Data-Driven Sensitivity Indices for Models With Dependent Inputs Using Polynomial Chaos Expansions

Zhanlin Liu[a], Youngjun Choe[a,*]

[a] *Department of Industrial and Systems Engineering, University of Washington, Seattle, WA 98195, USA*

Abstract

Uncertainties exist in both physics-based and data-driven models. Variance-based sensitivity analysis characterizes how the variance of a model output is propagated from the model inputs. The Sobol index is one of the most widely used sensitivity indices for models with independent inputs. For models with dependent inputs, different approaches have been explored to obtain sensitivity indices in the literature. Typical approaches are based on procedures of transforming the dependent inputs into independent inputs. However, such transformation requires additional information about the inputs, such as the dependency structure or the conditional probability density functions. In this paper, data-driven sensitivity indices are proposed for models with dependent inputs. We first construct ordered partitions of linearly independent polynomials of the inputs. The modified Gram-Schmidt algorithm is then applied to the ordered partitions to generate orthogonal polynomials with respect to the empirical measure based on observed data of model inputs and outputs. Using the polynomial chaos expansion with the orthogonal polynomials, we obtain the proposed data-driven sensitivity indices. The sensitivity indices provide intuitive interpretations of how the dependent inputs affect the variance of the output without a priori knowledge on the dependence structure of the inputs. Four numerical examples are used to validate the proposed approach.

[*] Corresponding author

Email address: ychoe@uw.edu (Youngjun Choe)

Keywords: uncertainty quantification, global sensitivity analysis, variance-based sensitivity analysis, Sobol index, polynomial chaos expansion, Gram-Schmidt orthogonalization

1. Introduction

Uncertainties exist in both physics-based and data-driven models. Uncertainty quantification (UQ) methods to characterize and reduce those uncertainties are increasingly popular in engineering studies. As an aspect of UQ, sensitivity analysis (SA) quantifies how output uncertainties are propagated from input uncertainties. Two general ways of conducting SA are local sensitivity analysis (LSA) and global sensitivity analysis (GSA). LSA analyzes how a small perturbation near an input space value could influence the output. On the contrary, GSA investigates how the input variability influences the output variability over the entire input space. In recent studies, variance-based sensitivity analysis, as a form of GSA, is utilized to understand system uncertainties in various applications such as material mechanics [1], building energy [2], structural mechanics [3], hydrogeology [4], and manufacturing [5].

Conducting variance-based sensitivity analysis for models with *independent* inputs has been studied widely. Monte Carlo simulation and surrogate models are two general ways to obtain sensitivity indices for models with independent inputs. Surrogate models have been shown to be more computationally efficient compared with Monte Carlo simulation [6]. Polynomial chaos expansion (PCE) and Kriging (also known as Gaussian process regression) are the two surrogate models which have been used to compute sensitivity indices most commonly in the literature [6, 7]. Thanks to the orthogonal property of a PCE model, sensitivity indices for *independent* inputs can be directly obtained using PCE coefficients [8, 9, 7]. PCE-based sensitivity indices appear in various fields including fluid dynamics [10], structural reliability [11], and vehicle dynamics [12].

For models with *dependent* inputs, a limited number of approaches are available in the literature to conduct variance-based sensitivity analyses. Generalized

Sobol sensitivity indices have been proposed in Chastaing et al. [13] based on the hierarchically orthogonal functional decomposition (HOFD). However, the unboundedness of the resulting sensitivity indices makes their interpretation for *dependent* inputs not as straightforward as the Sobol indices for models with *independent* inputs [14]. A different framework is proposed in [15] to obtain sensitivity indices for models with correlated inputs. However, it requires the knowledge of model structure between the inputs and the outputs. An alternative way of obtaining sensitivity indices for models with *dependent* inputs is to transform *dependent* inputs into *independent* inputs [16, 17, 18]. Even though the transformation-based methods generate interpretable sensitivity indices, they require strong assumptions on the dependency or distributions of the inputs.

The main contribution of this paper is the development of a data-driven method to obtain interpretable sensitivity indices for models with dependent inputs without invoking any assumptions on the inputs. We first propose the modified Gram-Schmidt based polynomial chaos expansion (mGS-PCE). The mGS-PCE increases the numerical robustness of constructing orthogonal polynomials for arbitrarily distributed inputs compared with the GS-PCE in [10]. Then, we propose a method to obtain data-driven sensitivity indices for models with dependent inputs by constructing ordered partitions of orthonormal polynomials of the inputs. This method estimates some of the sensitivity indices in [16] and [17] without invoking the assumptions therein. Lastly, we propose conditional order-based sensitivity indices, which explain the model output variability in a hierarchical manner.

The remainder of the paper is organized as follows: Section 2 reviews the background knowledge about Sobol indices and PCE models. Section 3 introduces the modified Gram-Schmidt algorithm and our data-driven method to obtain sensitivity indices for models with dependent inputs using PCE models. In Section 4, four numerical examples, where the inputs are dependent, are used to validate our proposed method. Section 5 provides a few concluding remarks and a discussion on future research directions.

2. Technical background

This section briefly reviews sensitivity indices in the existing literature for models with independent inputs and those with dependent inputs. We first introduce the Hoeffding functional decomposition and the Sobol indices for independent inputs. We then review the *full* sensitivity indices and the *uncorrelated* sensitivity indices defined for models with *dependent* inputs. Lastly, we introduce PCE models and explain how PCE coefficients can be used to calculate sensitivity indices.

2.1. Hoeffding decomposition and sensitivity indices for independent inputs

Suppose we have n independent random inputs $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ with their density $\mu(\boldsymbol{X})$. For the output $Y = f(\boldsymbol{X})$ that is square-integrable with respect to $\mu(\boldsymbol{X})$, its Hoeffding decomposition is defined as follows [19, 13]:

$$f(\boldsymbol{X}) = \sum_{u \subseteq \{1,2,\ldots,n\}} f_u(\boldsymbol{X}_u), \tag{1}$$

where $f_\emptyset = f_0$ and f_0 is a constant and $\boldsymbol{X}_u = (X_j)_{j \in u}$. Each summand $f_u(\boldsymbol{X}_u), u \neq \emptyset$, in Eq. (1) satisfies

$$\int f_u(\boldsymbol{X}_u)\mu(X_i)\,\mathrm{d}X_i = 0, \quad \forall i \in u.$$

such that

$$f_0 = \int f(\boldsymbol{X})\mu(\boldsymbol{X})\,\mathrm{d}\boldsymbol{X}.$$

In addition, the summands in Eq. (1) are orthogonal to each other as follows:

$$\int f_u(\boldsymbol{X}_u) f_v(\boldsymbol{X}_v)\mu(\boldsymbol{X})\,\mathrm{d}\boldsymbol{X} = 0, \quad \forall u, v \subseteq \{1, 2, \ldots, n\}, v \neq u.$$

Based on the Hoeffding decomposition, the variance of Y is decomposed as follows [8, 9]:

$$\begin{aligned} Var(Y) &= \int f^2(\boldsymbol{X})\mu(\boldsymbol{X})\,\mathrm{d}\boldsymbol{X} - f_0^2 \\ &= \sum_{\substack{u \subseteq \{1,2,\ldots,n\} \\ u \neq \emptyset}} D_u(Y), \end{aligned}$$

where

$$\begin{aligned} D_u(Y) &= \int f_u^2(\boldsymbol{X}_u)\mu(\boldsymbol{X}_u)\,\mathrm{d}\boldsymbol{X}_u \\ &= Var(E(Y \mid \boldsymbol{X}_u)) - \sum_{\substack{v\subset u\\ v\neq u\\ v\neq\emptyset}} D_v(Y). \end{aligned}$$

For example, $D_i(Y) = Var(E(Y \mid X_i))$ and $D_{ij}(Y) = Var(E(Y \mid X_i, X_j)) - D_i(Y) - D_j(Y)$.

Based on the variance decomposition, the Sobol index for set u is defined as

$$S_u = \frac{D_u(Y)}{Var(Y)},$$

which measures the sensitivity of the output variance with respect to the inputs in $\boldsymbol{X}_u$. For a particular input variable X_i, the *first-order* Sobol index S_{X_i} and *total* Sobol index ST_{X_i} are defined as follows:

$$\begin{aligned} S_{X_i} &= \frac{D_i(Y)}{Var(Y)}, \\ ST_{X_i} &= \sum_{u\ni i} S_u. \end{aligned}$$

S_{X_i} represents the percentage of the output variance that is propagated from the input X_i. ST_{X_i} represents the percentage of the output variance that is propagated from the input X_i and its interactions with the other variables.

2.2. Sensitivity indices for dependent inputs

This study focuses on sensitivity indices proposed in [16, 17, 18] because they are bounded and do not require the knowledge of the model structure between the inputs and the output in contrast to those considered in [20, 13, 15], as discussed earlier.

In [16], the Gram-Schmidt algorithm is employed to decorrelate the inputs when the dependences are characterized solely by the inputs' first-order conditional moments. Then the *full* sensitivity indices and the *uncorrelated* sensitivity indices (also called *independent* sensitivity indices in [17]) are defined. On the other hand, in order to calculate these sensitivity indices when conditional probability density functions (cPDFs) of the inputs are known, the inverse Rosenblatt transformation or the inverse Nataf transformation is applied to transform the *dependent* inputs into the *independent* inputs [17, 18].

