Probability Models for Data-Driven Global Sensitivity Analysis

Zhen Hu, Sankaran Mahadevan

PII: S0951-8320(17)30791-3
DOI: https://doi.org/10.1016/j.ress.2018.12.003
Reference: RESS 6324

To appear in: *Reliability Engineering and System Safety*

Received date: 30 June 2017
Revised date: 14 November 2018
Accepted date: 14 December 2018

Please cite this article as: Zhen Hu, Sankaran Mahadevan, Probability Models for Data-Driven Global Sensitivity Analysis, *Reliability Engineering and System Safety* (2018), doi: https://doi.org/10.1016/j.ress.2018.12.003

Highlights

- A data-driven GSA framework to compute first-order and total-effect Sobol' indices of single variables, grouped variables, and correlated variables, based on an available data matrix.
- A new Gaussian mixture copula model is proposed to model the joint probability density function of random variables.
- Implementation approaches are proposed for Sobol' indices computation using different probability models.
- Investigated the advantages and disadvantages of different probability models.

Probability Models for Data-Driven Global Sensitivity Analysis

Zhen Hu[a], Sankaran Mahadevan[b,*]

[a]*Department of Industrial and Manufacturing Systems Engineering, University of Michigan-Dearborn, Dearborn, Michigan, 48128*

[b]*Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, Tennessee, 37235*

Abstract

This paper presents a probability model-based global sensitivity analysis (PM-GSA)framework, to compute various Sobol' indices when only input-output data are available. The PM-GSA framework consists of two main elements, namely data extraction and probability model training. The data extraction step extracts data of the variables of interest (VoI) and quantity of interest (QoI) from an input-output data matrix. Following that, a probability model is built to approximate the joint probability density function between the VoI and QoI. The learned probability model is then used to compute various Sobol' indices. The implementation of the PM-GSA framework is investigated through three probability models including Gaussian copula model, Gaussian mixture model, and a new Gaussian mixture copula model. The number of dimensions of the probability model in the PM-GSA framework, is independent of the number of input variables and is always $N+1$ (e.g. 2 for the first-order index), where N is the order of the Sobol' index. In addition, the PM-GSA framework is applicable to global sensitivity analysis with not only independent input variables, but also with dependent input variables and for sets of variables. Four numerical examples are used to demonstrate the effectiveness of the proposed method and analyze the advantages and disadvantages of the different probability models.

Keywords: Global Sensitivity Analysis, Gaussian Mixture, Copula, High-dimensional, Data

*Corresponding author: Box 1831, Station B, Vanderbilt University, Nashville, Tennessee, 37235, USA. Tel.: +1-615-322-3040, E-mail: sankaran.mahadevan@vanderbilt.edu.

1. Introduction

Global sensitivity analysis (GSA), which quantifies the contributions of input random variables to the variability of an output quantity of interest (QoI) [1, 2, 3], has been widely used to rank the importance of input random variables and thus achieve the purpose of dimension reduction [4], uncertainty reduction [5], and resource allocation [6]. During the past decades, various approaches have been proposed to perform GSA, such as the Fourier amplitude sensitivity test (FAST) methods [7, 8], methods based on correlation ratio [9, 10], Kullback-Leibler divergence based approaches [11, 12], and Sobol' indices related methods [13, 14, 15]. Among these GSA methods, variance decomposition-based Sobol' indices, which is also the focus of this work, is one of the most widely used. Two types of Sobol' indices are usually computed, namely first-order Sobol' indices and total-effect Sobol' indices [6]. The first-order Sobol' index measures the individual contribution of each variable to the variability of the QoI without considering its interactions with other variables, while the interactions with other variables are included in the total-effect Sobol' indices [16, 17] and other joint effects indices of multiple variables. A straightforward way of computing Sobol's indices is to implement a double-loop Monte Carlo simulation (MCS). This double-loop procedure, however, requires a large number of evaluations of the prediction model and is unaffordable if the prediction model is expensive.

To overcome the computational effort challenge in GSA, various approaches have been proposed in recent years [18], which can be roughly grouped in three directions [19, 20]. Note that the classification of GSA methods here is not exhaustive. The *first* direction is to reduce the required number of samples in computing Sobol' indices using MCS by deriving efficient numerical algorithms [4, 21, 22, 19, 23]. For instance, Sobol' discussed how to efficiently estimate the Sobol' indices using MCS [24]. In this scheme, the required number of samples in the double-loop procedure is reduced to a number which is proportional to the dimension of the input variables. The accuracy of Sobol's

scheme was further improved by Homma and Saltelli in Ref. [4]. Similarly, Glen and Isaacs developed an approach to compute the Sobol' indices by switching the columns of two separately generated MCS sample matrices [25]. Instead of directly using the Monte Carlo samples, the *second* direction uses spectral approaches [26, 7, 8] or design of experiments methods [27, 28, 29, 30] to reduce the required number of samples in GSA. For example, the aforementioned FAST method [7, 8] is a spectral approach to perform GSA. Tissot and Prieur [29] developed a randomized orthogonal array-based procedure for the estimation of Sobol' indices. A detailed review of various sampling techniques and MCS approaches is available in Ref. [20]. The *third* direction is to replace the original (expensive) prediction model with a cheap algebraic surrogate model [31, 32], such as a regression model, Polynomial Chaos Expansion (PCE) model [33], or a Kriging model [34, 35]. Based on the surrogate model, the Sobol' indices are calculated using either analytical approaches or direct MCS-based methods. For PCE, Sudret [13] derived analytical expressions for the Sobol' indices by post-processing the PCE coefficients. For Kriging, Chen et al. [36] suggested analytical ways to compute Sobol' indices when the input variables follow normal or uniform distributions. Gratiet et al. [37] later developed a more generalized approach to estimate the Sobol' indices through a surrogate model by considering both estimation and surrogate model errors. They also extended the proposed approach to mltifidelity computer simulations [37].

In practical engineering applications, it is quite often that we may only have a group of numerical samples of input-output pairs, nothing more. The distributions, correlations, and interactions between different variables need to be learned purely based on the available numerical data. In that situation, direct MCS-based or sampling-based GSA methods cannot be adopted to rank the importance of variables for a given QoI due to the fact that the prediction model is not available and current MCS or sampling-based GSA approaches require at least two separated input-output data matrices [25]. Surrogate model-based GSA approaches [37, 13, 31] are still applicable in

this situation. However, this is not always the case. In some situations, it is observed that the surrogate model-based GSA approaches may become inapplicable or are not suggested for the following reasons [38, 39]:

- (1) Current surrogate modeling techniques are only applicable to moderate-sized problems (i.e., dimension of input variables less than 30). They cannot be applied to problems with high dimensions due to the curse of dimensionality [40].
- (2) Even if surrogate models (e.g. Kriging) sometimes can provide analytical expressions of Sobol' indices, as been pointed out in Ref. [39], most of these analytical expressions (except PCE-based expressions) involve multidimensional integrals that are tractable only when the conditional probability densities of the input random variables are known. This is apparently not the case when only a matrix of input-output data is given.
- (3) When algebraic surrogate model-based GSA approaches are applied to GSA for given data, an algebraic model is built first. The probability distributions of the input variables are then learned from the data. Based on the learned probability distributions, Sobol' indices are computed through the constructed algebraic surrogate model. This introduces several extra steps in computing Sobol' indices.

