chaos expansions

A PCE approximation of $Y = \mathcal{M}(\mathbf{X})$ has the form introduced by Xiu & Kaniadakis in [18] :

$$\widehat{Y} = \mathcal{M}^{\mathrm{PCE}}(\mathbf{X}) = \sum_{\alpha\in\mathcal{A}} Y_\alpha \Psi_\alpha\left(\mathbf{X}\right) \tag{19}$$

where $\{\Psi_\alpha,\ \alpha \in \mathcal{A}\}$ is a set of multivariate polynomials that are orthonormal with respect to $f_{\mathbf{X}}$, with multi-indices $\alpha = \{\alpha_1, \ldots, \alpha_M\}$ and S_α denoting the corresponding polynomial coefficients.

The multivariate polynomials that comprise the PCE basis are obtained by tensorizations of appropriate univariate polynomials,

$$\Psi_\alpha(\mathbf{X}) = \prod_{i=1}^{M} \psi_{\alpha_i^{(i)}}(X_i), \tag{20}$$

where $\psi_{\alpha_i^{(i)}}(X_i)$ is a polynomial of degree α_i in the i-th input variable belonging to a family of polynomials that are orthonormal with respect to f_{X_i}. For standard distributions, the associated family of orthonormal polynomials is well-known, for example, a uniform variable with support $[-1, 1]$ is associated with the family of Legendre polynomials. The general case can be treated through an isoprobabilistic transform of $\mathbf{X}$ to a basic random vector, this will not be the case in the present study as we will later take interest in uniformly distributed variables. The set of multi-indices $\mathcal{A}$ is determined by an appropriate truncation scheme. We will use a truncation scheme selecting all multi-indices satisfying

$$||\alpha||_q = \left(\sum_{i=1}^{M} \alpha_i^q\right)^{\frac{1}{q}} \leqslant p \tag{21}$$

where $0 < q \leqslant 1$ and p is the maximal degree of the polynomial expansion that is selected. In this paper, we will take $q = 1$. For more details on how to appropriately select q and p, as well as for the construction of the multi-index sequence, see [12].

The coefficients of the of PC-expansions may be calculated by determining the L^2-projection of the response Y onto the subspace spanned by the basis polynomials $\{\Psi_\alpha : \alpha \in \mathbb{N}^M, |\alpha| \leqslant p\}$. Assuming the following expression for a scalar response quantity Y:

$$Y = \mathcal{M}(\mathbf{X}) = \widetilde{Y}(\mathbf{X}) + \epsilon \tag{22}$$

$$\widetilde{Y}(\mathbf{X}) = \sum_{j=0}^{P-1} Y_j \Psi_j(\mathbf{X})) \tag{23}$$

where $P = \frac{(M+p)!}{M!p!}$ is the number of terms after which the series is truncated. Leading to the following problem, of minimizing the mean-square error of the approximation over a set of realizations of the input vector

$$\begin{aligned} \mathbf{y} &= \operatorname{argmin} \mathbb{E}\left[(\mathcal{M}(\mathbf{X}) - \tilde{Y}(\mathbf{X}))^2\right] \\ &= \operatorname*{argmin}_Y \frac{1}{N} \sum_{i=1}^{N} \left\{ \mathcal{M}(X^i) - \sum_{j=0}^{P-1} Y_j \Psi_j(X^i) \right\}^2 \end{aligned} \tag{24}$$

Denoting by Ψ the matrix whose coefficients are given by:

$$\Psi_{ij} = \Psi_j(X^i), i = 1, \ldots, N \text{ and } j = 0, \ldots, P-1$$

and by Y_{ex} the vector containing the exact response values computed by the model $Y_{ex} = \{f(X^i), i = 1, \ldots, M\}$, the solution to (24) reads:

$$\mathbf{y} = (\Psi^T \Psi)^{-1}.\Psi^T.\mathrm{Y}_{ex} \tag{25}$$

where $\Psi^T\Psi$ is called the information matrix. Computationally speaking, it may be ill-conditioned. For this purpose, the mean-square minimization problem will be solved computationally later on in the numerical study.

4.4 PCE-based Sobol' indices

In this subsection, the input parameters are supposed to be uniformly distributed in $[0, 1]$. As seen previously, the Legendre polynomials are orthogonal with respect to the uniform probability measure over $[-1, 1]$. Thus, the Legendre chaos will be used. Denoting by $\{P_n(X), n \in \mathbb{N}\}$ the family of univariate Legendre polynomials, the multivariate Legendre polynomial is given by

$$\Psi_j(\mathbf{X}) = \prod_{i=1}^{M} P_{\alpha_i}(X_i)$$

where α is the multi-indice sequence introduced earlier.

Let us now consider $\widehat{Y} = \mathcal{M}^{\mathrm{PCE}}(\mathbf{X})$, the PCE of the quantity of interest $Y = \mathcal{M}^{\mathrm{PCE}}(\mathbf{X})$. We have:

$$\mathcal{M}^{\mathrm{PCE}}(\mathbf{X}) = \sum_{j=0}^{P-1} Y_j \Psi_j(\mathbf{X}) \quad , \quad \mathbf{X} \sim \mathcal{U}([-1, 1])^M \tag{26}$$

Let us define by $\mathcal{I}^+_{i_1,\ldots,i_s}$ the set of α multi-indices such that only the indices $(i_1, \ldots, i_s)$ are non-zero:

$$\mathcal{I}^+_{i_1,\ldots,i_s} = \left\{ \alpha \in \mathbb{N}^M : \begin{array}{ll} \alpha_k > 0 & k \in (i_1, \ldots, i_s) \\ \alpha_k = 0 & k \notin (i_1, \ldots, i_s) \end{array} \right\}$$

Note that $\mathcal{I}_i^+$ corresponds to the polynomials depending only on parameter X_i. We can now gather the P terms in (26) corresponding to the polynomials according to the parameters they depend on:

$$\begin{aligned}\mathcal{M}^{\text{PCE}}(\mathbf{X}) = Y_0 + \sum_{i=1}^{M} \sum_{\alpha \in \mathcal{I}_i^+} Y_\alpha \Psi_\alpha(X_i) + \sum_{1 \leqslant i_1 < i_2 \leqslant M} \sum_{\alpha \in \mathcal{I}_{i_1,i_2}^+} Y_\alpha \Psi_\alpha(X_{i_1}, X_{i_2}) + \dots \\ + \sum_{1 \leqslant i_1 < \dots < i_s \leqslant M} \sum_{\alpha \in \mathcal{I}_{i_1,\dots,i_s}^+} Y_\alpha \Psi_\alpha(X_{i_1}, \dots, X_{i_s}) + \dots + \sum_{\alpha \in \mathcal{I}_{i_1,\dots,M}^+} Y_\alpha \Psi_\alpha(\mathbf{X})\end{aligned} \tag{27}$$

The statistical moments of the response PC-expansion be analytically derived from its coefficients. In particular, and due to the orthogonality of the basis, the mean and the variance, respectively, read:

$$\begin{aligned}\widetilde{Y} &= \mathbb{E}\left[\mathcal{M}(\mathbf{X})\right] = Y_0 \\ D_{\text{PCE}} &= \text{Var}\left[\sum_{j=0}^{P-1} Y_j \Psi_j(\mathbf{X})\right] \\ &= \sum_{j=0}^{P-1} Y_j^2 \mathbb{E}\left[\Psi_j^2(\mathbf{X})\right]\end{aligned} \tag{28}$$

Lastly, the orthonormality of the PC basis implies that the random summands on the right-hand side of (27) satisfy the properties (13) and (14). Assuming that the function $\mathcal{M}(\mathbf{X})$ is square-integrable with respect to the probability measure associated with $f_{\mathbf{X}}$, it's possible to uniquely identify each summand in (12) as follows:

$$\mathcal{M}_{i_1,\dots,i_s}(X_{i_1}, \dots, X_{i_s}) = \sum_{\alpha \in \mathcal{I}_{i_1,\dots,i_s}^+} Y_\alpha \Psi_\alpha(X_{i_1}, \dots, X_{i_s}) \tag{29}$$