Suppose dependent inputs $(X_1, X_2, \ldots, X_n)$ are transformed into independent inputs $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n)$, for example, under the assumptions of [16] such that $\bar{X}_1 = X_1$ and $\bar{X}_i = X_i - E\big(X_i \mid \bar{X}_1, \ldots, \bar{X}_{i-1}\big)$, $\forall i = 2, \ldots, n$. Intuitively speaking, $\bar{X}_1$ keeps all information concerning X_1 including its dependent part with the other inputs. $\bar{X}_i$ contains all information concerning X_i *except* its dependent part with $\bar{X}_1, \ldots, \bar{X}_{i-1}$. Thus, $\bar{X}_n$ only contains information of X_n *excluding* its dependent part with all the other inputs. These constructed *independent* inputs allow for calculating the first-order Sobol indices ($S_{\bar{X}_i}$) and total Sobol indices ($ST_{\bar{X}_i}$). Then the sensitivity indices with respect to the *dependent* inputs are defined as follows [16]:

$\bar{S}_{X_1} = S_{\bar{X}_1}$ is the first-order full contribution of X_1 to the variance of the output.

$\overline{ST}_{X_1} = ST_{\bar{X}_1}$ is the total full contribution of X_1 to the variance of the output.

$S^u_{X_n} = S_{\bar{X}_n}$ is the first-order uncorrelated contribution of X_n to the variance of the output.

$ST^u_{X_n} = ST_{\bar{X}_n}$ is the total uncorrelated contribution of X_n to the variance of the output.

By permuting the order of the inputs, different sensitivity indices can be further calculated. Suppose the initial input variables are ordered as $(X_i, X_{i+1}, \ldots, X_n,$-$X_1, \ldots, X_{i-1})$, and the constructed independent inputs are $(\bar{X}_i, \bar{X}_{i+1}, \ldots, \bar{X}_n, \bar{X}_1, \ldots,$-$\bar{X}_{i-1})$. Then the *full* sensitivity indices ($\bar{S}_{X_i} = S_{\bar{X}_i}$ and $\overline{ST}_{X_i} = ST_{\bar{X}_i}$) and the *uncorrelated* sensitivity indices ($S^u_{X_{i-1}} = S_{\bar{X}_{i-1}}$ and $ST^u_{X_i} = ST_{\bar{X}_{i-1}}$) are defined. $\bar{S}_{X_i}$ is called the first-order *full* sensitivity index and $\overline{ST}_{X_i}$ is called the total *full* sensitivity index. $S^u_{X_i}$ is called the first-order *uncorrelated* sensitivity index and $ST^u_{X_i}$ is called the total *uncorrelated* sensitivity index.

2.3. PCE and PCE-based sensitivity indices

As a way of calculating sensitivity indices, PCE is known to be more computationally efficient than Monte Carlo simulations [8, 9]. The original PCE, which is proposed in [21], provides Hermite polynomials for independent Gaussian random variables. Several types of PCE have been proposed under the assumption of *independence* between model inputs, including the generalized PCE (gPCE) [22], the multi-element generalized PCE (ME-gPCE) [23], the moment-based arbitrary PCE (aPCE) [24] and the Gram-Schmidt based PCE (GS-PCE) [10].

The GS-PCE for models with *independent* inputs is extended to models with multivariate *dependent* inputs in Navarro et al. [14]. It is regarded as the pioneering work in constructing an orthogonal polynomial basis for arbitrary *dependent* inputs. Rahman [25] theoretically validates the Gram-Schmidt orthogonalization process to construct an orthogonal polynomial basis for the PCE with dependent inputs.

2.3.1. PCE model

PCE uses a finite number of orthonormal polynomial terms of n random inputs in $\boldsymbol{X}$ to approximate the output Y as follows:

$$Y = f(\boldsymbol{X}) \approx \sum_{i=0}^{P} \theta_i \psi_i(\boldsymbol{X}), \tag{2}$$

where θ_i, $i = 0,1,2,\ldots,P$, are called PCE coefficients and ψ_i, $i = 1,2,\ldots,P$ are orthonormal polynomials.

$$P + 1 = \binom{n+p}{n} \tag{3}$$

is the number of polynomial terms, where p is the highest polynomial degree in the PCE model. As p increases, the accuracy of approximating a complex output function improves. In this paper, we estimate the PCE coefficients by solving an overdetermined linear system of equations in the least-squares sense as proposed in [26].

Thanks to the properties of orthonormal polynomials, we can approximate the lower order moments of output Y directly using the PCE coefficients in (2) as follows:

$$\begin{aligned} E(Y) &\approx \theta_0, \\ Var(Y) &\approx \sum_{i=1}^{P} \theta_i^2. \end{aligned} \tag{4}$$

The approximation errors converge to zero as P increases [27].

2.3.2. PCE-based sensitivity indices

For *independent* inputs, the multivariate orthonormal polynomials $\psi_i(\boldsymbol{X})$ can be directly constructed as the products of univariate orthonormal polynomials as follows [8, 9, 7]:

$$\psi_i(\boldsymbol{X}) = \psi_{\boldsymbol{\alpha}_i}(\boldsymbol{X}) = \prod_{j=1}^{n} \psi_{\alpha_{ij}}(X_j),$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in})$ and $\psi_{\alpha_{ij}}(X_j)$ represents the $\alpha_{ij}-$th order orthonormal polynomial in input X_j.

Define $\mathscr{A}_u$ as the set of multi-indices depending exactly on the subset of variables $\boldsymbol{X}_u, u \subseteq \{1, 2, \ldots, n\}$ as follows:

$$\mathscr{A}_u = \left\{ \boldsymbol{\alpha}_i \in \mathbb{N}^n : \alpha_{ij} \neq 0 \Leftrightarrow j \in u, |\boldsymbol{\alpha}_i| \leq p \right\},$$

where

$$|\boldsymbol{\alpha}_i| = \sum_{j=1}^{n} \alpha_{ij}.$$

Suppose $\theta_{\boldsymbol{\alpha}_i}$ is the PCE coefficient with respect to the polynomial term corresponding to $\boldsymbol{\alpha}_i$. Then the first-order Sobol index for X_j and the total Sobol index for X_j can be estimated for $j = 1, \ldots, n$ as follows [8, 9, 7]:

$$\begin{aligned} S_{X_j} &\approx \frac{\sum_{\boldsymbol{\alpha}_i \in \mathscr{A}_{\{j\}}} \theta_{\boldsymbol{\alpha}_i}^2}{\sum_{i=1}^{P} \theta_i^2}, \\ ST_{X_j} &\approx \sum_{\mathscr{A}_u \ni j} S_{\mathscr{A}_u}, \end{aligned}$$

where

$$S_{\mathscr{A}_u} \approx \frac{\sum_{\boldsymbol{\alpha}_i \in \mathscr{A}_u} \theta_{\boldsymbol{\alpha}_i}^2}{\sum_{i=1}^{P} \theta_i^2}.$$

Using these PCE-based Sobol indices, we can also obtain the sensitivity indices for *dependent* inputs, which were described earlier.

3. Methodology

In the previous section, we discussed the current methods proposed in [16] and [17] of obtaining the sensitivity indices for models with *dependent* inputs under certain assumptions on the inputs. In this section, we propose a data-driven method to estimate the sensitivity indices for models with *dependent* inputs using a PCE model based on the orthonormal polynomials constructed from the modified Gram-Schmidt algorithm. First, we show how to construct orthonormal polynomials using the modified Gram-Schmidt algorithm. Then we propose a data-driven method to estimate the *first-order full* sensitivity indices and the *total uncorrelated* sensitivity indices for models with *dependent* inputs. Then we propose an alternative *total full* sensitivity index and an alternative *first-order uncorrelated* sensitivity index, which can be also calculated using the proposed method. These alternative indices have different interpretations than those in [16] and [17] because our decorrelation process does not eliminate dependences in inputs. In addition, we propose conditional order-based sensitivity indices and illustrate how they can be used to reduce the PCE model complexity by excluding higher order interaction terms.

3.1. Modified Gram-Schmidt algorithm

In [14], orthonormal polynomials are constructed using the Gram-Schmidt algorithm for general multivariate correlated variables. Even though the Gram-Schmidt algorithm behaves the same as the modified Gram-Schmidt algorithm mathematically, the modified Gram-Schmidt algorithm is less sensitive to numeric rounding errors and performs more stably than the Gram-Schmidt algorithm [28]. Therefore, we propose to use the modified Gram-Schmidt algorithm to construct orthonormal polynomial basis $\{\psi_i(\boldsymbol{X})\}_{i=1}^{P}$ based on the initial P linearly independent polynomials $(e_i)_{i\in\{1,2,\dots,P\}}$ as follows [29]:

Algorithm 1 Modified Gram-Schmidt Algorithm

```
1: for $i = 1, 2, \ldots, P$ do
2:     $\phi_i(\boldsymbol{X}) \leftarrow e_i(\boldsymbol{X})$
3:     for $k = 1, 2, \ldots, i-1$ do
4:         $\phi_i(\boldsymbol{X}) \leftarrow \phi_i(\boldsymbol{X}) - \langle \phi_i(\boldsymbol{X}), \psi_k(\boldsymbol{X}) \rangle \psi_k(\boldsymbol{X})$
5:     end for
6:     $\psi_i(\boldsymbol{X}) \leftarrow \frac{\phi_i(\boldsymbol{X})}{||\phi_i(\boldsymbol{X})||_2}$
7: end for
```

The inner product in the algorithm is defined with respect to the empirical measure in this paper. The inner product is numerically evaluated using the observations of $\boldsymbol{X}$ in a given dataset. Note that the proposed data-driven method assumes neither any distributional knowledge of $\boldsymbol{X}$ nor the ability to easily sample from its distribution. Thus, we do not use a Monte Carlo approach to evaluate the inner product although it may be an option for the problems that permit the sampling.