Motivated by answering the question of *how to effectively perform GSA for given data*, various approaches have been proposed in recent years. For example, Li and Mahadevan [38] presented a modularized method to estimate the first-order Sobol' indices based on stratification of available samples; Jia and Taflanidis [41] developed an auxiliary probability density function approach to estimate the first-order Sobol' indices based on data; and Sparkman et al. [42] proposed an importance sampling approach to compute Sobol' indices from available data by introducing weights to different data points. In addition to these efforts, one of the dominant directions is to employ

smoothing-based methods [43, 44], such as locally weighted regressions, local polynomial smoothing [18, 39], and recursive portioning regression. The smoothing-based methods [43, 44, 18, 39] may require the smallest sample size when they are applied to compute the first-order Sobol' indices.

The above reviewed approaches for GSA with given samples [38, 42, 18, 39], are all limited to estimating the first-order Sobol' indices of individual variables (i.e. single variables), and have difficulty in dealing with GSA of sets of input variables and in computing the total-effect indices. This paper aims to overcome these drawbacks [38, 42] by developing a generalized GSA framework, which is able to perform various GSA computations (e.g. first-order, higher-order, grouped, and total-effects Sobol' indices) purely based on data. In the proposed method, a probability model is built first based on the available data to capture the joint probability distribution of the system inputs and outputs. Based on the learned probability model, approaches are developed to effectively compute different types of Sobol' indices. Three approaches, namely Gaussian copula model, Gaussian mixture model, and a new Gaussian mixture copula model, are explored in this paper to build the probability model for use in GSA.

The contributions of this paper can be summarized as: (1) A generalized probability model-based GSA (PM-GSA) framework is developed to efficiently compute different types of Sobol' indices (e.g. first-order, higher-order, grouped, and total-effect Sobol' indices); this overcomes the limitations of current approaches for GSA with given samples [38, 42, 43, 18, 39], which can only compute the first-order Sobol' indices of single input variables. (2) A new Gaussian mixture copula model is proposed and studied to build the probability model for GSA, in addition to Gaussian copula and Gaussian mixture models; The proposed Gaussian mixture copula model can benefit not only GSA, but also many other uncertainty quantification problems. (3) Implementation approaches are developed for GSA based on different types of probability models; and (4) Advantages and disadvantages of different probability models for GSA are investigated using numerical exam-

ples.

The remainder of this paper is organized as follows. Sec. 2 reviews background concepts of variance-based GSA and data-driven GSA. Sec. 3 develops the proposed method. Sec. 4 considers three numerical examples to illustrate the proposed method, and Sec. 5 gives concluding remarks.

2. Background

2.1. Variance-based Global Sensitivity Analysis

Defining $Y \in \mathbb{R}$ as a QoI and its underlying physics model or computer code given by $Y = g(\mathbf{X})$, where $\mathbf{X} = (X_1, X_2, \cdots, X_n) \in \mathbb{R}^n$ is a vector of random input variables, the variance $\mathrm{Var}(Y)$ of Y can be decomposed as follows [24, 13]:

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} V_i + \sum_{1 \le i < j}^{n} V_{ij} + \cdots + V_{12\ldots n}, \tag{1}$$

where $V_i = \mathrm{Var}_{X_i}(\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|X_i))$ is the variance of Y caused by X_i without considering its interactions with other input variables (i.e. $\mathbf{X}_{\sim i}$), $\mathbb{E}(\cdot)$ is the expectation operator, and $V_{1\ldots k}, \forall k = 2, \cdots, n$ represents the proportion of $\mathrm{Var}(Y)$ caused by variables $(X_1, \cdots, X_k)$.

Based on the above variance decomposition, the Sobol' indices are defined as [24, 13]

$$S_i = \frac{V_i}{Var(Y)}, S_{ij} = \frac{V_{ij}}{Var(Y)}, S_{1\ldots k} = \frac{V_{1\ldots k}}{Var(Y)}, \forall k = 2, \cdots, n, \tag{2}$$

where S_i is the first-order index, S_{ij} is the second-order index, and $S_{1\ldots k}$ is the higher-order index corresponding to input variables $(X_1, X_2, \cdots, X_k)$.

The number of indices will grow dramatically if the higher-order indices are used. For this reason, the first-order and total-effect Sobol' indices are commonly used and are given by

$$S_i = \frac{\mathrm{Var}_{X_i}(\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|X_i))}{\mathrm{Var}(Y)}, \forall i = i, \cdots, n, \tag{3}$$

$$S_{T_i} = 1 - \frac{\mathrm{Var}_{\mathbf{X}_{\sim i}}(\mathbb{E}_{X_i}(Y|\mathbf{X}_{\sim i}))}{\mathrm{Var}(Y)}, \forall i = i, \cdots, n, \tag{4}$$

or

$$S_{T_i} = \frac{\mathbb{E}_{\mathbf{X}_{\sim i}}(\mathrm{Var}_{X_i}(Y|\mathbf{X}_{\sim i}))}{\mathrm{Var}(Y)}, \forall i = i, \cdots, n, \tag{5}$$

where S_i and S_{T_i} are the first-order and total-effect Sobol' indices of X_i, respectively.

It should be noted that the above variance decomposition is derived based on the independence assumption of the input variables. When the input variables are correlated, $Var(Y)$ cannot be decomposed as in Eq. (1). However, as has been pointed out in Refs. [45, 46], S_i and S_{T_i} computed using the above formulas are still informative for the importance measure of dependent input variables. In addition, Mara and Tarantola [2] defined two types of sensitivity indices, namely *full* sensitivity index and *independent* sensitivity index, to perform GSA of model output with dependent random variables. The full sensitivity index includes the effects of the dependence of a VoI with other inputs while the independent sensitivity indices represent the effects of a VoI that are not due to its dependence with other variables [47]. According to the definitions given in Ref. [47], the indices given in Eqs. (3) and (4) are respectively the *full* first-order sensitivity index and the *independent* total-effect index when they are applied to GSA of model output with dependent random variables. In this paper, we therefore focus on how to compute Eqs. (3) or (4) for generalized problems with or without dependent input variables. Note that the proposed methods are not limited to Eqs. (3) or (4). They can be extended to other types of sensitivity indices if other definitions are used rather than Eqs. (3) or (4).

As discussed in Sec. 1, directly solving Eqs. (3) or (4) requires a double-loop MCS, which is computationally expensive. In some situations, we may only have a data matrix given as follows

$$\begin{pmatrix} \mathbf{z}^{(1)} \\ \mathbf{z}^{(2)} \\ \vdots \\ \mathbf{z}^{(s)} \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} & \cdots & y^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} & \cdots & y^{(2)} \\ \vdots & \vdots & \vdots & \cdots & \ddots & \vdots \\ x_1^{(s)} & x_2^{(s)} & \cdots & x_n^{(s)} & \cdots & y^{(s)} \end{pmatrix}, \tag{6}$$

in which $\mathbf{z}^{(q)} = [\mathbf{x}^{(q)}, y^{(q)}], \forall q = 1, \cdots, s$, is the q-th sample of $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ and Y and s is the number of samples in the data matrix. The data presented in the matrix may be polluted by the noise in the data collection process, and there are unknown interactions between different variables in the matrix. The question that needs to be answered is *how to effectively perform GSA based on the data matrix given in Eq. (6).*

As mentioned previously, GSA approach for given samples is a promising approach to answer the above equation. Next, we will briefly review several recently developed approaches for GSA with given samples [38, 42, 43, 18, 39] and analyze the advantages and drawbacks of these methods. Based on that, we present the proposed new method.