It is now easy to derive sensitivity indices from the above representation. The *Polynomial Chaos-based Sobol' indices*, denoted by $SU_{i_1,\dots,i_s}$, are given by

$$SU_{i_1,\dots,i_s} = \frac{1}{D_{\text{PCE}}} \sum_{\alpha \in \mathcal{I}_{i_1,\dots,i_s}^+} Y_\alpha^2 \mathbb{E}\left[\Psi_\alpha^2\right] \tag{30}$$

To retrieve first-order indices, it suffices to consider a singleton $\mathcal{I}_i^+$
The total sensitivity indices are also easy to compute. For a given integer sequence $(j_1, \dots, j_t)$, let us define the following set:

$$\mathcal{J}_{(j_1,\dots,j_t)} = \{(i_1, \dots, i_s),\ (j_1, \dots, j_t) \subset (i_1, \dots, i_s)\}$$

The *total PC-based sensitivity indices* read:

$$SU_{j_1,\dots,j_s}^T = \sum_{(i_1,\dots,i_s) \in \mathcal{J}_{(j_1,\dots,j_t)}} SU_{i_1,\dots,i_s}$$

Note: The previous sections show that, once the polynomial chaos representation of a model is available, a full list of Sobol' indices is available analytically with almost no additional cost. Indeed, only elementary mathematical operations are needed to compute these indices from the expansion coefficients.

4.5 Application Examples

Let us consider the so-called Sobol' function:

$$Y = \prod_{i=1}^{q} \frac{|4X_i - 2| + a_i}{1 + a_i} \tag{31}$$

where the input variables X_i, $i = 1, \ldots, q$ are uniformly distributed over [0,1] and a_i's are non negative. The variance D of Y and the first-order Sobol' sensitivity indices can be computed analytically:

$$D = \prod_{i=1}^{q}(D_i + 1) - 1, \qquad D_i = \frac{1}{3(1 + a_i)^2}$$

$$SU_{i_1,\ldots,i_s} = \frac{1}{D}\prod_{i=1}^{q} D_i$$

Due to its complexity(non-linear and non-monotonic correlations) and the analytical expression of the Sobol' indices, the Sobol' g-function is a classical text example commonly used in sensitivity analysis. For numerical application, we take $q = 8$ together with $a = [1, 2, 5, 10, 20, 50, 100, 500]$ and $N = 500$ regression points. As expected from the formula above, the lower the coefficient a_i, the more significant the variable X_i.

Index	Analytical Solution	PC-Based Solution
SU_1	0.6037	0.6031
SU_2	0.2683	0.2712
SU_3	0.0671	0.0527
SU_4	0.0200	0.0171
SU_5	0.0055	0.0051
SU_6	0.0009	0.0039
SU_7	0.0002	0.0001
SU_8	0.0000	0.0001

Table 1: g-Sobol' function, Analytical and PC-based Sobol' indices(p=2)

5 Sensitivity Analysis based on Cramér-Von Mises distances

5.1 Introduction

The main drawback of the Sobol' indices and their Monte-Carlo estimation is that they are order two methods since they derive from the L^2-Hoeffding functional decomposition, so they only take into account the second-order behavior which could hide the different contributions in some models. A new index was introduced in [19], based on the Cramér-Von Mises distance between the distribution of the output Y and its conditional law when the input is fixed. This leads to natural self-normalized indices. These indices take into account the whole output distribution instead of only the order two moments. Additionally, and in contrary to most of the other known indices, they are defined for multivariate outputs and thus, are well-tailored to perform sensitivity analysis.

5.2 Building the index based on Cramér-Von Mises distances

In this section, let $Z = f(X_1, \ldots, X_p) \in \mathbb{R}^k$ be the output of the numerical code and F be the cumulative distribution function of Z:

$$F(t) = \mathbb{P}(Z \leqslant t) = \mathbb{E}[\mathbb{1}_{\{Z \leqslant t\}}], \text{ for } t = (t_1, \ldots, t_k) \in \mathbb{R}^k \tag{32}$$

F^u denotes the conditional cumulative distribution function of Z given the u-th input X_u and is given by

$$F^u(t) = \mathbb{P}(Z \leqslant t | X_u) = \mathbb{E}[\mathbb{1}_{\{Z \leqslant t\}} | X_u], \text{ for } t = (t_1, \ldots, t_k) \in \mathbb{R}^k \tag{33}$$

where $\{Z \leqslant t\} = \{Z_1 \leqslant t_1, \ldots, Z_k \leqslant t_k\}$. By the law of total expectation and definition of mathematical expectation relative to indicator functions, We have:

$$\mathbb{E}[F^u(t)] = \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{Z \leqslant t\}} | X_u]] = \mathbb{E}[\mathbb{1}_{\{Z \leqslant t\}}] = F(t). \tag{34}$$

Let us set $Y(t) = \mathbb{1}_{\{Z \leqslant t\}}$. Since for any fixed $t \in \mathbb{R}^k$, $Y(t)$ is a real-valued random variable, we can apply the framework presented previously. More precisely, for any $u \in \mathbf{I}_p$, let $\tilde{u} \in \mathbf{I}_p \backslash \{u\}$. Applying Hoeffding decomposition on $Y(t)$ yields:

$$Y(t) = \mathbb{1}_{\{Z \leqslant t\}} = \mathbb{E}[Y(t)] + (\mathbb{E}[Y(t)|X_u] - \mathbb{E}[Y(t)]) + (\mathbb{E}[Y(t)|X_{\tilde{u}}] - \mathbb{E}[Y(t)]) + R(t, u) \tag{35}$$

where

$$R(t, u) = Y(t) - \mathbb{E}[Y(t)] - (\mathbb{E}[Y(t)|X_u)] - \mathbb{E}[Y(t)]) - (\mathbb{E}[Y(t)|X_{\tilde{u}})] - \mathbb{E}[Y(t)]);$$

Computing the variance of both sides in the previous equation leads to:

$$\begin{aligned} \mathrm{Var}(Y(t)) &= F(t)(1 - F(t)) \\ &= \mathrm{Var}\left(\mathbb{E}[Y(t)|X_u] - \mathbb{E}[Y(t)]\right) + \mathrm{Var}\left(\mathbb{E}[Y(t)|X_{\tilde{u}}] - \mathbb{E}[Y(t)]\right) + \mathrm{Var}(R(t, v)) \\ &= \mathrm{Var}(F^u(t)) + \mathrm{Var}(F^{\tilde{u}}(t)) + \mathrm{Var}(R(t, v)) \\ &= \mathbb{E}[(F^u(t) - F(t))^2] + \mathbb{E}[(F^{\tilde{u}}(t) - F(t))^2] + \mathrm{Var}(R(t, v)). \end{aligned} \tag{36}$$

The second order Cramér-Von Mises distance between the two empirical distributions, $\mathcal{L}(Z)$ and $\mathcal{L}(Z|X_u)$, is defined as:

$$\int_{\mathbb{R}^k} \mathbb{E}\left[(F^u(t) - F(t))^2\right] dF(t). \tag{37}$$

Integrating the terms in (36) with $t \in \mathbb{R}^k$ and with respect to the distribution of Z gives:

$$\begin{aligned} &\int_{\mathbb{R}^k} F(t)(1 - F(t)) dF(t) \\ &= \int_{\mathbb{R}^k} \mathbb{E}\left[(F^u(t) - F(t))^2\right] dF(t) + \int_{\mathbb{R}^k} \mathbb{E}\left[(F^{\tilde{u}}(t) - F(t))^2\right] dF(t) + \int_{\mathbb{R}^k} \mathrm{Var}(R(t, v)) dF(t) \end{aligned} \tag{38}$$