The difference between the standard Gram-Schmidt algorithm and the modified Gram-Schmidt algorithm is at the line 4 in Algorithm 1, where the standard Gram-Schmidt algorithm performs

$$\phi_i(\boldsymbol{X}) \leftarrow \phi_i(\boldsymbol{X}) - \langle e_i(\boldsymbol{X}), \psi_k(\boldsymbol{X}) \rangle \, \psi_k(\boldsymbol{X}).$$

Note that different orthonormal polynomials are constructed from different permutations of the initial polynomials. In the following section, we discuss how to permute the order of the initial polynomials in order to obtain data-driven sensitivity indices for models with *dependent* inputs.

3.2. Sensitivity indices

As we discussed in the previous section, PCE models can be constructed for models with *dependent* inputs based on the modified Gram-Schmidt algorithm. In this section, we first propose how to use PCE models to estimate the *full* sensitivity indices and the *uncorrelated* sensitivity indices based on data. Then we define the conditional order-based sensitivity indices and present how they can be used to exclude higher order interaction terms in a PCE model. For easy reference, we include in Appendix A.1 a list of sensitivity index symbols used in this paper.

3.2.1. Full sensitivity indices

Constructing orthonormal polynomials using the modified Gram-Schmidt algorithm requires a linearly independent set of polynomials. A PCE model with n inputs and the highest polynomial order p is composed of $P+1$ terms of polynomials as we defined in Eqs. (2) and (3). Assume polynomials in the set

$$S = \left\{ \prod_{l=1}^{n} X_l^{j_l} : j_l \in \{0, 1, \dots p\}, \sum_{l=1}^{n} j_l \leq p \right\} \tag{5}$$

are linearly independent.

Orthonormal polynomials can be constructed using the modified Gram-Schmidt algorithm with respect to a specific order of the polynomials. Suppose we order the polynomials in S as $(St_0, St_{11}, St_1 \backslash St_{11}, St_2, St_3, \dots, St_n)$, where $St_0 = \{1\}$, St_{11} and St_i are defined as follows:

$$St_{11} = \left\{ X_1^{j_1} : j_1 \in \{1, \dots p\} \right\},$$

$$St_i = \left(S \backslash \bigcup_{j=0}^{i-1} St_j \right) \bigcap \left\{ \prod_{l=1}^{n} X_l^{j_l} : j_i \in \{1, 2, \dots, p\}, j_{l \neq i} \in \{0, 1, \dots, p\}, \sum_{l=1}^{n} j_l \leq p \right\}. \tag{6}$$

$(St_0, St_{11}, St_1 \backslash St_{11}, St_2, St_3, \dots, St_n)$ is an ordered partition of the set S. In the partition, St_0, St_{11}, $St_1 \backslash St_{11}$, and $St_i, i = 2, 3, \dots, n$ are ordered in sequence but the polynomials in each set can be in any arbitrary order. Note that St_1 contains all the polynomial functions of X_1 and the interaction terms between X_1 and the rest of the inputs. St_2 contains all the polynomial functions of X_2 and the interaction terms between X_2 and the rest of the inputs *except* X_1. St_3 contains all the polynomial functions of X_3 and the interaction terms between X_3 and the rest of the inputs *except* X_1 and X_2.

For example, St_{11} and $St_i, i = 1, 2, 3$ for constructing a PCE model with inputs $\{X_1, X_2, X_3\}$ and the highest polynomial order $p = 3$ are defined as follows:

$$
\begin{aligned}
&St_{11} = \left\{X_1, X_1^2, X_1^3\right\}, \\
&St_1 = \left\{X_1, X_1^2, X_1^3, X_1X_2, X_1^2X_2, X_1X_2^2, X_1X_3, X_1^2X_3, X_1X_3^2, X_1X_2X_3\right\}, \\
&St_2 = \left\{X_2, X_2^2, X_2^3, X_2X_3, X_2^2X_3, X_2X_3^2\right\}, \\
&St_3 = \left\{X_3, X_3^2, X_3^3\right\}.
\end{aligned}
$$

After constructing the orthonormal polynomials using the ordered partition $(St_0, St_{11}, St_1 \backslash St_{11}, St_2, St_3, \dots, St_n)$, the first-order full sensitivity index $\bar{S}_{X_1}$ for X_1 can be estimated as follows:

$$\bar{S}_{X_1} \approx \frac{\sum_{j \in St_{11}} \theta_j^2}{Var(Y)}, \tag{7}$$

where θ_j's are the PCE coefficients corresponding to the orthonormal polynomials in the set St_{11}.

In addition, we propose an alternative total full sensitivity index

$$\overline{ST}_{X_1} = \frac{\sum_{u \ni X_1} D_u(Y)}{Var(Y)}$$

and estimate it using

$$\overline{ST}_{X_1} \approx \frac{\sum_{j \in St_1} \theta_j^2}{Var(Y)}. \tag{8}$$

This total full sensitivity index is different from the one defined in [16], which is obtained after transforming the dependent inputs into the independent inputs. Instead, the total full sensitivity index in Eq. (8) has dependent effects of X_1 with other inputs. By permuting the order of the input $(X_1, X_2, X_3, \dots, X_n)$ as $(X_i, X_{i+1}, \dots, X_n, X_1, \dots, X_{i-1})$, any $\bar{S}_{X_i}$ and $\overline{ST}_{X_i}$ can be estimated.

We also define the conditional total sensitivity indices for models with *dependent* inputs as follows:

$\overline{ST}_{X_2|X_1} = \frac{\sum_{u \ni \{X_1, X_2\}} D_u(Y) - \sum_{u \ni X_1} D_u(Y)}{Var(Y)}$ is the total contribution of input X_2 to the variance of output Y after taking account of the total full contribution of X_1.

$\overline{ST}_{X_3|X_1,X_2} = \frac{\sum_{u\ni\{X_1,X_2,X_3\}} D_u(Y) - \sum_{u\ni\{X_1,X_2\}} D_u(Y)}{Var(Y)}$ is the total contribution of input X_3 to the variance of output Y after taking account of the total full contributions of X_1 and X_2.

$\vdots$

$\overline{ST}_{X_n|X_1,X_2,\ldots,X_{n-1}} = \frac{\sum_{u\ni\{X_1,X_2,\ldots,X_n\}} D_u(Y) - \sum_{u\ni\{X_1,X_2,\ldots,X_{n-1}\}} D_u(Y)}{Var(Y)}$ is the total contribution of input X_n to the variance of output Y after taking account of the total full contributions of X_1, X_2, $\ldots$, X_{n-1}.

We estimate the conditional total sensitivity indices using

$$\overline{ST}_{X_i|X_1,X_2,\ldots,X_{i-1}} \approx \frac{\sum_{j\in St_i} \theta_j^2}{Var(Y)}$$

for $i = 2, 3, \ldots, n$

When inputs can be grouped such that inputs from different groups have neither dependence nor interaction across groups, the total Sobol index of a group can be estimated based on the conditional total sensitivity indices. For example, if the first d inputs are independent of and have no interactions with the rest of the inputs, $\sum_{i=1}^{d} \overline{ST}_{X_i}$ estimates the total Sobol index of the first d inputs. Eq. (10) for Example 3 in Section 4 illustrates how this sensitivity index can be used in practice.

3.2.2. Uncorrelated sensitivity indices

In order to estimate the total *uncorrelated* sensitivity index of X_1, we consider the ordered partition of S as $(St_0, St_{-1}, St_{11}, St_1\backslash St_{11})$, where $St_0 = \{1\}$ and $St_{-1} = \bigcup_{j=2}^{n} St_j$. St_{11} and $St_i, i = 1, 2, \ldots, n$ are defined in Eq. (6). Then the orthonormal polynomials can be constructed with respect to this ordered partition. As we obtain the PCE coefficients corresponding to the orthonormal polynomials in the set St_1, the total *uncorrelated* sensitivity index ($ST^u_{X_1}$) can be estimated using Eq. (8).

In addition, we propose an alternative first-order *uncorrelated* sensitivity index ($S^u_{X_1}$) and estimate it using Eq. (7). The proposed first-order *uncorrelated* sensitivity index is different from the one defined in [16] because the latter is estimated after decorrelating X_1 with all the other inputs. Note that the proposed first-order *uncorrelated* sensitivity index is estimated by decorrelating the polynomials of the inputs. By permuting the order of the inputs $(X_1, X_2, X_3, \dots, X_n)$ as $(X_i, X_{i+1}, \dots, X_n, X_1, \dots, X_{i-1})$, any $S^u_{X_i}$ and $ST^u_{X_i}$ can be estimated.