2.2. GSA with Given Samples

To answer the above question of GSA with given samples, several approaches have been proposed recently, such as smoothing-based methods [43, 44, 18, 39], the auxiliary probability density function approach [41], the importance sampling-based method [42], and the modularized GSA method [38]. In smoothing-based methods [43, 44, 18, 39], regression models are first built for the conditional moments using various local smoothing approaches. Based on the conditional moments regression models, the first-order Sobol' indices are estimated efficiently. In the auxiliary probability density function approach [41], an auxiliary probability density function is defined first. From the auxiliary probability density function, marginal distributions of interest are approximated using Kernel Density Estimation. Based on that, the first-order Sobol' indices are computed. In the importance sampling-based method [42], the moments of conditional distributions are computed by assigning weights to different samples using the importance sampling concept and kernel density function. In the modularized GSA method (referred to as MGSA in Ref. [38]), the data of Y is first divided into M segments and the first-order Sobol' indices (S_i) are then computed based on mean and variance of Y in each segment. Two algorithms are proposed in Ref. [38] for MGSA.

The fundamental idea of the MGSA method is similar to the smoothing-based method since both of them focus on the approximation of conditional moments. Considering that the MGSA method has shown better accuracy than the local smoothing approach [38] (see the results of Example one in the numerical example section) and will be compared with the proposed method in this paper, we first analyze the fundamental principle of the MGSA method and then discuss the limitations of current methods for GSA with given samples.

The fundamental principle of the MGSA method can be explained as discretized numerical integration. The first-order indices of $X_i, \forall i = 1, \cdots, n$, given in Eq. (3) can be written as

$$S_i = \frac{\mathrm{Var}_{X_i}(\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|X_i))}{\mathrm{Var}(Y)} = \frac{1}{\mathrm{Var}(Y)} \times \left\{ \int_{\Omega_X} f_{X_i}(x_i) \mathbb{E}^2_{\mathbf{X}_{\sim i}}(Y|x_i) dx_i - \left(\int_{\Omega_X} f_{X_i}(x_i) \mathbb{E}_{\mathbf{X}_{\sim i}}(Y|x_i) dx_i \right)^2 \right\}, \quad (7)$$

in which Ω_X is the domain of X_i, $f_{X_i}(x_i)$ is the probability density function (PDF) of X_i, and

$$\int_{\Omega_X} f_{X_i}(x_i) \mathbb{E}^2_{\mathbf{X}_{\sim i}}(Y|x_i) dx_i = \int_{\Omega_X} \mathbb{E}^2_{\mathbf{X}_{\sim i}}(Y|x_i) dF_{X_i}(x_i) = \int_0^1 \mathbb{E}^2_{\mathbf{X}_{\sim i}}(Y|F^{-1}_{X_i}(u_i)) du_i, \quad (8)$$

$$\int_{\Omega_X} f_{X_i}(x_i) \mathbb{E}_{\mathbf{X}_{\sim i}}(Y|x_i) dx_i = \int_0^1 \mathbb{E}_{\mathbf{X}_{\sim i}}(Y|F^{-1}_{X_i}(u_i)) du_i, \quad (9)$$

where $u_i = F_{X_i}(x_i)$ and $x_i = F^{-1}_{X_i}(u_i)$ are respectively the cumulative density function (CDF) and inverse CDF of X_i.

Based on the above transformation, the MCS method is then used to estimate Eq.(9) by discretizing the interval $[0, 1]$ into M segments as follows

$$\int_{\Omega_X} f_{X_i}(x_i) \mathbb{E}_{\mathbf{X}_{\sim i}}(Y|x_i) dx_i \approx \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}_{\mathbf{X}_{\sim i}}(Y|F^{-1}_{X_i}(u_i^{(j)})), \quad (10)$$

where $u_i^{(j)}$ is the j-th sample of the CDF of X_i and $\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|F^{-1}_{X_i}(u_i^{(j)}))$ is mean of Y for given $u_i^{(j)}$. Similarly, Eq. (8) can also be approximated using the MCS method. After that, we have Eq. (7)

as

$$S_i = \frac{\mathrm{Var}_{X_i}(\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|X_i))}{\mathrm{Var}(Y)} \approx \frac{1}{\mathrm{Var}(Y)M^2}\sum_{j=1}^{M}\sum_{k>j}^{M}(\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|F_{X_i}^{-1}(u_i^{(j)})) - \mathbb{E}_{\mathbf{X}_{\sim i}}(Y|F_{X_i}^{-1}(u_i^{(k)})))^2. \quad (11)$$

Eq. (11) implies that $\mathrm{Var}_{X_i}(\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|X_i))$ can be computed by estimating the variances of $\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|F_{X_i}^{-1}(u_i^{(j)})), \forall j = 1, 2, \cdots M$. Since it is difficult to directly compute $\mathbb{E}_{\mathbf{X}_{\sim i}}(Y|F_{X_i}^{-1}(u_i^{(j)}))$, Li and Mahadevan [38] proposed to use the samples of Y corresponding to X_i in the interval $[F_{X_i}^{-1}(u_i^{(j-1)}), F_{X_i}^{-1}(u_i^{(j+1)})]$ to compute the conditional expectation. Based on this, the first-order Sobol' indices are estimated.

Alternatively, the first-order Sobol' indices can be computed by

$$S_i = 1 - \frac{\mathbb{E}_{X_i}(\mathrm{Var}_{\mathbf{X}_{\sim i}}(Y|X_i))}{\mathrm{Var}(Y)}. \quad (12)$$

Based on Eq. (12), a second algorithm is also proposed in Ref.[38] to estimate S_i as

$$S_i \approx 1 - \frac{1}{M\mathrm{Var}(Y)}\sum_{j=1}^{M}\mathrm{Var}_{\mathbf{X}_{\sim i}}(Y|F_{X_i}^{-1}(u_i^{(j)})), \quad (13)$$

where $\mathrm{Var}_{\mathbf{X}_{\sim i}}(Y|F_{X_i}^{-1}(u_i^{(j)}))$ is computed using the samples of Y corresponding to X_i in the segment $[F_{X_i}^{-1}(u_i^{(j-1)}), F_{X_i}^{-1}(u_i^{(j+1)})]$ based on the data matrix given in Eq. (6). More details of the MGSA method are available in Ref.[38].

From the above presented fundamental principle of MGSA method [38], it can be found that the method is capable of estimating the first-order Sobol' indices purely based on given samples. It does not make assumptions about the function form or the distributions of variables. It can also be used to compute S_i for problems with correlated input variables. However, the method also has several disadvantages. For instance, the result of GSA is affected by M (i.e. the number of segments that the CDF is divided into); the method cannot be used to compute the total-effects Sobol' indices; and it is not capable of evaluating the first-order and total-effect Sobol' indices of a set of input variables (i.e. more than one input variables). The other methods for GSA with given

samples [41, 44, 39, 42] share the same disadvantages. In the next section, we will discuss how to overcome the aforementioned disadvantages by using different types of probability models.