Following the classical way of defining Sobol' indices, we normalize the previous equation by

$$\int_{\mathbb{R}^k} F(t)(1 - F(t)) dF(t)$$

which leads to

$$1 = \frac{\int_{\mathbb{R}^k} \mathbb{E}\left[(F^u(t) - F(t))^2\right] dF(t)}{\int_{\mathbb{R}^k} F(t)(1 - F(t)) dF(t)} + \frac{\int_{\mathbb{R}^k} \mathbb{E}\left[(F^{\tilde{u}}(t) - F(t))^2\right] dF(t)}{\int_{\mathbb{R}^k} F(t)(1 - F(t)) dF(t)} + \frac{\int_{\mathbb{R}^k} \mathrm{Var}(R(t, v)) dF(t)}{\int_{\mathbb{R}^k} F(t)(1 - F(t)) dF(t)}$$

Allowing us to define the Cramér-Von Mises indices with respect to u and $\tilde{u}$ as the authors in [19] defined them:

$$S^u_{2,CVM} = \frac{\int_{\mathbb{R}^k} \mathbb{E}\left[(F^u(t) - F(t))^2\right] dF(t)}{\int_{\mathbb{R}^k} F(t)(1-F(t))dF(t)} \quad \text{and} \quad S^{\tilde{u}}_{2,CVM} = \frac{\int_{\mathbb{R}^k} \mathbb{E}\left[(F^{\tilde{u}}(t) - F(t))^2\right] dF(t)}{\int_{\mathbb{R}^k} F(t)(1-F(t))dF(t)} \tag{39}$$

These indices verify the classical Sobol' indices properties as seen in 2.0.1.

Remark
In this paper, only first-order sensitivity indices are considered. However, one may define higher order and total Cramér-Von Mises indices in a similar way to Sobol' ones. It suffices to take $\mathbf{u}$ as any desired subset, $\tilde{\mathbf{u}}$ being its complement in $\mathbf{I}_p$, and use the previous formula. The total Cramér-Von Mises index $S^{Tot,\mathbf{u}}_{2,CVM}$ is defined as:

$$S^{Tot,\mathbf{u}}_{2,CVM} := 1 - S^{\tilde{\mathbf{u}}}_{2,CVM} = 1 - \frac{\int_{\mathbb{R}^k} \mathbb{E}\left[(F^{\tilde{\mathbf{u}}}(t) - F(t))^2\right] dF(t)}{\int_{\mathbb{R}^k} F(t)(1-F(t))dF(t)}. \tag{40}$$

Note 5.2.1 If the coordinates of the output Z are independent and are absolutely continuous with respect to the Lebesgue measure, the normalizing factor is reduced as in the formula below:

$$\int_{\mathbb{R}^k} F(t)(1-F(t))dF(t) = \frac{1}{2^k} - \frac{1}{3^k} \quad , (k \in \mathbb{N}^*) \tag{41}$$

5.3 General index estimation

Denoting the numerator of $S^u_{2,CVM}$ by $N^u_{2,CVM}$, we can rewrite it as :

$$N^u_{2,CVM} = \mathbb{E}_{\widetilde{Z}}\left[\mathbb{E}_{X_u}\left[(F^u(\widetilde{Z}) - F(\widetilde{Z}))^2\right]\right]$$

where $\widetilde{Z}$ is an independent copy of Z. We then proceed to estimate following a double Monte-Carlo scheme as detailed below:

1. We generate a Pick-Freeze sample of Z: two N-samples $(Z^{u,1}_j, Z^{u,2}_j), 1 \leqslant j \leqslant N$;
2. We create a third N-sample from Z, independent of $(Z^{u,1}_j, Z^{u,2}_j)$: $W_k, 1 \leqslant k \leqslant N$.

For the sake of clarity, let $\mathcal{A}^i_{j,k}$ denote the event $\{Z^{u,i}_j \leqslant W_k\}$. The empirical estimator of $N^u_{2,CVM}$ is thus given by:

$$\widehat{N}^u_{2,CVM} = \frac{1}{N}\sum_{k=1}^{N}\left\{\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}_{\mathcal{A}^1_{j,k}}\mathbb{1}_{\mathcal{A}^2_{j,k}} - \left[\frac{1}{2N}\sum_{j=1}^{N}\left(\mathbb{1}_{\mathcal{A}^1_{j,k}} + \mathbb{1}_{\mathcal{A}^2_{j,k}}\right)\right]^2\right\}. \tag{42}$$

It remains now to estimate the denominator $D^u_{2,CVM}$. It can be rewritten as such:

$$D^u_{2,CVM} = \mathbb{E}[F(Z)(1-F(Z))]$$

Using the exact same procedure as above, it is estimated by:

$$\widehat{D}^u_{2,CVM} = \frac{1}{N}\sum_{k=1}^{N}\left\{\frac{1}{2N}\sum_{j=1}^{N}\left(\mathbb{1}_{\mathcal{A}^1_{j,k}} + \mathbb{1}_{\mathcal{A}^2_{j,k}}\right) - \left[\frac{1}{2N}\sum_{j=1}^{N}\left(\mathbb{1}_{\mathcal{A}^1_{j,k}} + \mathbb{1}_{\mathcal{A}^2_{j,k}}\right)\right]^2\right\}. \tag{43}$$

5.4 Asymptotic properties

Lemma 5.0.1 *$\widehat{N}^u_{2,CVM}$ is strongly consistent as N goes to infinity.*

Theorem 5.1 *The sequence of estimators $\widehat{N}^u_{2,CVM}$ is asymptotically Gaussian in estimating $N^u_{2,CVM}$. That is, $\sqrt{N}\left(\widehat{N}_{2,CVM} - N^u_{2,CVM}\right)$ converges in distribution towards the centered Gaussian law with limiting variance that can be computed.*

Lemma 5.1.1 *$\widehat{S}^u_{2,CVM}$ is strongly consistent as N goes to infinity.*

Theorem 5.2 *The sequence of estimators $\widehat{S}^u_{2,CVM}$ is asymptotically Gaussian in estimating $S^u_{2,CVM}$. That is, $\sqrt{N}\left(\widehat{S}_{2,CVM} - S^u_{2,CVM}\right)$ converges in distribution towards the centered Gaussian law with limiting variance ξ^2, whose explicit expression can be found in [19].*

Remark Considering a sample with an appropriate size, we can estimate both Cramér-Von Mises and Sobol' indices. More precisely, estimating p Sobol' indices requires a sample size of $(p+1)N$. Only N more output evaluation are required to get the Cramér-Von Mises ones. Confidence intervals controlling the accuracy of the estimations are provided by the theorems. This makes the use of these easy-to-implement indices quite efficient.

5.5 Numerical Application

Let us consider the following linear model

$$Y = \alpha X_1 + X_2,\ \alpha > 0,$$

where X_1 is a Bernoulli random variable with success probability $0 < p < 1$ and X_2 is a random variable independent of X_1. Let us further assume that X_2 has a continuous distribution F on $\mathbb{R}$ with $\mathbb{E}[X_2] = \alpha p$ and is of finite variance $\mathrm{Var}(X_2) = \alpha^2 p(1-p)$. This way, the random variables αX_1 and X_2 share the same expectation and variance, and thus the same first order Sobol' indices value which is equal to $1/2$.

The aim of such a construction is to point at the inability of the classical Sobol' indices to detect differences in influence in some cases, while the Cramér-Von Mises ones permit that since they take into consideration the whole distribution of the output as we've mentioned earlier.