Note that the first-order full (resp. uncorrelated) sensitivity indices are always smaller or equal to the total full (resp. uncorrelated) sensitivity indices, but the first-order full (resp. uncorrelated) sensitivity indices are not necessarily larger or smaller than the total uncorrelated (resp. full) sensitivity indices.

3.2.3. Conditional order-based sensitivity indices

In order to reduce the complexity of a PCE model and select appropriate interaction terms in the PCE model, we propose the conditional order-based sensitivity indices.

Suppose we order the polynomials in the polynomial set S in Eq. (5) as $(Sc_0, Sc_{11}, Sc_{12}, \dots, Sc_{1p}, Sc_{22}, Sc_{23}, \dots, Sc_{2p}, \dots, Sc_{kk}, Sc_{kk+1}, \dots, Sc_{kp})$, where $k = \min(n, p)$, $Sc_0 = \{1\}$, and $Sc_{ij}, i = 1, 2, \dots, k;\ j = i, i+1, \dots, p$, are defined as follows:

$$Sc_{ij} = \left\{ \prod_{l=1}^{n} X_l^{j_l} : j_l \in \{0, 1, \dots p\}, \sum_{l=1}^{n} \mathbb{1}_{j_l \neq 0} = i, \sum_{l=1}^{n} j_l = j \right\},$$

where

$$\mathbb{1}_{j_l \neq 0} = \begin{cases} 1 & j_l \neq 0 \\ 0 & j_l = 0 \end{cases}.$$

Define $Sc_i = \cup_{j=i}^{p} Sc_{ij}, i \leq \min(n, p)$, then Sc_1 contains all the polynomial functions of $X_i, i = 1, 2, \dots, n$. Sc_2 contains all the two-way interaction terms. Sc_3 contains all the three-way interaction terms. Note that $(Sc_0, Sc_{11}, Sc_{12}, \dots, Sc_{1p}, Sc_{22}, Sc_{23}, \dots, Sc_{2p}, \dots, Sc_{kk}, Sc_{kk+1}, \dots, Sc_{kp})$, where $k = \min(n, p)$, is an ordered partition of the set S. In the partition, $Sc_0, Sc_{11}, Sc_{12}, \dots, Sc_{1p}, Sc_{22}$,-$Sc_{23}, \dots, Sc_{2p}, \dots, Sc_{kk}, Sc_{kk+1}, \dots, Sc_{kp}$, are ordered in sequence but the polynomials in each set Sc_{ij} can be in any arbitrary order.

For example, $Sc_{11}, Sc_{12}, Sc_{13}, Sc_{22}, Sc_{23},$ and Sc_{33} for constructing a PCE model with inputs $\{X_1, X_2, X_3\}$ and the highest polynomial order $p = 3$ are defined as follows:

$$\begin{aligned}
&Sc_{11} = \{X_1, X_2, X_3\}\,, \\
&Sc_{12} = \left\{X_1^2, X_2^2, X_3^2\right\}, \\
&Sc_{13} = \left\{X_1^3, X_2^3, X_3^3\right\}, \\
&Sc_{22} = \{X_1X_2, X_2X_3, X_1X_3\}\,, \\
&Sc_{23} = \left\{X_1^2X_2, X_1^2X_3, X_1X_2^2, X_1X_3^2, X_2^2X_3, X_2X_3^2,\right\}, \\
&Sc_{33} = \{X_1X_2X_3\}\,.
\end{aligned}$$

We define the conditional order-based sensitivity indices as follows:

$\tilde{S}_1 = \frac{\sum_{i=1}^{n} D_i(Y)}{Var(Y)}$ is the first order sensitivity index of the output Y with respect to the inputs $\boldsymbol{X}$.
$\tilde{S}_{2|1} = \frac{\sum_{i<j} D_{ij}(Y)}{Var(Y)}$ is the second order sensitivity index of the output Y with respect to the inputs $\boldsymbol{X}$ after taking account of the first order sensitivity index.
$\tilde{S}_{3|1,2} = \frac{\sum_{i<j<k} D_{ijk}(Y)}{Var(Y)}$ is the third order sensitivity index of the output Y with respect to the inputs $\boldsymbol{X}$ after taking account of the first order sensitivity index and the second order sensitivity index.
$\vdots$
$\tilde{S}_{k|1,2,\ldots,k-1} = \frac{D_{1,2,3,\ldots,k}(Y)}{Var(Y)}$ is the k^{th} order sensitivity index of the output Y with respect to the inputs $\boldsymbol{X}$ after taking account of the first order sensitivity index through the $(k-1)^{th}$ order sensitivity index.

As we can obtain the PCE coefficients corresponding to the orthonormal polynomials constructed from $(Sc_0, Sc_{11}, Sc_{12}, \ldots, Sc_{1p}, Sc_{22}, Sc_{23}, \ldots, Sc_{2p}, \ldots,$ $Sc_{kk}, Sc_{kk+1}, \ldots, Sc_{kp})$, where $k = \min(n, p)$, using the modified Gram-Schmidt algorithm, the above sensitivity indices can be estimated as follows:

$$\tilde{S}_1 \approx \frac{\sum_{j \in Sc_1} \theta_j^2}{Var(\boldsymbol{Y})},$$
$$\tilde{S}_{i|1,\ldots,i-1} \approx \frac{\sum_{j \in Sc_i} \theta_j^2}{Var(\boldsymbol{Y})}, \;\; i = 2, \ldots, \min(n, p),$$

where θ_j's are the PCE coefficients corresponding to the orthonormal polynomials in the set Sc_i for $i \leq \min(n,p)$. Note that for a full PCE model, Sc_i contains $\binom{n}{i}\binom{p}{i}$ polynomial terms.

The conditional order-based sensitivity indices serve the purpose of identifying up to which *order of interaction* of inputs significantly influences the output variance. Specifically, if the cumulative sum of conditional order-based sensitivity indices, $\sum_{i=1}^{d} \tilde{S}_i$, is close to one for a certain polynomial order, $d \leq \min(n,p)$, it indicates that the interaction terms of orders higher than d can be excluded in the PCE model. Using a simple procedure of inspecting the cumulative sum for different d's, we can identify and remove unnecessary high-order interaction terms from the PCE model. In constrast to the existing methods of constructing a sparse PCE model [30, 31, 32, 33], this procedure keeps the effect hierarchy principle [34] while improving the parsimony of the PCE model. Example 3 in Section 4 illustrates how this procedure can be used to determine the highest polynomial order.

While the conditional order-based sensitivity indices are useful for effective PCE modeling and, in turn, PCE-based sensitivity analyses, the new sensitivity indices do not directly serve the traditional purposes of sensitivity indices. In contrast, for example, the first-order full sensitivity indices and the total uncorrelated sensitivity indices directly help determine influential and non-influential inputs, respectively, in terms of their contributions to the output variance (also known as factor prioritization and factor fixing, respectively, in [35]; see also [16]).

4. Numerical examples

To validate the proposed data-driven sensitivity indices, this section presents four examples where inputs are dependent. We present our experiment results based on 500 replications with 95% confidence intervals wherever applicable. The confidence intervals are computed using $10,000$ bootstrap samples of the 500 replications to improve upon the accuracy of the empirical confidence interval computed from the 500 replications [36].

4.1. Example 1

We use a benchmark example in [16] as our first validation case. In this case, inputs follow a three-dimensional multivariate normal distribution as follows:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}\right].$$

The output Y is simply modeled using a linear model $Y = X_1 + X_2 + X_3$.

Table 1: Sample mean of sensitivity indices and 95% confidence intervals.

$(\rho_{12}, \rho_{13}, \rho_{23})$	Input	$\bar{S}_{X_i}$		$ST^u_{X_i}$	
		Analytical method[†]	Proposed method[‡]	Analytical method[†]	Proposed method[‡]
(0.5,0.8,0)	X_1	0.945	$0.945\ (0.945, 0.945)$	0.020	$0.020\ (0.020, 0.020)$
	X_2	0.402	$0.401\ (0.400, 0.402)$	0.055	$0.055\ (0.054, 0.055)$
	X_3	0.579	$0.579\ (0.578, 0.579)$	0.026	$0.026\ (0.026, 0.026)$
(-0.5,0.2,-0.7)	X_1	0.490	$0.491\ (0.490, 0.492)$	0.706	$0.707\ (0.706, 0.707)$
	X_2	0.040	$0.041\ (0.040, 0.041)$	0.375	$0.374\ (0.373, 0.374)$
	X_3	0.250	$0.250\ (0.249, 0.251)$	0.480	$0.480\ (0.479, 0.481)$
(-0.49,-0.49,-0.49)	X_1	0.007	$0.007\ (0.007, 0.007)$	0.974	$0.974\ (0.974, 0.974)$
	X_2	0.007	$0.007\ (0.006, 0.007)$	0.974	$0.974\ (0.974, 0.974)$
	X_3	0.007	$0.007\ (0.007, 0.007)$	0.974	$0.974\ (0.973, 0.974)$

Note: †The value presented in [16] differs by up to 0.01 due to rounding. ‡The proposed method does not require any distributional assumption. The sample means of the sensitivity indices and 95% bootstrap confidence intervals (using 10,000 bootstrap samples) are calculated based on 500 simulation replications where each replication uses 500 random observations.