3. Proposed Method

Before discussing the details of the proposed method, we define two sets of random input variables: $\mathbf{X}_c \in \mathbb{R}^{n_c}$ and $\mathbf{X}_r \in \mathbb{R}^{n-n_c}$, where $\mathbf{X} = (\mathbf{X}_c, \mathbf{X}_r)$, $\mathbf{X}_c$ represents the random input variables of interest (VoI), n_c is the number of variables in $\mathbf{X}_c$, for which the Sobol' indices need to be computed, and $\mathbf{X}_r$ represents the remaining random input variables including remaining known and unknown random input variables, where the unknown random variables could be any unknown noise sources in the data.

Based on this definition and Eqs. (3) and (4), it can be concluded that most of the GSA computations such as first-order, higher-order, or total-effect Sobol' indices, GSA of set of variables, and GSA of dependent random variables can be connected to the estimation of $\mathrm{Var}(Y)$, $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$, or $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ (i.e. Eq. (12)) with given samples. For given samples (see Eq. (6)), $\mathrm{Var}(Y)$ can be directly estimated based on $[y^{(1)}, \cdots, y^{(s)}]^T$. The main challenge is how to effectively estimate the other two elements.

In order to estimate $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ or $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using the data given in Eq. (6), in this paper, a probability model-based method is explored. Assuming that we have a probability model $f_{\mathbf{X}Y}(\mathbf{x}_c, \mathbf{x}_r, y)$ which can accurately represent the joint probability density function (PDF) between $\mathbf{X}_c = \mathbf{x}_c$, $\mathbf{X}_r = \mathbf{x}_r$, and $Y = y$. For given $\mathbf{X}_c = \mathbf{x}_c$, we can compute $\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c)$ as

$$\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) = \int\int y \frac{f_{\mathbf{X}Y}(\mathbf{x}_c, \mathbf{x}_r, y)}{f_{\mathbf{X_c}}(\mathbf{x}_c)} \mathbf{d}\mathbf{x}_r dy = \int_{\Omega_Y} y \frac{f_{\mathbf{X_c}Y}(\mathbf{x}_c, y)}{f_{\mathbf{X_c}}(\mathbf{x}_c)} dy, \tag{14}$$

where Ω_Y is the domain of Y and $f_{\mathbf{X_c}Y}(\mathbf{x}_c, y)$ is the joint PDF of $\mathbf{X}_c$ and Y.

The above equation implies that we can accurately compute $\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c)$ if the joint PDF

$f_{\mathbf{X}Y}(\mathbf{x}_c, y)$ is accurately represented based on the data given in Eq. (6), so is $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$. This is also applicable to the evaluation of $\mathbb{E}_{\mathbf{X}_c}(Var_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$. Based on this observation, we develop a new data-driven GSA framework. As shown in Fig.1, there are mainly three steps. Here we briefly introduce the task of each step and the implementation procedures of each step are discussed in detail in subsequent sections.

- Step 1: Extract data of $\mathbf{X}_c$ and Y from the data matrix given in Eq.(6). The implementation of this step is very straightforward and is general to any probability models. It should be noted that the data should be extracted differently depending on what type of Sobol' indices need to be computed.

- Step 2: Build probability models to approximate $f_{\mathbf{X}Y}(\mathbf{x}_c, y)$ based on the data extracted from the Step 1. In this paper, three types of probability models, including Gaussian copula, Gaussian mixture, and a new Gaussian mixture copula, are explored to build such a model.

- Step 3: Compute the Sobol' indices using the probability model learned from Step 2. Depending on the probability model type, the way of computing Sobol' indices may be quite different. In this paper, we investigate different ways of computing Sobol' indices based on the studied three probability models.

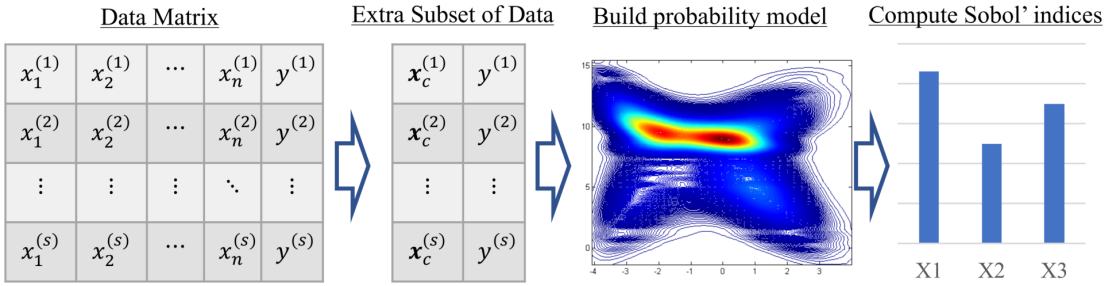


Figure 1: Flowchart of data-driven GSA

As discussed in Sec. 2, GSA can be performed through Eqs. (3) and (4) or Eqs. (12) and (5). In

the subsequent sections, we call GSA based on $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ (i.e. Eqs. (3) and (4)) as *approach one* and GSA based on $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ (i.e. Eqs. (12) and (5)) as *approach two*. In what follows, we provide details of the implementation procedure of these two approaches using Gaussian copula model, Gaussian mixture model, and a new Gaussian mixture copula model.

3.1. Gaussian Copula GSA (GC-GSA)

A copula function describes the dependence between random variables by connecting the marginal CDFs to the joint cumulative distribution function [48, 49, 50]. For a subset of random input variables $\mathbf{X}_c$ and Y, the joint CDF $F(\mathbf{x}_c, y)$ is connected to the marginal CDFs, $F_c(\mathbf{x}_c)$ and $F_Y(y)$, through the copula function $C(\cdot, \cdot)$ as follows

$$F(\mathbf{x}_c, y) = \mathbb{P}\{\mathbf{X}_c \leq \mathbf{x}_c, Y \leq y\} = C(\mathbf{F}_c(\mathbf{x}_c), F_Y(y); \boldsymbol{\theta}) = C(\mathbf{u}_c, u_Y; \boldsymbol{\theta}), \tag{15}$$

where $\mathbb{P}\{\cdot\}$ is the probability operator, $\mathbf{F}_c(\mathbf{x}_c) = [F_{X_1}(x_1), \cdots F_{X_{n_c}}(x_{n_c})] = \mathbf{u}_c = [u_1, \cdots, u_{n_c}]$ is a vector of marginal CDF values of $\mathbf{X}_c$, $F_Y(y) = u_Y$ is the CDF of the QoI, n_c is the number of variables in $\mathbf{X}_c$, and $\boldsymbol{\theta}$ is a vector of parameters of the copula function.