5.5.1 General closed formula

First, since X_1 is a Bernoulli random variable, thus taking values as either 0 or 1, we take interest in the distributions of Y given $\{X_1 = 0\}$ and of Y given $\{X_1 = 1\}$

$$\begin{cases} \mathcal{L}(Y|X_1 = 0) = \mathcal{L}(X_2) \\ \mathcal{L}(Y|X_1 = 1) = \mathcal{L}(X_2 + \alpha) \end{cases}$$

The conditional distribution of Y given X_2 is

$$\mathbb{P}(Y = \alpha X_1 + X_2 | X_2) = 1 - \mathbb{P}(Y = X_2 | X_2) = p$$

Thus, the distribution function of Y is

$$pF(\cdot - \alpha) + (1-p)F(\cdot) \tag{44}$$

It remains now to compute $S^1_{2,CVM}$ and $S^2_{2,CVM}$. They are given by

$$S^1_{2,CVM} = 6p(1-p)\int_{\mathbb{R}} \left(F(t) - F(t-\alpha)\right)^2 \left[(1-p)dF(t) + pdF(t-\alpha)\right] \tag{45}$$

And

$$S^2_{2,CVM} = 1 - 6p(1-p)\left[\frac{1}{2} - \int_{\mathbb{R}} F(t-\alpha)dF(t)\right] \tag{46}$$

(the 6 is the inverse of the normalizing factor, refer to note 5.2.1 with $k = 1$). Intuitively, as $p \to 0$, we'd be able to tell that X_2 has more influence than X_1. However, the Sobol' indices remain equal to $\frac{1}{2}$, whereas $(S^1_{2,CVM}, S^2_{2,CVM}) \to (0, 1)$.
The following numerical illustration has a sample size of $N = 1000$. The variable X_2 is uniformly distributed in $[0, 3/4]$ in order to guarantee αX_1 and X_2 share the same expectation and variance. The two indices dependence on $p(1-p)$ results in extreme values taken for $p = 1/2$.

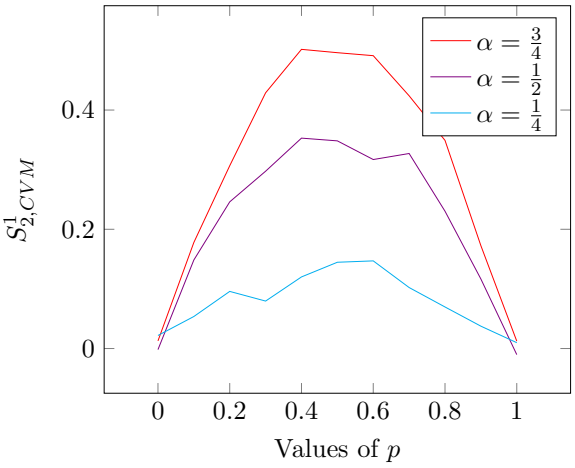

Figure 1: First index variation with p

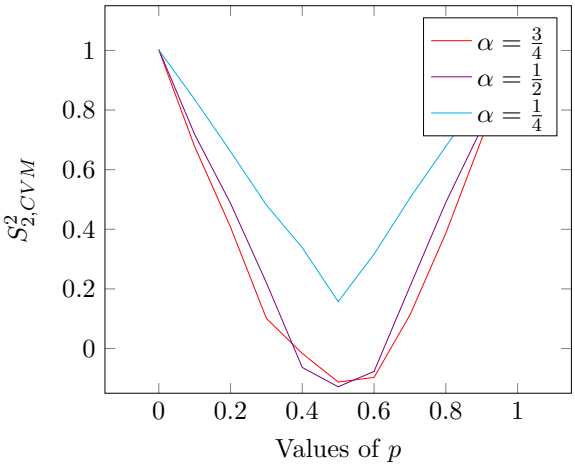

Figure 2: Second index variation with p

6 Sensitivity analysis using rank statistics

6.1 Introduction

The aim of this section is to introduce a new estimator based on the work of S.Chatterjee. In [20], he introduced a new empirical correlation coefficient that is directly related to both Cramér-Von Mises and Sobol' indices. Their main advantage lies within the small size of data required to get "good" performances. We will present the two the estimators and their asymptotic properties.

6.2 Estimators

Let $Y = f(V_1, \ldots, V_n) \in \mathbb{R}$ be the output of a numerical code. Let us consider a pair of real-valued random variables (X, Y) as an i.i.d sample $(X_j, Y_j)_{1 \leqslant j \leqslant n}$. First, we rearrange the pairs $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ in such a way that

$$X_{(1)} \leqslant \ldots \leqslant X_{(n)}. \tag{47}$$

Considering r_i to be the rank of $Y_{(i)}$ - that is, the number of j such that $Y_{(j)} \leqslant Y_{(i)}$ - the new coefficient of correlation introduced by S.Chatterjee [20] is defined as

$$\xi_n(X, Y) = 1 - \frac{3\sum_{i=1}^{n-1} |r_{i+1} - r_i|}{n^2 - 1}. \tag{48}$$

In the presence of ties, ξ_n is defined as follows: if there are ties among the X_i's, then choose an increasing rearrangement as above by breaking ties uniformly at random. Additionally, let r_i be

as previously introduced and define l_i the number of j such that $Y_{(j)} \geqslant Y_{(i)}$. The coefficient is thus defined by

$$\xi_n(X,Y) = 1 - \frac{n\sum_{i=1}^{n-1}|r_{i+1}-r_i|}{2\sum_{i=1}^{n} l_i(n-l_i)}$$

Theorem 6.1 (Chatterjee) *If Y is not almost surely a constant, then as $n \to \infty, \xi_n(X,Y)$ converges almost surely to the deterministic limit*

$$\xi(X,Y) := \frac{\int \mathrm{Var}(\mathbb{E}[\mathbb{1}_{\{Y\geqslant t\}}|X])d\mu(t)}{\int \mathrm{Var}(\mathbb{1}_{\{Y\geqslant t\}})d\mu(t)}, \tag{49}$$

where μ is the law of Y. This limit belongs to the interval $[0,1]$. It is 0 if and only if X and Y are independent, and it is 1 if and only if there is a measurable function $f : \mathbb{R} \to \mathbb{R}$ such that $Y = f(X)$ almost surely.

This limit is equal to the Cramér-Von Mises sensitivity index $S^X_{2,\mathrm{CVM}}$ with respect to X, if it's one of the random real-valued variables $X_1, \ldots, X_n$ in the considered numerical code. This estimator will not be taken into consideration as the hypothesis of the input and output being i.i.d. is almost never granted in real models. The idea behind it will be used however in order to provide an estimator of the Sobol' index, as we will see next. Let $\pi(i)$ be the rank of X_i, breaking ties at random so that π is permutation of $\{1, \ldots, n\}$. Define

$$N(i) := \begin{cases} \pi^{-1}(\pi(i)+1) & \text{if } \pi(i) < n, \\ i & \text{if } \pi(i) = n. \end{cases} \tag{50}$$

For any $t \in \mathbb{R}$, let

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i\leqslant t\}}, \quad G_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i\geqslant t\}}. \tag{51}$$

Additionaly define

$$Q_n := \frac{1}{n}\sum_{i=1}^{n} \min\{F_n(Y_i)F_N(Y_{N(i)})\} - \frac{1}{n}\sum_{i=1}^{n} G_n(Y_i)^2 \tag{52}$$

And

$$S_n := \frac{1}{n}\sum_{i=1}^{n} G_n(Y_i)(1-G_n(Y_i)), \tag{53}$$

Taking π as the permutation introduced earlier, we notice that $nF_n(Y_i) = r_{\pi(i)}$ for all i, and $nF(Y_n(N(i)) = r_{\pi(i)+1}$ for $i \neq \pi^{-1}(n)$. If $i = \pi(n)$, then $nF_n(Y_i) = nF_n(Y_{N(i)}) = r_n$. Therefore

$$Q_n = \frac{1}{n}\sum_{i=1}^{n} \min\{F_n(Y_i), F_N(Y_{N(i)})\} = \frac{1}{n}^2 \sum_{i\neq\pi^{-1}(n)} \min\{r_{\pi(i)}, r_{\pi(i+1)}\} + \frac{r_n}{n^2}. \tag{54}$$

The identity $\min\{a,b\} = \frac{1}{2}(a+b-|a-b|)$ gives

$$\frac{1}{n}\sum_{i=1}^{n} \min\{F_n(Y_i), F_N(Y_{N(i)})\} = \frac{1}{n^2}\sum_{i=1}^{n} r_i - \frac{1}{2n^2}\sum_{i=1}^{n-1}|r_{i+1}-r_i| + \frac{r_n - r_1}{2n^2}. \tag{55}$$

Combining all of the above, we get

$$\frac{Q_n}{S_n} = \xi_n + \frac{r_n - r_1}{2n^2 S_n} \tag{56}$$

And in particular,

$$\left|\frac{Q_n}{S_n} - \xi_n\right| \leqslant \frac{1}{2nS_n}. \tag{57}$$

Since S_n converges to a non-zero limit, we can rewrite ξ_n as Q_n/S_n. [21]