Table 1 shows the first-order full sensitivity index $\bar{S}_{X_i}$ and total uncorrelated sensitivity index $ST^u_{X_i}$ for each input based on the analytical method [16]. These true indices are compared with the estimated indices from the proposed method. This example validates that the proposed method can estimate the first-order full sensitivity index and total uncorrelated sensitivity index based only on data without the knowledge of the distribution of dependent inputs and the model structure. Note that in this example, the *total* full sensitivity index is the same as the *first-order* full sensitivity index (i.e., $\overline{ST}_{X_i} = \bar{S}_{X_i}$) and that the *first-order* uncorrelated sensitivity index equals the *total* uncorrelated sensitivity index (i.e., $S^u_{X_i} = ST^u_{X_i}$) because there is no interaction effect. Thus, $\overline{ST}_{X_i}$ and $S^u_{X_i}$ are not presented.

4.2. Example 2

In this example from [18], the output Y is a non-linear function of four dependent inputs: $Y = X_1X_2 + X_3X_4$. Here, $(X_1, X_2) \in \left[0,1\right]^2$ is uniformly distributed within the triangle $X_1 + X_2 \leq 1$ and $(X_3, X_4) \in \left[0,1\right]^2$ is uniformly distributed within the triangle $X_1+X_2 \geq 1$. Due to the symmetry of the model, the sensitivity indices of Y with respect to X_1 and X_3 are equal to those with respect to X_2 and X_4, respectively.

As shown in Table 2, the proposed method yields the estimates that are close to the analytical values of $\bar{S}_{X_i}$ and $ST^u_{X_i}$. In contrast to the benchmark method in [18] that requires the knowledge of joint probability distribution of the inputs, the proposed method is purely data-driven. $ST_{\{X_1,X_2\}}$ (or, $ST_{\{X_3,X_4\}}$) is estimated using $\sum_{i=1}^{2} \overline{ST}_{X_i}$ by permuting the inputs as (X_1, X_2, X_3, X_4) (or, (X_3, X_4, X_1, X_2)).

Figure 1 shows intricate working of the model by revealing how each input influences the output variance. The *total* full (uncorrelated) sensitivity index $\overline{ST}_{X_i}$ ($ST^u_{X_i}$) can be decomposed into the *first-order* full (uncorrelated) sensitivity index $\bar{S}_{X_i}$ ($S^u_{X_i}$) and the *rest of the total* effect, $\overline{ST}_{X_i} - \bar{S}_{X_i}$ ($ST^u_{X_i} - S^u_{X_i}$), which accounts for all the interactions of X_i. The gap between the two lines on the left (right) graph in Figure 1 shows the magnitude of $\overline{ST}_{X_i} - \bar{S}_{X_i}$

Table 2: Sample mean of sensitivity indices and 95% confidence intervals.

Method	$\bar{S}_{X_1}$	$ST^u_{X_2}$	$ST_{\{X_1,X_2\}}$	$\bar{S}_{X_3}$	$ST^u_{X_4}$	$ST_{\{X_3,X_4\}}$
Analytical method[†]	0.033	0.067	0.100	0.233	0.666	0.900
Benchmark method[‡]	0.032	0.071	0.103	0.226	0.669	0.895
	(0.028, 0.037)	(0.066,0.077)	(0.095,0.114)	(0.209,0.248)	(0.639,0.705)	(0.848, 953)
Proposed method[*]	0.035	0.066	0.101	0.233	0.663	0.896
	(0.034, 0.036)	(0.066,0.067)	(0.100,0.103)	(0.231,0.235)	(0.661,0.666)	(0.892, 901)

Note: †The value is provided in [18]. ‡The value is estimated using the method proposed in [18] and the confidence interval is calculated based on 16,380 random observations under the assumption that the joint probability distribution of the inputs is *known*. *The proposed method does not require any distributional assumption. The sample means of sensitivity indices and 95% bootstrap confidence intervals (using $10,000$ bootstrap samples) are calculated based on 500 replications. In each replication, sensitivity indices are calculated using 500 random observations.

$(ST^u_{X_i} - S^u_{X_i})$, indicating how much of the total effect of X_i is attributed to the interaction effects compared to the first-order effect when we consider the full (uncorrelated) contribution of X_i.

4.3. Example 3

For the third example, we modified the example in [37] to have a more complex structure and involve multiple types of probability distributions as follows:

$$\begin{aligned}
\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} &\sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.3 \\ 0 & 0 & 0.3 & 1 \end{pmatrix}\right], \\
X &\sim \mathcal{U}(0,\,1), \\
X_5 &= \theta_1 X + \mathcal{U}(0,\,1), \\
X_6 &= \theta_2 X + \theta_3 X^2 + \mathcal{U}(0,\,1), \\
Y &= X_1X_2 + X_3X_4 + X_5X_6.
\end{aligned} \tag{9}$$

Here, (X_1, X_2, X_3, X_4) follows a multivariate Gaussian distribution with the parameters as above. The inputs X_5 and X_6 are dependent on each other, but

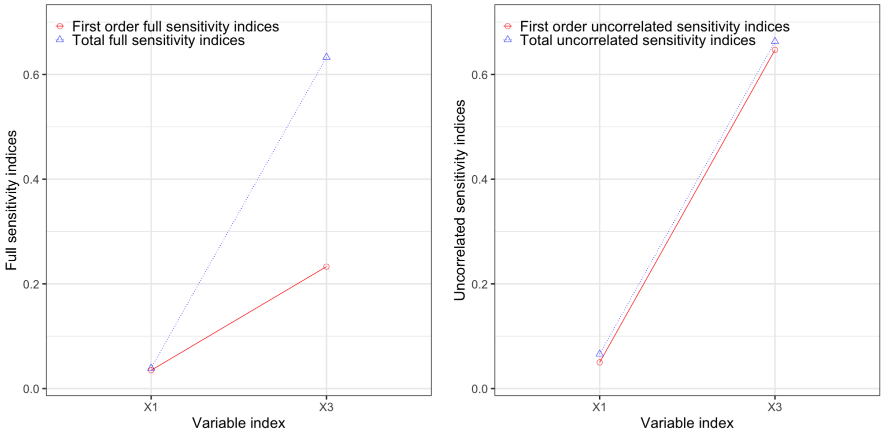

Figure 1: The left-hand side graph shows the *full* sensitivity indices for dependent input variables and the right-hand side graph shows the *uncorrelated* sensitivity indices for dependent input variables in Example 2.

their dependency cannot be explained by their first-order conditional moments. In this experiment, we set $(\theta_1, \theta_2, \theta_3) = (0.4, 0.6, 1)$ and obtain 10,000 random observations. Because the cumulative sum of the first two conditional order-based sensitivity indices is $\sum_{i=1}^{2} \tilde{S}_i = 1$, we exclude third- and higher-order interaction terms in the PCE model to make it sparse. As for the two-way interaction terms, as shown in Figure 2, most of the corresponding PCE coefficients are nearly zero except for the polynomial terms, X_1X_2, X_3X_4, and X_5X_6. It indicates that X_1X_2, X_3X_4, and X_5X_6 are the only interaction terms in the true model. Various sparse PCE approaches [30, 31, 32, 33] can be additionally applied here to construct a sparser PCE model with only the important orthonormal polynomials of inputs.

Because $\{X_1, X_2\}$, $\{X_3, X_4\}$, and $\{X_5, X_6\}$ are mutually independent, we can infer from the conditional order-based sensitivity indices that the output Y is composed of three non-interacting functions $f_{12}(X_1, X_2)$, $f_{34}(X_3, X_4)$, and $f_{56}(X_5, X_6)$. Thanks to this special structure, we can directly calculate the total sensitivity indices for $\{X_1, X_2\}$, $\{X_3, X_4\}$, and $\{X_5, X_6\}$. Without permuting

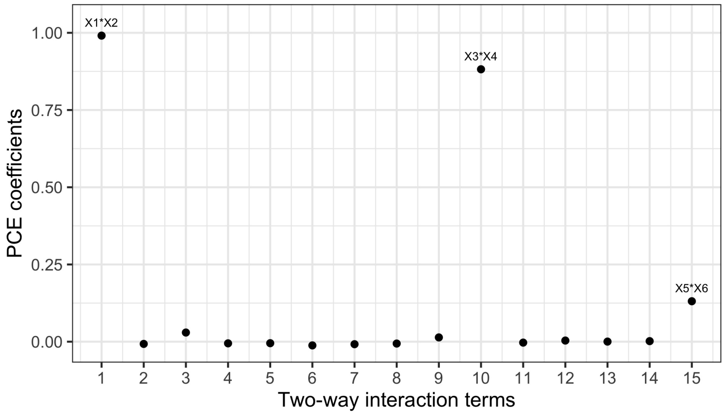

Figure 2: PCE coefficients v.s. the two-way interaction terms in the PCE model in Example 3. The significant interaction terms, X_1X_2, X_3X_4, and X_5X_6, are identified.

the order of the input variables, the total Sobol indices can be calculated as follows:

$$\begin{aligned} ST_{\{X_1,X_2\}} &= \overline{ST}_{X_1} + \overline{ST}_{X_2}, \\ ST_{\{X_3,X_4\}} &= \overline{ST}_{X_3} + \overline{ST}_{X_4}, \\ ST_{\{X_5,X_6\}} &= \overline{ST}_{X_5} + \overline{ST}_{X_6}, \end{aligned} \tag{10}$$

where $\overline{ST}_{X_i}$ is the conditional *total full* sensitivity index defined in Section 3.2.1.