The joint PDF of $\mathbf{X}_c = \mathbf{x}_c$ and $Y = y$ is computed by

$$\begin{aligned} f_{\mathbf{X}_c Y}(\mathbf{x}_c, y) &= \frac{\partial^{n_c+1} C(\mathbf{F}_c(\mathbf{x}_c), F_Y(y); \boldsymbol{\theta})}{\partial F_{X_1}(x_1) \cdots \partial F_{X_{nc}}(x_{nc}) \partial F_Y(y)} \frac{\partial F_{X_1}(x_1)}{\partial x_1} \cdots \frac{\partial F_{X_n}(x_n)}{\partial x_n} \frac{\partial F_Y(Y)}{\partial Y}, \\ &= \frac{\partial^{n_c+1} C(\mathbf{u}_c, u_Y; \boldsymbol{\theta})}{\partial u_1 \cdots \partial u_{nc} \partial u_Y} f_{X_1}(x_1) \cdots f_{X_n}(x_n) f_Y(y), \end{aligned} \tag{16}$$

and

$$f_{\mathbf{X}_c Y}(\mathbf{x}_c, y) = f_{X_1}(x_1) \cdots f_{X_n}(x_n) f_Y(y) c(\mathbf{u}_c, u_Y; \boldsymbol{\theta}), \tag{17}$$

where $c(\cdot)$ is the copula PDF function, $f_Y(y)$ is the PDF of the QoI, Y, and $f_{X_i}(x_i)$ is the PDF of $X_i, \forall i = 1, 2, \cdots, n_c$.

The copula functions are usually defined for bivariate problems. Only a few copula functions, such as Gaussian copula and student's t copula, are well-studied for the multi-variate high-

dimensional case [51]. Here, the Gaussian copula is used as an example to illustrate the application of a copula function to GSA with given samples. For Gaussian copula, we have $\boldsymbol{\theta} = \boldsymbol{\rho}$ and the Gaussian copula function is given by

$$F(\mathbf{x}_c, y) = C(\mathbf{u}_c, u_Y; \boldsymbol{\theta}) = \Phi(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_Y); \boldsymbol{\rho}), \tag{18}$$

in which $\Phi^{-1}(\cdot)$ is the inverse CDF function of a standard normal variable and $\boldsymbol{\rho}$ is the correlation matrix between variables, $(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_Y))$. Since the linear correlation matrix $\boldsymbol{\rho}$ given in Eq. (18) cannot be estimated directly from data, it is usually estimated based on the Kendall's tau rank correlation matrix based on the transformation of the Kendall's tau [51, 52]. More detailed discussion of various copula functions is available in Ref. [51].

The PDF function given in Eq. (17) for the Gaussian copula case is given by [53]

$$\begin{aligned} f(\mathbf{x}_c, y) = f_{X_1}(x_1) \cdots f_{X_{n_c}}(x_{n_c}) f_Y(y) \frac{\partial \Phi^{-1}(u_1)}{\partial u_1} \cdots \frac{\partial \Phi^{-1}(u_{n_c})}{\partial u_{n_c}} \\ \times \frac{\partial \Phi^{-1}(u_Y)}{\partial u_Y} \phi(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_Y); \boldsymbol{\rho}), \end{aligned} \tag{19}$$

where $\phi(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_Y); \boldsymbol{\rho})$ is the PDF of multivariate Gaussian distribution.

The above equation indicates that estimating $\boldsymbol{\rho}$ (i.e. Kendall's tau since there is a relationship between Kendall's tau and $\boldsymbol{\rho}$) is the most important part for the modeling of a Gaussian copula. This correlation matrix can be solved using either optimization using the maximum likelihood estimate or empirical estimation from data based on the estimation of Kendall's tau. Here, the command of *copulafit* in the MATLAB Statistics Toolbox (MATLAB R2017a) [54] is employed. Using the data matrix given in Eq. (6), *copulafit* returns following empirical correlation matrix

$$\boldsymbol{\rho} = \begin{pmatrix} \boldsymbol{\rho}_{cc} & \boldsymbol{\rho}_{cy} \\ \boldsymbol{\rho}_{yc} & 1 \end{pmatrix}, \tag{20}$$

in which $\boldsymbol{\rho}_{cc}$, $\boldsymbol{\rho}_{yc}$, and $\boldsymbol{\rho}_{cy}$ are the correlation matrices between $(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_{n_c}))$ and $(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_{n_c}))$, Y and $(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_{n_c}))$, and $(\Phi^{-1}(u_1), \cdots, \Phi^{-1}(u_{n_c}))$ and Y, re-

spectively. Based on $\boldsymbol{\rho}$ and the Gaussian copula, we then compute the Sobol' indices using *approach one* (i.e. Eq. (3)) and *approach two* (i.e. Eq. (12)).

3.1.1. GC-GSA approach one

The task of approach one is to compute $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using the Gaussian copula. For given $\mathbf{X}_c = \mathbf{x}_c$, we have $\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c)$ as

$$\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) = \int_{\Omega_Y} y f_{Y|\mathbf{X}_c}(y|\mathbf{x}_c)dy = \int_{\Omega_Y} y f_Y(y)\frac{c(\mathbf{u}_c, u_Y; \boldsymbol{\rho})}{c(\mathbf{u}_c; \boldsymbol{\rho}_{cc})}dy, \tag{21}$$

where $c(\mathbf{u}_c; \boldsymbol{\rho}_{cc})$ is the Gaussian copula of $\mathbf{X}_c$.

Eq. (21) can be further written as

$$\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) = \frac{1}{c(\mathbf{u}_c; \boldsymbol{\rho}_{cc})}\int_0^1 F_Y^{-1}(u_Y)c(\mathbf{u}_c, u_Y; \boldsymbol{\rho})du_Y. \tag{22}$$

For some special cases, such as when the marginal distribution of Y is Gaussian, an analytical solution may be derived for Eq. (22). For the purpose of generalization, here, an MCS-based method is adopted to estimate Eq. (22) as below

$$\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) \approx \frac{1}{c(\mathbf{u}_c; \boldsymbol{\rho}_{cc})N_{MCS}}\sum_{k=1}^{N_{MCS}} F_Y^{-1}(u_Y^{(k)})c(\mathbf{u}_c, u_Y^{(k)}; \boldsymbol{\rho}), \tag{23}$$

where $u_Y^{(k)}$ is the k-th sample of U_Y, N_{MCS} is the number of MCS samples used for integration, and $F_Y^{-1}(\cdot)$ is the inverse CDF of Y computed using kernel density function [55]. Since the integration will not evaluate the original computer simulation model or physical model, it is computationally cheap to solve Eq. (23) using MCS and a samples size of $N_{MCS} \geq 1 \times 10^4$ is used to reduce the integration error.

Based on the learned Gaussian copula model (i.e. Eq. (20)), we also generate N_{MCS} samples of $\mathbf{u}_c^{(i)}, i = 1, 2, \cdots, N_{MCS}$. Using Eq. (23) and the data of $\mathbf{u}_c$, we then estimate $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c =$

$\mathbf{x}_c$)) as

$$\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c=\mathbf{x}_c)) \approx \frac{1}{2N_{MCS}^3}\sum_{i=1}^{N_{MCS}}\sum_{j=1}^{N_{MCS}}\Big(\frac{1}{c(\mathbf{u}_c^{(i)};\boldsymbol{\rho}_{cc})}\sum_{k=1}^{N_{MCS}}F_Y^{-1}(u_Y^{(k)})c(\mathbf{u}_c^{(i)},u_Y^{(k)};\boldsymbol{\rho}) - \frac{1}{c(\mathbf{u}_c^{(j)};\boldsymbol{\rho}_{cc})}\sum_{k=1}^{N_{MCS}}F_Y^{-1}(u_Y^{(k)})c(\mathbf{u}_c^{(j)},u_Y^{(k)};\boldsymbol{\rho})\Big)^2. \quad (24)$$

Once we can compute $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$, various Sobol' indices such as first-order, second-order, and total-effect indices, and indices of a set of variables can be computed using Eqs. (3) and (4).