Let $\mathbf{u}$ be a subset of $\{1,\dots,n\}$, taking $Y^{\mathbf{u}} = f(X^{\mathbf{u}})$ we have the following result

$$\mathrm{Var}(\mathbb{E}\left[\mathbb{1}_{\{Y\geqslant t\}}|X^{\mathbf{u}}\right]) = \mathrm{Cov}\left(\mathbb{1}_{\{Y\geqslant t\}}, \mathbb{1}_{\{Y^{\mathbf{u}}\geqslant t\}}\right)$$

Lemma 6.1.1 *Let G_X be the conditional survival function: $G_X = \mathbb{P}(Y \geqslant t|X)$. We can rewrite the previous formula as*

$$\mathrm{Var}\left(\mathbb{E}\left[\mathbb{1}_{\{Y_j\geqslant t\}}\mathbb{1}_{\{Y_{N(j)}\geqslant t\}}|X_1,\dots,X_n\right]\right) = G_{X_j}(t)G_{X_{N(j)}}(t). \tag{58}$$

For g and h two integrable functions, the authors in [21] propose a universal estimation procedure of expectation of the form

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

Let us first define Ψ_X as

$$\Psi_X(g) := \mathbb{E}\left[g(Y)|X\right], \tag{59}$$

Lemma 6.1.2 *Let g and h be two integrable functions such that gh is also integrable. Let (X_j, Y_j) be an n-sample of (X, Y). Consider an $\mathcal{F}_n$-measurable random permutation σ_n with no fix point (i.e. $\sigma_n(j) \neq j$) for all $j = 1,\dots,n$. Then*

$$\mathbb{E}\left[g(Y_j)h(Y_{\sigma_n(j)}|X_1,\dots,X_n\right] = \Psi_{X_j}(g)\Psi_{X_{\sigma_n(j)}}(h). \tag{60}$$

Theorem 6.2 *Let g and h be two bounded measurable functions. Consider an $\mathcal{F}_n$-measurable random permutation σ_n with no fix point for all $j = 1,\dots,n$. In addition, assume that for any $j = 1,\dots,n$, $V_{\sigma_n} \to V_j$ as $n \to \infty$ with probability one. Then $\chi_n(X, Y; g, h)$ defined by*

$$\chi_n(X, Y; g, h) := \frac{1}{n}\sum_{j=1}^{n} g(Y_j)h(Y_{\sigma_n(j)}) \tag{61}$$

converges almost surely as $n \to \infty$ to $\chi(X, Y; g, h)$, given by

$$\chi(X, Y; g, h) = \Psi_{X_j}(g)\Psi_{X_{\sigma_n(j)}}(h) \tag{62}$$

where Ψ_X is given in (59).+

We can now provide an estimator to the first-order Sobol' index S^i with respect to $X = V_i$. Taking $g(x) = h(x) = x$ and $\sigma_n = N$ provides the analogue to ξ_n to estimate first order Sobol' indices, given by:

$$\xi_n^{\mathrm{Sobol}}(X, Y) = \frac{\frac{1}{n}\sum_{j=1}^{n} Y_i Y_{N(j)} - (\frac{1}{n}\sum_{j=1}^{n} Y_j)^2}{\frac{1}{n}\sum_{j=1}^{n} Y_j^2 - (\frac{1}{n}\sum_{j=1}^{n} Y_j)^2}. \tag{63}$$

6.3 Asymptotic properties

Under some mild assumptions on the model f and the variable X_i, a CLT for the first-order Sobol' index S^i estimator ξ_n^{Sobol} is established, where S^i is given by:

$$\frac{\text{Var}(\mathbb{E}[Y|X_i])}{\text{Var}(Y)}.$$

Theorem 6.3 *Assuming X_i is real-valued and uniformly distributed in $[0,1]$, f to be a twice differentiable function with respect to its i-th coordinate and that both f and its first two derivatives are bounded, with respect to the i-th coordinate. Then:*

$$\sqrt{n}\left(\xi_n^{Sobol}(X_i, Y) - S^i\right) \tag{64}$$

is asymptotically Gaussian with zero mean and explicit variance σ^2.

Remark The boundedness of f implies it has a fourth moment, which is the minimal assumption to to get a CLT.

6.4 Numerical application

Let us consider the *Ishigami* model given by

$$Y = \sin X_1 + 7\sin^2 X_2 + 0.1X_3^4 \sin X_1$$

where the X_i's are i.i.d. uniform random variables in $[-\pi;\pi]$. This model is often used as an example for uncertainty and sensitivity analysis methods, because it exhibits strong non-linearity and non-monotonicity. It also has a peculiar dependence on X_3, as described by Sobol' & Levitan in [22].
Since rank based sensitivity analysis main's strength is the low number of oberservations used, compared to e.g. a Monte Carlo Pick-Freeze scheme, we will focus on showcasing this advantage.

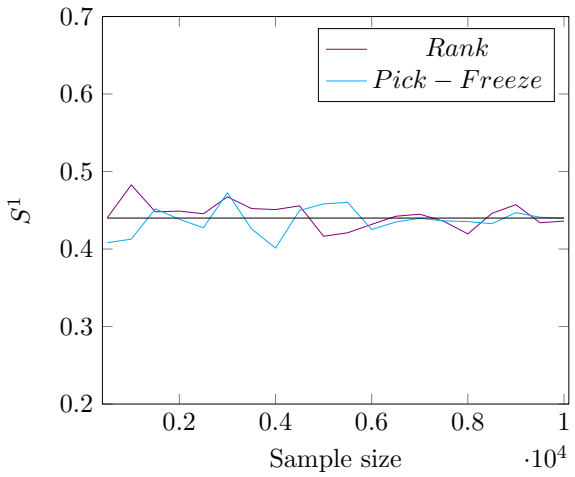

Figure 1: First index convergence

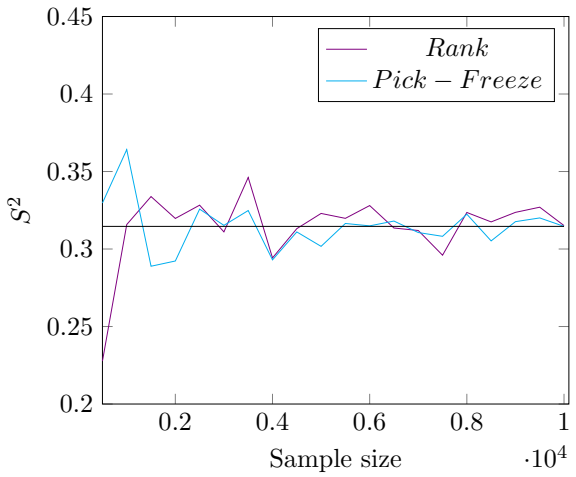

Figure 2: Second index convergence

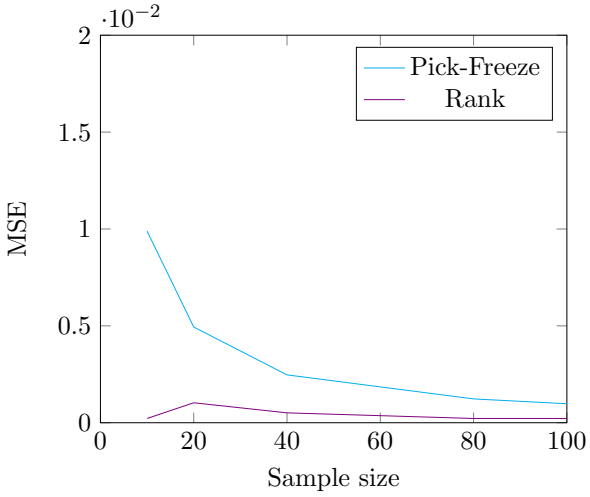

Figure 3: MSE^1 with growing sample size

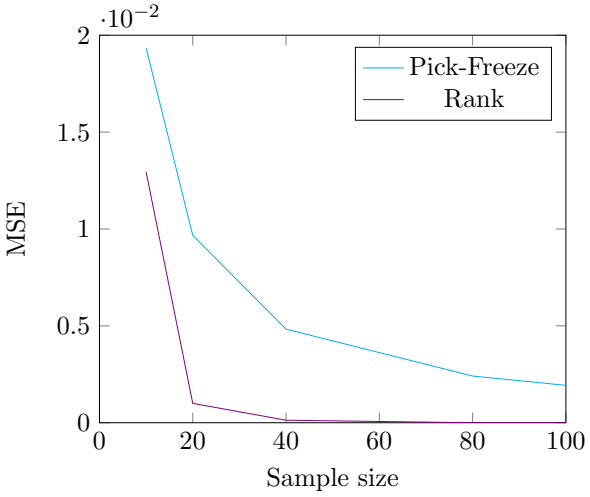

Figure 4: MSE^2 with growing sample size

Performances for small sample sizes : The rank-based scheme proceeds much better for small sample sizes as expected, and with no significant difference for bigger samples.. We consider for our small-sized model a sample size varying from 10 to 100. The mean square error for the first two indices are then shown. For a model where data might be scarce, the rank-based estimators can come in very handy.