We validate the sensitivity indices from the proposed method with the values from an analytical method (see Appendix A.2). We also calculate $ST_{\{X_1,X_2\}}$ and $ST_{\{X_3,X_4\}}$ using the benchmark method in [16] assuming the knowledge that (X_1, X_2, X_3, X_4) is multivariate Gaussian distributed and $\{X_1, X_2\}$, $\{X_3, X_4\}$, and $\{X_5, X_6\}$ are mutually independent. Then, $ST_{\{X_5,X_6\}}$ is calculated based on the fact that $ST_{\{X_1,X_2\}} + ST_{\{X_3,X_4\}} + ST_{\{X_5,X_6\}} = 1$.

As shown in Table 3, the sensitivity indices from our method are close to . those from the benchmark method and the analytical values. Note that, in contrast to the benchmark method, the proposed method is a data-driven approach that does not impose any assumption on the inputs.

4.4. Example 4

As the fourth example, we consider the 23-bar horizontal truss example in [38]. The output of interest, Y, is a downward vertical displacement at the mid span of the structure subject to random loads. As depicted in Figure 3, the uncertainty of Y depends on the ten random inputs in $\boldsymbol{X} = (E_1, E_2, A_1, A_2, P_1, \ldots, P_6)$:

Table 3: Sample means of sensitivity indices and 95% confidence intervals.

Input set Method	$ST_{\{X_1,X_2\}}$	$ST_{\{X_3,X_4\}}$	$ST_{\{X_5,X_6\}}$
Analytical method	0.402	0.438	0.160
Proposed method*	0.402	0.438	0.160
	(0.401, 0.404)	(0.437, 0.439)	(0.159, 0.160)
Benchmark method[16]	0.403†	0.439†	(0.158)‡
	(0.402, 0.404)	(0.438, 0.440)	(0.157, 0.160)

Note: †The value is obtained using the sample variance to estimate $Var(Y)$ in the denominator of the sensitivity index instead of using the PCE coefficients from the benchmark method (see Eq. (4)) because the latter estimation suffers a non-negligible bias in this example that does not satisfy the assumption of the benchmark method. ‡The value cannot be obtained directly from the benchmark method, but we calculate the value based on the assumption that the user knows that X_5 and X_6 are independent of the rest of the inputs and that (X_1, X_2, X_3, X_4) follows a multivariate Gaussian distribution. *The proposed method does not require any assumption. The sample means of sensitivity indices and 95% bootstrap confidence intervals (using 10,000 bootstrap samples) are calculated based on 500 replications. In each replication, sensitivity indices are calculated using 5,000 random observations of the inputs and output in Eq. (9).

namely, uncertain Young modulus $E_i, i = 1, 2$, and uncertain cross-sectional area $A_i, i = 1, 2$, for two different groups of bars (horizontal for $i = 1$ and diagonal for $i = 2$), and the random loads $P_i, i = 1, 2, \cdots, 6$. The inputs $E_i, i = 1, 2$, and $A_i, i = 1, 2$, are assumed to be mutually independent and follow the following distributions:

$$E_1, E_2 \sim \mathcal{LN}(2.1 \times 10^{11}, 2.1 \times 10^{10}) \text{ [Pa]},$$

$$A_1 \sim \mathcal{LN}(2.0 \times 10^{-3}, 2.0 \times 10^{-4}) \text{ [m}^2\text{]},$$

$$A_2 \sim \mathcal{LN}(1.0 \times 10^{-3}, 1.0 \times 10^{-4}) \text{ [m}^2\text{]},$$

where $\mathcal{LN}(\mu, \sigma)$ denotes the lognormal distribution with mean μ and standard deviation σ.

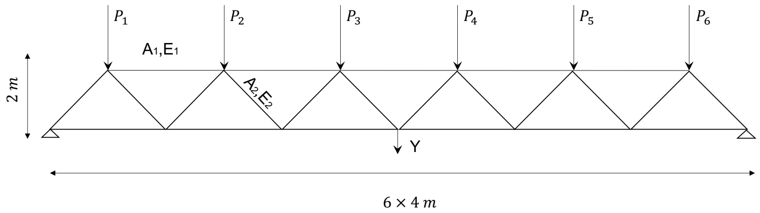

Figure 3: Scheme of the horizontal truss model modified from [39]. The downward vertical displacement at the mid span of the structure Y depends on Young modulus $E_i, i = 1, 2$, cross-sectional area $A_i, i = 1, 2$ for both horizontal and diagonal bars, and the random loads $P_i, i = 1, 2, \cdots, 6$.

The dependent inputs $P_i, i = 1, 2, \cdots, 6$, have the following Gumbel marginal distribution function with mean $\mu = 5 \times 10^4$ [N] and standard deviation $\sigma = 7.5 \times 10^3$ [N]:

$$F_i(x; \alpha, \beta) = e^{-e^{-(x-\alpha)/\beta}}, i = 1, 2, \ldots, 6,$$

where $\beta = \sqrt{6}\sigma/\pi$, $\alpha = \mu - \gamma\beta$, and $\gamma \approx 0.5772$ is the Euler-Mascheroni constant. The dependence between the inputs $P_i, i = 1, \ldots, 6$, is encoded in the C-vine copula with the following density:

$$c_{\boldsymbol{X}}^{(\mathcal{G})}(u_1, \ldots, u_6) = \prod_{j=2}^{6} c_{1j;\theta=1.1}^{(\mathcal{GH})}(u_1, u_j), \tag{11}$$

where $c_{1j;\theta=1.1}^{(\mathcal{GH})}$ is the density of the pair-copula between P_1 and P_j, $j = 2, \ldots, 6$. $\mathcal{GH}$ represents the Gumbel-Hougaard family whose bivariate copula is

$$C_\theta^{(\mathcal{GH})}(u, v) = \exp\left(-\left((-\log u)^\theta + (-\log v)^\theta\right)^{1/\theta}\right), \quad \theta \in [1, \infty).$$

The parameter $\boldsymbol{\theta}$ decides the dependence between two loads (i.e., the larger parameter $\boldsymbol{\theta}$ the stronger dependence). In this case, P_1 is equally and positively correlated with each of $P_2, \ldots, P_6$. Thus, $P_2, \ldots, P_6$ are positively correlated with each other although they are conditionally independent given P_1 (see a sample correlation matrix for the loads in Appendix A.3). The output Y is simulated using the response surface model (see Appendix A.4) in [39], which was constructed based on a finite element analysis. The explicit relationship between Y and $\boldsymbol{X}$ in the response surface model allows us to evaluate the estimated sensitivity indices.

This realistic problem with dependent inputs has neither analytically known sensitivity indices nor any benchmark methods that attempted to estimate the sensitivity indices (see [31] for a related sensitivity analysis with *independent* inputs). Implementing a brute-force Monte Carlo approach is computationally challenging, if not infeasible, because the analytical expressions of the sensitivity indices involve variances of conditional expectations that condition on (multiple combinations of) multiple inputs. A similar challenge lies in even estimating Sobol indices for *independent* inputs and is studied extensively in the literature [40, 41, 42]. No Monte Carlo method has satisfactorily addressed the computational challenge yet. Extending the existing Monte Carlo methods for independent inputs to handle dependent inputs is left for future work.

In this study, we examine the estimated sensitivity indices to confirm that they agree with their expected physical interpretations. As it is shown in Table 4, P_2 (resp. P_3) and P_5 (resp. P_4) have almost the same sensitivity indices for $\bar{S}$, $\overline{ST}$, S^u, and ST^u. This can be explained by a) the physical symmetry between P_2 (resp. P_3) and P_5 (resp. P_4) with respect to the location at which the output Y is measured (see Figure 3 and the response surface model in Appendix A.4) and b) their symmetric correlations with other inputs (recall the copula density in Eq. (11) and see Appendix A.3). On the other hand, the differences between the *full* sensitivity indices (i.e., $\bar{S}$ and $\overline{ST}$) for P_1 and P_6 can be explained by the fact that P_1 has over 5 times stronger correlations than P_6 with $P_2, \ldots, P_5$ (see Appendix A.3). In contrast, the *uncorrelated* sensitivity indices (i.e., S^u and ST^u) for P_1 and P_6 are the same (up to 4 decimal places) because the uncorrelated effects of P_1 and P_6 on Y should be very similar (see the coefficients of the response surface model in Appendix A.4.). The sensitivity indices for all the other inputs are similarly confirmed to be consistent with their expected physical interpretations based on the response surface model, which reflects the physical relationship between Y and $\boldsymbol{X}$, and the dependence structure of $\boldsymbol{X}$.

In addition, although not directly comparable due to different settings, still the first and total uncorrelated sensitivity indices (i.e., S^u and ST^u) have similar magnitudes as the first and total Sobol indices (i.e., S and ST) reported in Table 5 of [7] and Table 2 of [31], respectively, for all ten inputs. Both articles [7, 31] assumed that P_1 through P_6 are *independent*, and directly computed Y using a

Table 4: Sample means of sensitivity indices and 95% confidence intervals.