3.1.2. GC-GSA approach two

In approach two, we focus on computing $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using the Gaussian copula. Similar to Eq. (22), we have

$$\mathbb{E}_{\mathbf{X}_r}(Y^2|\mathbf{X}_c=\mathbf{x}_c) = \frac{1}{c(\mathbf{u}_c;\boldsymbol{\rho}_{cc})}\int_0^1 (F_Y^{-1}(u_Y))^2 c(\mathbf{u}_c,u_Y;\boldsymbol{\rho})du_Y. \quad (25)$$

Eq. (25) is estimated as

$$\mathbb{E}_{\mathbf{X}_r}(Y^2|\mathbf{X}_c=\mathbf{x}_c) \approx \frac{1}{c(\mathbf{u}_c;\boldsymbol{\rho}_{cc})N_{MCS}}\sum_{k=1}^{N_{MCS}}(F_Y^{-1}(u_Y^{(k)}))^2 c(\mathbf{u}_c,u_Y^{(k)};\boldsymbol{\rho}). \quad (26)$$

Combining Eqs. (23) and (26) yields

$$\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c=\mathbf{x}_c) \approx \frac{1}{c(\mathbf{u}_c;\boldsymbol{\rho}_{cc})N_{MCS}}\sum_{k=1}^{N_{MCS}}(F_Y^{-1}(u_Y^{(k)}))^2 c(\mathbf{u}_c,u_Y^{(k)};\boldsymbol{\rho}) - \Big(\frac{1}{c(\mathbf{u}_c;\boldsymbol{\rho}_{cc})N_{MCS}}\sum_{k=1}^{N_{MCS}}F_Y^{-1}(u_Y^{(k)})c(\mathbf{u}_c,u_Y^{(k)};\boldsymbol{\rho})\Big)^2. \quad (27)$$

Same to the GC-GSA approach one, $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ is estimated using the random samples generated using the learned Gaussian coupla model as

$$\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c=\mathbf{x}_c)) \approx \frac{1}{N_{MCS}}\sum_{i=1}^{N_{MCS}}\Big[\frac{1}{c(\mathbf{u}_c^{(i)};\boldsymbol{\rho}_{cc})N_{MCS}}\sum_{k=1}^{N_{MCS}}(F_Y^{-1}(u_Y^{(k)}))^2 c(\mathbf{u}_c^{(i)},u_Y^{(k)};\boldsymbol{\rho}) - \Big(\frac{1}{c(\mathbf{u}_c^{(i)};\boldsymbol{\rho}_{cc})N_{MCS}}\sum_{k=1}^{N_{MCS}}F_Y^{-1}(u_Y^{(k)})c(\mathbf{u}_c^{(i)},u_Y^{(k)};\boldsymbol{\rho})\Big)^2\Big]. \quad (28)$$

Once we have $\mathbb{E}_{\mathbf{X}_c}(\text{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$, the Sobol' indices can be computed using Eqs. (12) and (5). The above discussed method is based on the Gaussian copula assumption. It can be easily extended to other copula functions when the multi-variate copula function is available. For instance, the Vine copula [48] can be employed to overcome the limitations of current bivariate copula functions to represent complicated joint PDFs of multivariate random variables. Extension of the presented method to Vine copula, however, is not the focus of this paper. Next, we summarize the implementation procedure of Gaussian copula-based GSA approaches.

In the Gaussian copula-based GSA method, there are mainly five steps explained as below:

- Step 1: Compute $\text{Var}(Y)$ using data of Y given in Eq. (6) and convert the data matrix given in Eq. (6) into the data of $\mathbf{u}_x$ and u_Y using kernel density function [55].
- Step 2: Extract data of $\mathbf{u}_c$ and u_Y from the data matrix obtained from Step 1. For example, if the first-order Sobol' index of X_i needs to be computed, extract data of $\mathbf{u}_{X_i}$ and u_Y.
- Step 3: Fit Gaussian copulas $c(\mathbf{u}_c, u_Y)$ and $c(\mathbf{u}_c)$ using the data of $\mathbf{u}_c$ and u_Y.
- Step 4: Compute $\text{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using Eq. (24) or $\mathbb{E}_{\mathbf{X}_c}(\text{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using Eq. (28).
- Step 5: Compute Sobol' indices (e.g. first-order indices using Eq. (3) or (12) and total-effect indices using Eq. (4)) or (5).

Since the Gaussian copula function may not be able to accurately approximate $f_{\mathbf{X}_c Y}(\mathbf{x}_c, y)$, in the following section, we investigate GSA based on Gaussian mixture model (GMM), which is more flexible than the Gaussian copula in modeling the joint PDF based on data.

3.2. Gaussian Mixture Model GSA (GMM-GSA)

The GMM represents an arbitrary probability distribution using mixtures of Gaussian components. For a random variable X_i, its PDF $f_{X_i}(x_i)$ is approximated using a Q component Gaussian distribution as follows [56, 57, 58]

$$f_{X_i}(x) = \sum_{i=1}^{Q} \lambda_i \phi(x, \mu_i, \sigma_i^2), \tag{29}$$

where Q is the number of Gaussian components, $\phi(\cdot)$ is the PDF of a Gaussian random variable, λ_i, μ_i, and σ_i are the weight, mean, and standard deviation of the i-th Gaussian component.

For $\mathbf{Z} = (\mathbf{X}_c, Y) \in \mathbb{R}^{n_c+1}$, the joint PDF $f_{\mathbf{X}_c Y}(\mathbf{x}_c, y)$ is approximated using a multi-variate GMM as

$$f_{\mathbf{Z}}(\mathbf{z}) = \sum_{i=1}^{Q} \lambda_i \phi(\mathbf{z}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{30}$$

where $\mathbf{z} = [\mathbf{x}_c, y]^T$, $\boldsymbol{\mu}_i = [\boldsymbol{\mu}_{i,\mathbf{x}_c}, \mu_{i,y}]^T$, and

$$\boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_{i,cc} & \boldsymbol{\Sigma}_{i,cy} \\ \boldsymbol{\Sigma}_{i,yc} & \sigma_{i,y}^2 \end{bmatrix}. \tag{31}$$

The expectation maximization (EM) method [59] is commonly used to estimate the parameters of the GMM model. In the past decades, various approaches [60, 61, 62] have been proposed for learning of GMM from data. For example, Nanty et al. [63] recently proposed a new sparse method based on a Lasso penalization algorithm to estimate the GMM parameters and reduce the number of components in the model. For the sake of illustration, in this paper, the Matlab toolbox is used directly to learn the GMM from data of $\mathbf{X}_c$ and Y, and the Bayesian information criterion (BIC) [64] is adopted to select the number of components (i.e. Q) in GMM for GSA. Next, we discuss how to implement approach one and approach two using GMM to compute Sobol' indices.