7 Application : Aircraft take-off performance design requirements.

7.1 Introduction

In aircraft design process, Top Level Aircraft Requirements (TLARs) summarize the expected performance of future aircraft. Setting these requirements might be challenging and can result in an over-designed aircraft, and thus increased fuel consumption and overall cost. It is therefore necessary to do readjustments through a negotiation process. In designing an aircraft, some of these requirements, such as take-off requirements, can be modeled as constraints in an optimization problem. In this section, we will try to incorporate more information in the renegotiation process by considering some of these requirements as design variables rather than constraints, then perform sensitivity analysis to assess their impact. The results tell that combining airplane design and operations can provide new perspectives.

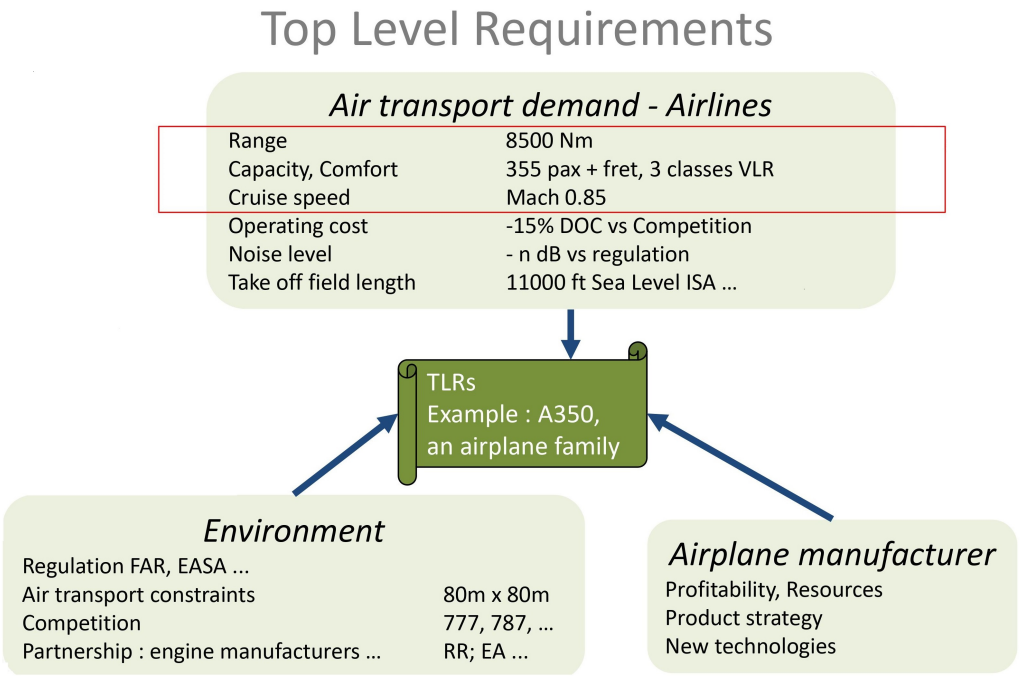

Figure 1: Top level requirements example scheme

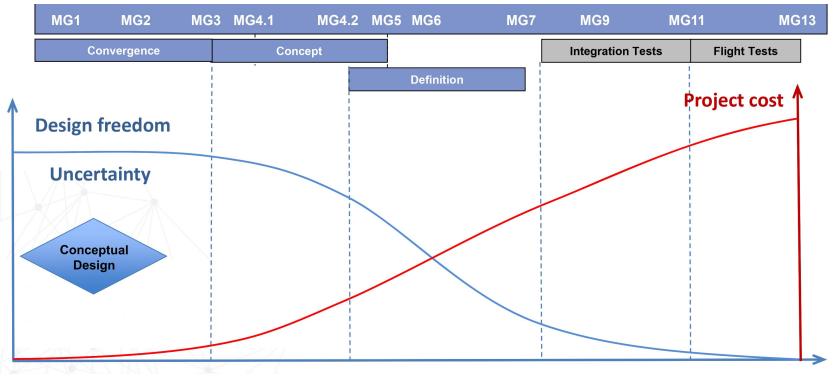

Figure 2: The curse of conceptual design

7.2 Airplane design tool and data

In this work, the airplane design tool used is based on the MARILib library [23]. It contains a set of models dedicated to airplane conceptual design, that can therefore be used for multi-disciplinary design optimization (MDO). The generic low fidelity models offered enable to size modern aircraft based on a reduced number of TLARs, such as cruise Mach, and also constraints such as take-off requirements.

The available database - gathered in the frame of the MOZAIC project [24] - offers detailed flight data from take-off to landing for four A340-400 aircraft over 20 years. The focus will therefore mainly be on the take-off requirements of this series. However, the data were needed to be tuned [2] for the model to match a real A340-400, given an acceptable margin of error.

The take-off requirements are usually taken into account in the optimization process as constraints. The parameters involved are among others, take-off mass, take-off field length, pressure altitude of the airport, and local temperature. The requirements within this framework are stated for maximum take-off mass (MTOM). The next section will define the process used to calculate the impact of take-off requirements on operational costs.

7.3 Computation of take-off requirements' impact on operational costs

The minimization problem is summarized in (65), while the process used for the calculation is presented in figure 3 below

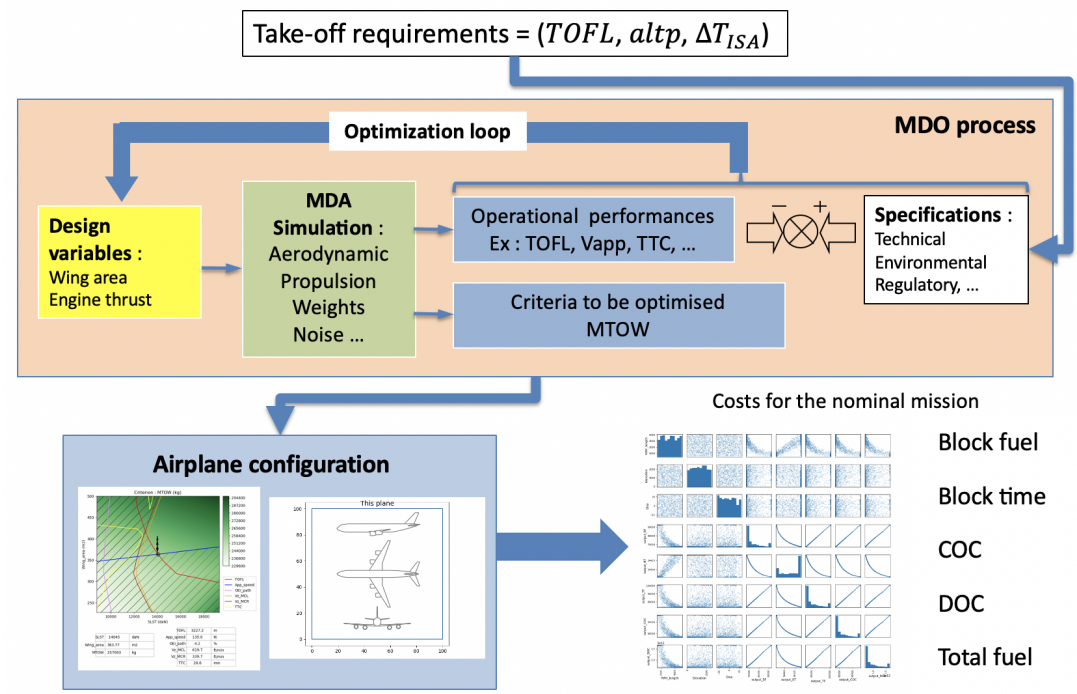

Figure 3: General process to calculate take-off requirements' impact on operational costs

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{with respect to} \quad & x \\ \text{subject to} \quad & g(x) \geqslant 0; \end{aligned} \tag{65}$$

where:
. x is the design variable vector, including the wing reference area and the engine sea level static thrust.