Sensitivity indices / Input	$\bar{S}$	$\overline{ST}$	S^u	ST^u
E_1	0.324 (0.321, 0.327)	0.371 (0.367, 0.374)	0.286 (0.284, 0.288)	0.312 (0.310, 0.314)
E_2	0.013 (0.012, 0.014)	0.036 (0.035, 0.037)	0.009 (0.008, 0.009)	0.009 (0.008, 0.009)
A_1	0.325 (0.322, 0.328)	0.370 (0.367, 0.374)	0.285 (0.283, 0.287)	0.310 (0.308, 0.312)
A_2	0.013 (0.012, 0.014)	0.037 (0.036, 0.038)	0.008 (0.008, 0.008)	0.008 (0.008, 0.008)
P_1	0.065 (0.063, 0.068)	0.096 (0.093, 0.099)	0.004 (0.004, 0.004)	0.004 (0.004, 0.004)
P_2	0.060 (0.058, 0.062)	0.088 (0.085, 0.090)	0.033 (0.033, 0.033)	0.035 (0.035, 0.036)
P_3	0.105 (0.103, 0.108)	0.135 (0.132, 0.138)	0.068 (0.067, 0.068)	0.073 (0.072, 0.073)
P_4	0.102 (0.010, 0.105)	0.130 (0.128, 0.133)	0.068 (0.067, 0.068)	0.073 (0.072, 0.073)
P_5	0.057 (0.055, 0.059)	0.084 (0.082, 0.087)	0.033 (0.033, 0.033)	0.035 (0.035, 0.036)
P_6	0.017 (0.016, 0.018)	0.043 (0.042, 0.045)	0.004 (0.004, 0.004)	0.004 (0.004, 0.004)

Note: The sample means of sensitivity indices and 95% bootstrap confidence intervals (using $10,000$ bootstrap samples) are calculated based on 500 replications. In each replication, sensitivity indices are calculated using 500 random observations of the inputs and outputs in Eq. (12).

finite element model.

5. Conclusion

In this paper, data-driven sensitivity indices for a model with *dependent* inputs are proposed using the PCE without imposing any strong assumptions on the model inputs. The modified Gram-Schmidt algorithm with the empirical measure is utilized to construct orthonormal polynomials for a PCE model on the merit of numerical stability. The proposed data-driven method yields the *full* sensitivity indices and the *uncorrelated* sensitivity indices by constructing ordered partitions of orthonormal polynomials of inputs for a PCE model. The proposed conditional order-based sensitivity indices for a model with *dependent* inputs help reduce the complexity of a PCE model while keeping the effect hierarchy principle. Four numerical examples validate the proposed method.

The proposed method requires polynomials of inputs, which are fed into the modified Gram-Schmidt algorithm, to be linearly independent. This suggests a future research direction because there are multiple ways of constructing linearly independent polynomials from a linearly dependent polynomial basis. How to build a theoretically and practically desirable basis warrants more investigation.

Appendix

A.1. List of sensitivity indices

$ST^u_{X_i}$ Total *uncorrelated* sensitivity index of the output Y with respect to the input X_i.

ST_{X_i} Total Sobol index of the output Y with respect to the input X_i.

$S^u_{X_i}$ First-order *uncorrelated* sensitivity index of the output Y with respect to the input X_i.

S_{X_i} First-order Sobol index of the output Y with respect to the input X_i.

$\bar{S}_{X_i}$ First-order *full* sensitivity index of the output Y with respect to the input X_i.

$\overline{ST}_{X_i}$ Total *full* sensitivity index of the output Y with respect to the input X_i.

$\tilde{S}_{k|1,2,\ldots,k-1}$ The k^{th} order sensitivity index of the output Y with respect to the inputs $\boldsymbol{X}$ after taking account of the first order sensitivity index through the $(k-1)^{th}$ order sensitivity index.

A.2. Analytical method for calculating the sensitivity indices in Example 3 in Section 4

The following lemma is used for the analytical method in Example 3 in Section 4.

Lemma 1. *Suppose*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right].$$

Then, $Var(X_1X_2) = \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + \sigma_1^2\sigma_2^2 + 2\mu_1\mu_2\rho\sigma_1\sigma_2 + \rho^2\sigma_1^2\sigma_2^2$.

Proof. We can express X_1 and X_2 as follows:

$$X_1 = \mu_1 + r\sigma_1 Z + (1 - r^2)^{\frac{1}{2}}\sigma_1 Y_1,$$
$$X_2 = \mu_2 + r\sigma_2 Z + (1 - r^2)^{\frac{1}{2}}\sigma_2 Y_2,$$

where $r = \sqrt{\rho}$, $Z \sim \mathcal{N}(0, 1)$, and $Y_i \sim \mathcal{N}(0, 1), i = 1, 2$. We have the covariance among Z and $Y_i, i = 1, 2$ as follows:

$$Cov(Z, Y_i) = 0, \text{ for } i = 1, 2,$$
$$Cov(Y_1, Y_2) = 0.$$

Therefore, we have

$$\begin{aligned}
Var(X_1X_2) &= E(X_1^2X_2^2) - [E(X_1X_2)]^2\\
&= E(X_1^2)E(X_2^2) + Cov(X_1^2, X_2^2) - [E(X_1)E(X_2) + Cov(X_1, X_2)]^2\\
&= (\sigma_1^2 + \mu_1^2)(\sigma_2^2 + \mu_2^2) + Cov(r^2\sigma_1^2Z^2 + 2\mu_1 r\sigma_1 Z, r^2\sigma_2^2Z^2 + 2\mu_2 r\sigma_2 Z)-\\
&\quad (\mu_1\mu_2 + r^2\sigma_1\sigma_2)^2\\
&= \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 + \sigma_1^2\sigma_2^2 + 2\mu_1\mu_2\rho\sigma_1\sigma_2 + \rho^2\sigma_1^2\sigma_2^2.
\end{aligned}$$

□

We now present how to analytically calculate the sensitivity indices in Example 3. From (9), $\{X_1, X_2\}$, $\{X_3, X_4\}$, and $\{X_5, X_6\}$ are mutually independent. We have

$$\begin{aligned}
Var(Y) &= Var(X_1X_2 + X_3X_4 + X_5X_6)\\
&= Var(X_1X_2) + Var(X_3X_4) + Var(X_5X_6).
\end{aligned}$$

Based on Lemma 1, we can easily obtain $Var(X_1X_2)$ and $Var(X_3X_4)$. As for $Var(X_5X_6)$, X_5 and X_6 can be expressed as follows:

$$X_5 = \theta_1 U_1 + U_2,$$
$$X_6 = \theta_2 U_1 + \theta_3 U_1^2 + U_3,$$

where $U_i \sim U(0, 1), i = 1, 2, 3$ and U_i's are mutually independent for $i = 1, 2, 3$. Because

$$\begin{aligned}
Var(X_5X_6) &= E(X_5^2X_6^2) - [E(X_5X_6)]^2\\
&= Cov(X_5^2, X_6^2) + E(X_5^2)E(X_6^2) - [Cov(X_5, X_6) + E(X_5)E(X_6)]^2,
\end{aligned}$$

using the property that U_i's are mutually independent for $i = 1, 2, 3$, it is straightforward to express $Var(X_5 X_6)$ as a function of $(\theta_1, \theta_2, \theta_3)$ and the moments of $U_i, i = 1, 2, 3$.

After calculating $Var(X_1 X_2)$, $Var(X_3 X_4)$, and $Var(X_5 X_6)$, we can calculate $ST_{\{X_1 X_2\}}$, $ST_{\{X_3 X_4\}}$, and $ST_{\{X_5 X_6\}}$ as follows:

$$\begin{aligned} ST_{\{X_1 X_2\}} &= \frac{Var(X_1 X_2)}{Var(X_1 X_2) + Var(X_3 X_4) + Var(X_5 X_6)}, \\ ST_{\{X_3 X_4\}} &= \frac{Var(X_3 X_4)}{Var(X_1 X_2) + Var(X_3 X_4) + Var(X_5 X_6)}, \\ ST_{\{X_5 X_6\}} &= \frac{Var(X_5 X_6)}{Var(X_1 X_2) + Var(X_3 X_4) + Var(X_5 X_6)}. \end{aligned}$$

A.3. The correlation matrix for the loads in Example 4 in Section 4

The correlation matrix for the loads listed below is estimated using 10^6 random observations.

Table 5: A correlation matrix for the six loads estimated based on 10^6 random observations generated based on the C-vine copula in Eq. (11).