3.2.1. GMM-GSA approach one

After $f_{\mathbf{X}_c Y}(\mathbf{x}_c, y)$ is approximated using GMM, for given $\mathbf{X}_c = \mathbf{x}_c$, the conditional PDF $f_{Y|\mathbf{X}_c}(y|\mathbf{x}_c)$ is given by

$$f_{Y|\mathbf{X}_c}(y|\mathbf{x}_c) = \sum_{i=1}^{Q} \lambda_i(\mathbf{x}_c)\phi(y, \mu_{i,y|\mathbf{x}_c}, \sigma^2_{i,y|\mathbf{x}_c}), \tag{32}$$

where

$$\mu_{i,y|\mathbf{x}_c} = \mu_{i,y} + \mathbf{\Sigma}_{i,yc}\mathbf{\Sigma}^{-1}_{i,cc}(\mathbf{x}_c - \mu_{i,\mathbf{x}_c}), \tag{33}$$

$$\sigma^2_{i,y|\mathbf{x}_c} = \sigma^2_{i,y} - \mathbf{\Sigma}_{i,yc}\mathbf{\Sigma}^{-1}_{i,cc}\mathbf{\Sigma}^T_{i,yc}, \tag{34}$$

and

$$\lambda_i(\mathbf{x}_c) = \frac{\lambda_i \phi(\mathbf{x}_c, \mu_{i,\mathbf{x}_c}, \mathbf{\Sigma}_{i,\mathbf{x}_c})}{\sum_{k=1}^{Q} \lambda_k \phi(\mathbf{x}_c, \mu_{k,\mathbf{x}_c}, \mathbf{\Sigma}_{k,\mathbf{x}_c})}. \tag{35}$$

Based on Eq. (32), $\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c)$ is computed as

$$\begin{aligned}\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) &= \int_{\Omega_Y} y f_{Y|\mathbf{X}_c}(y|\mathbf{x}_c)dy \\ &= \int_{\Omega_Y} y \sum_{i=1}^{Q} \lambda_i(\mathbf{x}_c)\phi(y, \mu_{i,y|\mathbf{x}_c}, \sigma^2_{i,y|\mathbf{x}_c})dy = \sum_{i=1}^{Q} \lambda_i(\mathbf{x}_c)\mu_{i,y|\mathbf{x}_c},\end{aligned} \tag{36}$$

With the GMM of $f_{\mathbf{X}_c Y}(\mathbf{x}_c, y)$, we generate N_{MCS} samples of $\mathbf{X}_c$. Based on the generated samples, we then have

$$\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c)) \approx \frac{1}{2N^2_{MCS}} \sum_{i=1}^{N_{MCS}} \sum_{j=1}^{N_{MCS}} \big(\sum_{k=1}^{Q} \lambda_k(\mathbf{x}_c^{(i)})\mu_{k,y|\mathbf{x}_c^{(i)}} - \sum_{k=1}^{Q} \lambda_k(\mathbf{x}_c^{(j)})\mu_{k,y|\mathbf{x}_c^{(j)}} \big)^2. \tag{37}$$

With Eq. (37) and Eqs. (3) and (4), the first-order and total-effect Sobol' indices can be computed.

3.2.2. GMM-GSA approach two

Similar to GC-GSA approach two, we first compute $\mathbb{E}_{\mathbf{X}_r}(Y^2|\mathbf{X}_c = \mathbf{x}_c)$. Using Eq. (32), we have

$$\begin{aligned}\mathbb{E}_{\mathbf{X}_r}(Y^2|\mathbf{X}_c = \mathbf{x}_c) &= \int_{\Omega_Y} y^2 f_{Y|\mathbf{X}_c}(y|\mathbf{x}_c)dy = \sum_{i=1}^{Q}\int_{\Omega_Y} y^2\lambda_i(\mathbf{x}_c)\phi(y, \mu_{i,y|\mathbf{x}_c}, \sigma^2_{i,y|\mathbf{x}_c})dy \\ &= \sum_{i=1}^{Q}\lambda_i(\mathbf{x}_c)(\mu^2_{i,y|\mathbf{x}_c} + \sigma^2_{i,y|\mathbf{x}_c}). \end{aligned} \tag{38}$$

Combining Eqs. (36) and (38) yields

$$\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) = \sum_{i=1}^{Q}\lambda_i(\mathbf{x}_c)(\mu^2_{i,y|\mathbf{x}_c} + \sigma^2_{i,y|\mathbf{x}_c}) - \Big(\sum_{i=1}^{Q}\lambda_i(\mathbf{x}_c)\mu_{i,y|\mathbf{x}_c}\Big)^2. \tag{39}$$

Based on the expression given in the above equation and random samples of $\mathbf{X}_c$ generated from the GMM model, $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ is then approximated as

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c)) \\ &\qquad\approx \frac{1}{N_{MCS}}\sum_{i=1}^{N_{MCS}}\Big[\sum_{k=1}^{Q}\lambda_k(\mathbf{x}_c^{(i)})(\mu^2_{k,y|\mathbf{x}_c^{(i)}} + \sigma^2_{k,y|\mathbf{x}_c^{(i)}}) - \Big(\sum_{k=1}^{Q}\lambda_k(\mathbf{x}_c^{(i)})\mu_{k,y|\mathbf{x}_c^{(i)}}\Big)^2\Big]. \end{aligned} \tag{40}$$

With Eq. (40), different types of Sobol' indices can be computed. Since $\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c)$ is *computed analytically* in Eq. (39), the variance of the estimation of $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ due to the limited number of samples can be quantified [65]. The general implementation procedure of the GMM-based GSA methods can be summarized as below:

- Step 1: Compute $\mathrm{Var}(Y)$ using data of Y given in Eq. (6) and extract data of $\mathbf{x}_c$ and y from the data matrix given in Eq. (6).
- Step 2: Learn GMM model based on the data of $\mathbf{x}_c$ and y.
- Step 3: Compute $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using Eq. (36) or $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using Eq. (40).
- Step 4: Compute the Sobol' indices.

In reality, it is quite possible that both the Gaussian copula model (i.e. Sec. 3.1) and GMM cannot well-represent the joint PDF $f_{\mathbf{X}_c Y}(\mathbf{x}_c, y)$. In addition to the above discussed GC-GSA and GMM-GSA approaches, in the subsequent section, we propose a new Gaussian mixture copula (GMC) model for GSA with given samples.