. $f(x)$ is the objection function, it can either be MTOM, cash operating cost or any other value of interest.
. $g(x)$ is the inequality constraint vector, which includes the take-off field length, the approach speed etc.

7.4 Sensitivity analysis

The data used is based on a combination of an airplane design tool and a database. The input parameters taken into consideration are the take-off field length ($TOFL$), the pressure altitude of the airport (Alt_p) and the difference of the local temperature relative to the ISA model of the atmosphere ΔT_{ISA}. The operational costs considered as inputs are the following: block fuel (BF), block time (BT), cash operating cost (COC), direct operating cost (DOC), and the total fuel (TF). The following table presents the results obtained for Sobol' and Cramér-Von Mises indices. They are calculated using Polynomial Chaos expansions, Pick-Freeze, and Rank Statistics. The CVM-index estimator provided using Rank Statistics has not been included as the available data does not resemble an i.i.d. sample, which is the condition for Chatterjee's theorem to be applied. The results displayed show only the direct influence of the inputs on the outputs (first-order indices) and do not take into account the combined influences. They are calculated using a design of experiments having 1000 points uniformly distributed in the following ranges:

$TOFL$: between 1500 and 5000 meters.
Alt_p: between 0 and 2500 meters.
ΔT_{ISA}: between -30 and +30 Celsius degrees.

Operation considered		Design cost missions			
Sensitivity indices		Sobol			CVM
Calculation method		PCE	Rank	P&F	
Output	Input	Index value			
Block Fuel	$TOFL$	81.3%	73.5%	70.7%	48.5%
	Alt_p	6.2%	8.8%	13.5%	8.6%
	ΔT_{ISA}	6.1%	8.2%	14.7%	16.2%
Block Time	$TOFL$	79.1%	69.4%	67.5%	53.4%
	Alt_p	7.4%	6.5%	12.4%	8.6%
	ΔT_{ISA}	8.7%	10.6%	18.3%	12.5%
COC	$TOFL$	81.3%	73.5%	70.7%	48.5%
	Alt_p	6.2%	8.8%	13.5%	8.7%
	ΔT_{ISA}	6.1%	8.2%	14.7%	12.5%
DOC	$TOFL$	81.3%	73.5%	70.7%	48.5%
	Alt_p	6.2%	8.8%	13.5%	8.6%
	ΔT_{ISA}	6.1%	8.2%	14.7%	12.5%
Total Fuel	$TOFL$	81.6%	73.6%	70.8%	48.5%
	Alt_p	5.9%	8.8%	13.4%	8.6%
	ΔT_{ISA}	5.8%	7.8%	14.5%	12.5%

Table 2: Sensitivity indices for the different costs related to the design cost mission

For all outputs in question, all indices present the same classification of the influence of each input. The take-off field length has by far the greatest influence while the elevation and temperature are almost equally involved with a small percentage each, except in the case of Cramér-Von Mises indices where it shows that the temperature has slightly more influence.

However, the CVM indices display a constant classification of the inputs for four outputs out of five. The reason is that a CVM index takes into account the whole distribution of the output regardless of the deviation of a variable from its mean, that is the variance. Something else worth noticing - since we only considered first-order indices for our study - is that the sums of Sobol' indices for each output are around 94-98 %, whereas the Cramér-Von Mises ones go as low as 73 %. This difference indicates that when the whole distribution is taken into consideration, the combined influences(second- and third-order indices) can significantly increase. Lastly, for the three calculation methods considered, block time shows different results than the other results when these come close to one another. This may be explained by the fact that the flight duration does not depend on take-off requirements as much as for the other four outputs.

To conclude, since $TOFL$ has the greatest impact on the outputs, it will be kept to specify the take-off requirements while the airport altitude and temperature will be set to the following values: $Alt_p = 0$ ft. and $\Delta T_{ISA} = 0$ Celsius degrees. This results in a modified design process with take-off requirements turned into design variables. In addition to $TOFL$, the design variables would now also include the wing reference area $WING_{area}$ and the engine sea-level static thrust $SLST$. The final results of the optimization process - for different payload disembarking payload costs (c_{pld}, expressed in \$/kg) - are displayed in Table 3. (for more details on the optimization process, refer to [2])

c_{pld}	$TOFL$	$WING_{area}$	$SLST$	$Total\ cost$
0.0	8.48%	-1.19%	-9.53%	-1.02 %
0.5	8.48%	-1.19%	-9.53%	-0.63 %
1.0	5.93%	-0.86%	-6.85%	-0.30 %
1.5	3.42%	-0.51%	-4.07%	-0.15 %
2.0	1.75%	-0.27%	-2.13%	-0.09 %
2.5	1.21%	-0.19%	-1.49%	-0.07 %
3.0	0.93%	-0.14%	-1.14%	-0.05 %
4.0	0.74%	-0.11%	-0.92%	-0.04 %
5.0	0.57%	-0.09%	-0.69%	-0.03 %
6.0	0.41%	-0.06%	-0.51%	-0.02 %
8.0	0.31%	-0.05%	-0.39%	-0.01 %
10.0	0.24%	-0.04%	-0.30%	-0.01 %

Table 3: Results of optimization presented as relative difference to the reference aircraft

We can notice that the total cost decreases with c_{pld}. This is explained by the fact the airlines work to maximize the use of their fleet, once they are aware of their operational limits. The values c_{pld} cover are from 0 to 10 \$/kg. Assuming a passenger represents 100 kg of payload, then the range the penalty we cover is up to 1000 \$ per passenger, which seems reasonably high for the price of a long-range flight ticket. It is important to note that these costs will greatly depend on the region of the world and airlines operated, the period of the year, the day of the week, and the additional services related.

Also, as the $TOFL$ increases, both the $WING_{area}$ and $SLST$ reduce. We would expect this to make the airplane more affordable.

And finally, though the gains in COC might seem marginal (up to 0.63%), when scaled to a full fleet over a long period of time, they present a large sum.

8 Conclusion

In this paper, I present multiple sensitivity analysis indices with various ways of estimating them. The classical Sobol' indices can show several drawbacks that can lead to consider the use of Cramér-Von Mises ones. The latter contains all of the distribution information. It is

up to the practitioner to select the rightest combination of index and method of calculation relative to the data desired to analyse. The paper concludes with a concrete example applied to the aircraft design process in which parameters were tuned in order to improve the overall optimization process, resulting in significant savings.

Acknowledgments

I would like to express my sincere gratitude to Pr. Thierry Klein and Pr. Nicolas Peteilh from Ecole Nationale de l'Aviation Civile - for their clear-cut remarks, their constructive criticism, and most importantly their support and consideration throughout the whole period of this project.