	P_1	P_2	P_3	P_4	P_5	P_6
P_1	1.000	0.172	0.173	0.171	0.176	0.173
P_2	0.172	1.000	0.032	0.031	0.033	0.032
P_3	0.173	0.032	1.000	0.032	0.034	0.031
P_4	0.171	0.031	0.032	1.000	0.033	0.031
P_5	0.176	0.033	0.034	0.033	1.000	0.032
P_6	0.173	0.032	0.031	0.031	0.032	1.000

A.4. The response surface model used for simulating the output Y in Example 4 in Section 4

The response surface model used to simulate the output Y is provided in [39] as follows:

$$\begin{aligned}
Y = {} & 2.8070 + 1.2598E_1' + 0.2147E_2' + 1.2559A_1' + 0.2133A_2' - 0.1510P_1' - 0.4238P_2' - \\
& 0.6100P_3' - 0.6100P_4' - 0.4238P_5' - 0.1510P_6' - 0.1978E_1'^2 - 0.0362E_2'^2 - 0.2016A_1'^2 - \\
& 0.0346A_2'^2 + 0.0023P_1'^2 + 0.0008P_2'^2 + 0.0036P_3'^2 + 0.0036P_4'^2 + 0.0008P_5'^2 + 0.0023P_6'^2 - \\
& 0.0042E_1'E_2' - 0.3022E_1'A_1' - 0.0110E_1'A_2' + 0.0381E_1'P_1' + 0.0871E_1'P_2' + 0.1232E_1'P_3' + \\
& 0.1232E_1'P_4' + 0.0871E_1'P_5' + 0.0346E_1'P_6' + 0.0041E_2'A_1' + 0.0110A_1'A_2' + 0.0261A_1'P_1' + \\
& 0.0831A_1'P_2' + 0.1172A_1'P_3' + 0.1172A_1'P_4' + 0.0832A_1'P_5' + 0.0296A_1'P_6',
\end{aligned} \tag{12}$$

where $E_i', i = 1, 2$, $A_i', i = 1, 2$, and $P_i', i = 1, 2, 3, 4, 5, 6$ are the standardized inputs. For example, $E_1' = \frac{E - \mu_{E_1}}{\sigma_{E_1}}$, where μ_{E_1} is the mean of E_1 and σ_{E_1} is the standard deviation of E_1.

Acknowledgements

The authors would like to thank the Editor and two anonymous reviewers for their feedback that helped significantly improve this article. This work was supported in part by the National Science Foundation (NSF) under Grant CMMI-1824681.

References

References

[1] Z. Kala, Sensitivity and reliability analyses of lateral-torsional buckling resistance of steel beams, Archives of Civil and Mechanical Engineering 15 (4) (2015) 1098–1107.

[2] D. G. Sanchez, B. Lacarrière, M. Musy, B. Bourges, Application of sensitivity analysis in building energy simulations: Combining first- and second-order elementary effects methods, Energy and Buildings 68 (2014) 741–750.

[3] J. Xu, F. Kong, A cubature collocation based sparse polynomial chaos expansion for efficient structural reliability analysis, Structural Safety 74 (2018) 24–31.

[4] G. Deman, K. Konakli, B. Sudret, J. Kerrou, P. Perrochet, H. Benabderrahmane, Using sparse polynomial chaos expansions for the global sensitivity analysis of groundwater lifetime expectancy in a multi-layered hydrogeological model, Reliability Engineering & System Safety 147 (2016) 156–169.

[5] M. Fesanghary, E. Damangir, I. Soleimani, Design optimization of shell and tube heat exchangers using global sensitivity analysis and harmony search algorithm, Applied Thermal Engineering 29 (5) (2009) 1026–1031.

[6] B. Sudret, Meta-models for structural reliability and uncertainty quantification, in: Asian-Pacific Symposium on Structural Reliability and its Applications, Singapore, 2012, pp. 1–24.

[7] L. Le Gratiet, S. Marelli, B. Sudret, Metamodel-based sensitivity analysis: polynomial chaos expansions and Gaussian processes, in: R. Ghanem, D. Higdon, H. Owhadi (Eds.), Handbook of Uncertainty Quantification, Springer International Publishing, 2017, pp. 1289–1325.

[8] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, in: P. Spanos, G. Deodatis (Eds.), Proc. 5th Int. Conf. on Comp. Stoch. Mech (CSM5), Rhodos, Greece, 2006.

[9] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Reliability Engineering & System Safety 93 (7) (2008) 964–979.

[10] J. A. Witteveen, S. Sarkar, H. Bijl, Modeling physical uncertainties in dynamic stall induced fluid–structure interaction of turbine blades using arbitrary polynomial chaos, Computers & Structures 85 (11) (2007) 866–878.

[11] S. Marelli, B. Sudret, An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis, Structural Safety 75 (2018) 67–74.

[12] G. Kewlani, J. Crawford, K. Iagnemma, A polynomial chaos approach to the analysis of vehicle dynamics under uncertainty, Vehicle System Dynamics 50 (5) (2012) 749–774.

[13] G. Chastaing, F. Gamboa, C. Prieur, Generalized Hoeffding-Sobol decomposition for dependent variables-application to sensitivity analysis, Electronic Journal of Statistics 6 (2012) 2420–2448.

[14] M. Navarro, J. Witteveen, J. Blom, Polynomial chaos expansion for general multivariate distributions with correlated variables, arXiv:1406.5483 (2014) 1–24.

[15] K. Zhang, Z. Lu, L. Cheng, F. Xu, A new framework of variance based global sensitivity analysis for models with correlated inputs, Structural Safety 55 (2015) 1–9.

[16] T. A. Mara, S. Tarantola, Variance-based sensitivity indices for models with dependent inputs, Reliability Engineering & System Safety 107 (2012) 115–121.

[17] T. A. Mara, S. Tarantola, P. Annoni, Non-parametric methods for global sensitivity analysis of model output with dependent inputs, Environmental Modelling & Software 72 (2015) 173–183.

[18] S. Tarantola, T. A. Mara, Variance-based sensitivity indices of computer models with dependent inputs: The Fourier amplitude sensitivity test, International Journal for Uncertainty Quantification 7 (6) (2017) 511–523.

[19] I. M. Sobol, Sensitivity estimates for nonlinear mathematical models, Mathematical Modelling and Computational Experiments 1 (4) (1993) 407–414.

[20] S. Kucherenko, S. Tarantola, P. Annoni, Estimation of global sensitivity indices for models with dependent variables, Computer Physics Communications 183 (4) (2012) 937–946.

[21] N. Wiener, The homogeneous chaos, American Journal of Mathematics 60 (4) (1938) 897–936.

[22] D. Xiu, G. E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM Journal on Scientific Computing 24 (2) (2002) 619–644.

[23] X. Wan, G. E. Karniadakis, Multi-element generalized polynomial chaos for arbitrary probability measures, SIAM Journal on Scientific Computing 28 (3) (2006) 901–928.

[24] S. Oladyshkin, W. Nowak, Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion, Reliability Engineering & System Safety 106 (2012) 179–190.

[25] S. Rahman, A polynomial chaos expansion in dependent random variables, Journal of Mathematical Analysis and Applications 464 (1) (2018) 749–775.

[26] M. Berveiller, B. Sudret, M. Lemaire, Stochastic finite element: a non intrusive approach by regression, European Journal of Computational Mechanics/Revue Européenne de Mécanique Numérique 15 (1-3) (2006) 81–92.

[27] R. H. Cameron, W. T. Martin, The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals, Annals of Mathematics (1947) 385–392.

[28] Å. Björck, Numerics of Gram-Schmidt orthogonalization, Linear Algebra and its Applications 197 (1994) 297–316.

[29] Å. Björck, C. C. Paige, Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm, SIAM Journal on Matrix Analysis and Applications 13 (1) (1992) 176–190.

[30] G. Blatman, B. Sudret, Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach, Comptes Rendus Mécanique 336 (6) (2008) 518–523.

[31] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, Journal of Computational Physics 230 (6) (2011) 2345–2367.

[32] J. D. Jakeman, M. S. Eldred, K. Sargsyan, Enhancing $\ell 1$-minimization estimates of polynomial chaos expansions using basis selection, Journal of Computational Physics 289 (2015) 18–34.

[33] J. Peng, J. Hampton, A. Doostan, On polynomial chaos expansion via gradient-enhanced $\ell 1$-minimization, Journal of Computational Physics 310 (2016) 440–458.

[34] C. F. J. Wu, M. S. Hamada, Experiments: Planning, Analysis, and Optimization, John Wiley & Sons, 2011.

[35] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global sensitivity analysis: the primer, John Wiley & Sons, 2008.

[36] T. J. DiCiccio, B. Efron, Bootstrap confidence intervals, Statistical Science 11 (3) (1996) 189–228.

[37] J. Jacques, C. Lavergne, N. Devictor, Sensitivity analysis in presence of model uncertainty and correlated inputs, Reliability Engineering & System Safety 91 (10) (2006) 1126–1134.

[38] E. Torre, S. Marelli, P. Embrechts, B. Sudret, Data-driven polynomial chaos expansion for machine learning regression, Journal of Computational Physics 388 (2019) 601–623.

[39] S. H. Lee, B. M. Kwak, Response surface augmented moment method for efficient reliability analysis, Structural Safety 28 (3) (2006) 261–272.

[40] E. Myshetskaya, Monte Carlo estimators for small sensitivity indices, Monte Carlo Methods and Applications 13 (5-6) (2008) 455–465.

[41] J.-Y. Tissot, C. Prieur, A randomized orthogonal array-based procedure for the estimation of first-and second-order sobol'indices, Journal of Statistical Computation and Simulation 85 (7) (2015) 1358–1381.

[42] A. B. Owen, Better estimation of small Sobol' sensitivity indices, ACM Transactions on Modeling and Computer Simulation 23 (2) (2013) 11–17.