Appendix

Theorems

Theorem 8.1 (Central Limit Theorem) *Let $\{X_1, \ldots, X_n\}$ be a random n-sample size- that is, a sequence of independent and identically distributed (i.i.d.) random variables drawn from a distribution of expectation (or expected value) given by μ and finite variance given by σ^2. Then*

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2) \tag{66}$$

Theorem 8.2 (Strong law of large numbers) *Let $\{X_1, \ldots, X_n\}$ be a random n-sample size- that is, a sequence of independent and identically distributed (i.i.d.) random variables drawn from a distribution of expectation (or expected value) given by μ. Then*

$$\bar{X}_n := \frac{X_1 + \ldots + X_n}{n} \xrightarrow[n\to\infty]{\text{a.s}} \mu \tag{67}$$

Legendre polynomials

The Legendre polynomials $P_n(x)$ are solution of the following differential equation:

$$(1 - x^2)y" - 2xy' + n(n+1)y = 0 \quad (n \in \mathbb{N}) \tag{68}$$

They may be generated in practice by the following recurrence relationship:

$$\begin{aligned} P_0(x) &= 1 \\ (n+1)P_{n+1}(x) &= (2n+1)xP_n(x) - nP_{n-1}(x); \end{aligned} \tag{69}$$

They are orthogonal with respect to the uniform probability measure over [-1, 1]:

$$\int_{-1}^{1} P_m(x)P_n(x)dx = \frac{2}{2n+1}\delta_{mn} \tag{70}$$

where δ_{mn} is the Kronecker symbol. If U is a random variable with a uniform probability distribution function over [-1,1], the following relationship holds:

$$\mathbb{E}\left[P_m(U)P_n(U)\right] = \frac{2}{2n+1}\delta_{mn} \tag{71}$$

Additionally, the first three Legendre polynomials are:

$$P_1(x) = x \ , \ P_2(x) = \frac{1}{2}(3x^2 - 1) \ , \ P_3(x) = \frac{1}{2}(5x^3 - 3x) \tag{72}$$

References

[1] Anil Variyar, Thomas D. Economon, and Juan J. Alonso. "Design and Optimization of Unconventional Aircraft Configurations with Aeroelastic Constraints". In: *55th AIAA Aerospace Sciences Meeting*. DOI: 10.2514/6.2017-0463. eprint: https://arc.aiaa.org/doi/pdf/10.2514/6.2017-0463. URL: https://arc.aiaa.org/doi/abs/10.2514/6.2017-0463.

[2] Nicolas Peteilh et al. "Correction: Challenging Top Level Aircraft Requirements based on operations analysis and data-driven models, application to takeoff performance design requirements". In: *AIAA AVIATION 2020 FORUM*. DOI: 10.2514/6.2020-3171.c1. eprint: https://arc.aiaa.org/doi/pdf/10.2514/6.2020-3171.c1. URL: https://arc.aiaa.org/doi/abs/10.2514/6.2020-3171.c1.

[3] B. Williams T. J. Santner and W. Notz. "The Design and Analysis of Computer Experiments". In: 2003. DOI: 10.1007/978-1-4757-3799-8.

[4] "Uncertainty Modelling Methods". In: *Uncertainty in Industrial Practice*. John Wiley Sons, Ltd, 2008. Chap. 16, pp. 225–238. ISBN: 9780470770733. DOI: https://doi.org/10.1002/9780470770733.ch16. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470770733.ch16. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470770733.ch16.

[5] K. Chan A. Saltelli and E. Scott. "Sensitivity analysis". In: 2000. ISBN: 978-0-470-74382-9.

[6] Xiaobo Zhou and Henry Lin. "Local Sensitivity Analysis". In: *Encyclopedia of GIS*. Ed. by Shashi Shekhar and Hui Xiong. Boston, MA: Springer US, 2008, pp. 616–616. ISBN: 978-0-387-35973-1. DOI: 10.1007/978-0-387-35973-1_703. URL: https://doi.org/10.1007/978-0-387-35973-1_703.

[7] K. Pearson. "On the partial correlation ratio". In: *Proceedings of the Royal Society of London. Series A* (June 1914). DOI: 10.1098/rspa.1915.0041.

[8] I. M. SOBOL'. "Sensitivity analysis for non-linear mathematical models". In: *Mathematical Modelling and Computational Experiment* 1 (1993), pp. 407–414. URL: https://ci.nii.ac.jp/naid/10027137978/en/.

[9] Wassily Hoeffding. "A Class of Statistics with Asymptotically Normal Distribution". In: (Jan. 1948), pp. 293–325.

[10] Sergei Kucherenko and Song Shufang. "Comparison of different numerical estimators for main effect global sensitivity indices". In: *UNCECOMP 2015 - 1st ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering* (Jan. 2015), pp. 607–638.

[11] Art B. Owen. "Better Estimation of Small Sobol' Sensitivity Indices". In: *ACM Trans. Model. Comput. Simul.* 23.2 (May 2013). ISSN: 1049-3301. DOI: 10.1145/2457459.2457460. URL: https://doi.org/10.1145/2457459.2457460.

[12] Bruno Sudret. "Global sensitivity analysis using polynomial chaos expansions". In: *Reliability Engineering System Safety* 93.7 (2008). Bayesian Networks in Dependability, pp. 964–979. ISSN: 0951-8320. DOI: https://doi.org/10.1016/j.ress.2007.04.002. URL: http://www.sciencedirect.com/science/article/pii/S0951832007001329.

[13] Fabrice Gamboa et al. "Statistical inference for Sobol pick freeze Monte Carlo method". In: *Statistics* 50.4 (2016), pp. 881–902. DOI: 10.1080/02331888.2015.1105803. URL: https://hal.inria.fr/hal-00804668.

[14] Art Owen, Josef Dick, and Su Chen. *Higher order Sobol' indices*. 2013. arXiv: 1306.4068 [math.NA].

[15] Sébastien Da Veiga. "Global Sensitivity Analysis with Dependence Measures". working paper or preprint. Nov. 2013. URL: https://hal.archives-ouvertes.fr/hal-00903283.

[16] Toshimitsu Homma and Andrea Saltelli. "Importance measures in global sensitivity analysis of nonlinear models". In: *Reliability Engineering System Safety* 52.1 (1996), pp. 1–17. ISSN: 0951-8320. DOI: https://doi.org/10.1016/0951-8320(96)00002-6. URL: http://www.sciencedirect.com/science/article/pii/0951832096000026.

[17] Hervé Monod, C. Naud, and David Makowski. "Uncertainty and sensitivity analysis for crop models". In: *Working with Dynamic Crop Models* (Jan. 2006), pp. 55–100.

[18] Wei Zhao and Lingze Bu. "Global sensitivity analysis with a hierarchical sparse metamodeling method". In: *Mechanical Systems and Signal Processing* 115 (2019), pp. 769–781. ISSN: 0888-3270. DOI: https://doi.org/10.1016/j.ymssp.2018.06.044. URL: http://www.sciencedirect.com/science/article/pii/S0888327018303832.

[19] Fabrice Gamboa, Thierry Klein, and Agnès Lagnoux. "Sensitivity analysis based on Cramér von Mises distance". In: *SIAM/ASA Journal on Uncertainty Quantification* 6.2 (Apr. 2018), pp. 522–548. DOI: 10.1137/15M1025621. URL: https://hal.archives-ouvertes.fr/hal-01163393.

[20] Sourav Chatterjee. "A new coefficient of correlation". In: (2020). arXiv: 1909.10140 [math.ST].

[21] Fabrice Gamboa et al. "Global Sensitivity Analysis: a new generation of mighty estimators based on rank statistics". working paper or preprint. Nov. 2020. URL: https://hal.archives-ouvertes.fr/hal-02474902.

[22] I.M. Sobol' and Yu.L. Levitan. "On the use of variance reducing multipliers in Monte Carlo computations of a global sensitivity index". In: *Computer Physics Communications* 117.1 (1999), pp. 52–61. ISSN: 0010-4655. DOI: https://doi.org/10.1016/S0010-4655(98)00156-8. URL: http://www.sciencedirect.com/science/article/pii/S0010465598001568.

[23] Thierry Druot et al. "A Multidisciplinary Airplane Research Integrated Library With Applications To Partial Turboelectric Propulsion". In: June 2019. DOI: 10.2514/6.2019-3243.

[24] European Research Infrastructure IAGOS (In-service Aircraft of a Global Observing System). "MOZAIC IAGOS Database". In: 2016. URL: https://iagos.org.