3.3. Gaussian Mixture Copula GSA (GMC-GSA)

Recall that in Sklar's Theorem [66], an n-dimensional distribution function, $F(x_1, x_2, \cdots, x_n)$ is connected with the marginal distributions, $F_1(x_1), \cdots, F_n(x_n)$ as follows

$$F(x_1, x_2, \cdots, x_n) = C(F_1(x_1), F_2(x_2), \cdots, F_n(x_n)) = C(u_1, u_2, \cdots, u_n). \tag{41}$$

In the above equation, $C(u_1, u_2, \cdots, u_n)$ is a unique function if $F_1(x_1), \cdots, F_n(x_n)$ are all continuous [66]. In addition, $C(u_1, u_2, \cdots, u_n)$ is well studied for few copula functions, such as Gaussian copula, student's t copula, and several other bivariate copulas [51]. In many practical applications, the current available copula functions, however, cannot accurately capture the complicated dependences between multiple variables. Inspired by the fact that kennel density function or GMM model can be employed to approximate an arbitrary probability distribution if the parametric distributions (e.g. Gaussian, Lognormal, or Weibull) are not applicable, this paper develops a GMC model by integrating non-parametric GMM model with the copula function given in Eq. (41). We refer the developed GMC model as *non-parametric copula* and the Gaussian copula used in Sec. 3.1 and other well-studied copula functions as *parametric copula*.

The non-parametric GMC model inherits the characteristics of the copula function given in Eq. (41) (i.e. the copula function is independent from the marginal distribution). Similar to Eq. (15),

we connect the marginal CDFs with the CDF $F(\mathbf{x}_c, y)$ through a generalized copula as follows

$$\begin{aligned} F(\mathbf{x}_c, y) &= \mathbb{P}\{\mathbf{X}_c \leq \mathbf{x}_c, Y \leq y\}, \\ &= C(F_1(x_1), F_2(x_2), \cdots, F_{n_c}(x_{n_c}), F_Y(y); \boldsymbol{\theta}), \\ &= C(u_1, u_2, \cdots, u_{n_c}, u_Y; \boldsymbol{\theta}), \\ &= C_{new}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_y); \boldsymbol{\theta}), \end{aligned} \tag{42}$$

where $C_{new}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_y); \boldsymbol{\theta})$ is an *unknown* new copula function (i.e. CDF function). Note that the unknown new copula is independent from the marginal distributions. This preserves the property of copula function defined in Sklar's Theorem in Eq. (41). The unknown new copula function will reduce to a Gaussian copula if the underlying model $C_{new}(\cdot)$ is a multivariate Gaussian distribution (see Eq. (18)).

Computing the derivatives of Eq. (42), we have the joint PDF as

$$\begin{aligned} f_{\mathbf{X}_c Y}(\mathbf{x}_c, y) &= \frac{\partial^{n_c+1} C_{new}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_y); \boldsymbol{\theta})}{\partial u_1 \partial u_2 \cdots \partial u_{n_c} \partial u_Y} \frac{\partial u_1}{\partial x_1} \frac{\partial u_2}{\partial x_2} \cdots \frac{\partial u_{n_c}}{\partial x_{n_c}} \frac{\partial u_Y}{\partial y}, \\ &= c_{new}(u_1, u_2, \cdots, u_{n_c}, u_y; \boldsymbol{\theta}) f_{X_1}(x_1) \cdots f_{X_n}(x_n) f_Y(y), \end{aligned} \tag{43}$$

in which $\frac{f_{\mathbf{X}_c Y}(\mathbf{x}_c, y)}{f_{X_1}(x_1) \cdots f_{X_n}(x_n) f_Y(y)} = c_{new}(u_1, u_2, \cdots, u_{n_c}, u_y; \boldsymbol{\theta})$ is the PDF of the new copula function and is given by

$$\begin{aligned} c_{new}(u_1, u_2, \cdots, u_{n_c}, u_y; \boldsymbol{\theta}) &= \frac{\partial^{n_c+1} C_{new}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_y); \boldsymbol{\theta})}{\partial u_1 \partial u_2 \cdots \partial u_{n_c} \partial u_Y}, \\ &= \frac{\partial^{n_c+1} C_{new}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_y); \boldsymbol{\theta})}{\partial \Phi^{-1}(u_1) \partial \Phi^{-1}(u_2) \cdots \partial \Phi^{-1}(u_{n_c}) \partial \Phi^{-1}(u_Y)} \\ &\quad \times \frac{\partial \Phi^{-1}(u_1)}{\partial u_1} \frac{\partial \Phi^{-1}(u_2)}{\partial u_2} \cdots \frac{\partial \Phi^{-1}(u_{n_c})}{\partial u_{n_c}} \frac{\partial \Phi^{-1}(u_Y)}{\partial u_Y}, \\ &= \frac{\partial^{n_c+1} C_{new}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_y); \boldsymbol{\theta})}{\partial \Phi^{-1}(u_1) \partial \Phi^{-1}(u_2) \cdots \partial \Phi^{-1}(u_{n_c}) \partial \Phi^{-1}(u_Y)} \\ &\quad \times \frac{1}{\phi(\Phi^{-1}(u_1))} \frac{1}{\phi(\Phi^{-1}(u_2))} \cdots \frac{1}{\phi(\Phi^{-1}(u_{n_c}))} \frac{1}{\phi(\Phi^{-1}(u_Y))}. \end{aligned} \tag{44}$$

Defining $f_{new}(u_1, u_2, \cdots, u_{n_c,u_Y}; \boldsymbol{\theta}) = \frac{\partial^{n_c+1} C_{new}(\Phi^{-1}(u_1),\Phi^{-1}(u_2),\cdots,\Phi^{-1}(u_{n_c}),\Phi^{-1}(u_y);\boldsymbol{\theta})}{\partial\Phi^{-1}(u_1)\partial\Phi^{-1}(u_2)\cdots\partial\Phi^{-1}(u_{n_c})\partial\Phi^{-1}(u_Y)}$, Eq. (43) can be rewritten as follows

$$\begin{aligned} f_{\mathbf{X}_c Y}(\mathbf{x}_c, y) &= c_{new}(u_1, u_2, \cdots, u_{n_c}, u_y; \boldsymbol{\theta}) f_{X_1}(x_1) \cdots f_{X_n}(x_n) f_Y(y), \\ &= \frac{1}{\phi(\Phi^{-1}(u_1))} \cdots \frac{1}{\phi(\Phi^{-1}(u_{n_c}))} \frac{1}{\phi(\Phi^{-1}(u_Y))} f_{new}(u_1, u_2, \cdots, u_{n_c}, u_Y; \boldsymbol{\theta}) \\ &\quad \times f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) f_Y(y). \end{aligned} \tag{45}$$

in which $f_{new}(u_1, u_2, \cdots, u_{n_c}, u_Y)$ is the PDF function of the unknown CDF function $C_{new}(\cdot)$. If the underlying function of $C_{new}(\cdot)$ is a multivariate Gaussian distribution, $f_{new}(u_1, u_2, \cdots, u_{n_c})$ is then the PDF of multivariate Gaussian distribution (see Eq. (19)). It should be noted that $f_{new}(u_1, u_2, \cdots, u_{n_c}; \boldsymbol{\theta})$ and $c_{new}(u_1, u_2, \cdots, u_{n_c}, u_y; \boldsymbol{\theta})$ are different functions.

A critical step to constructe the unknown non-parametric copula is the learning of the unknown PDF function $f_{new}(u_1, u_2, \cdots, u_{n_c}, u_Y; \boldsymbol{\theta})$. In this paper, the GMM model is employed to approximate $f_{new}(u_1, u_2, \cdots, u_{n_c}, u_Y; \boldsymbol{\theta})$ by taking advantage of the capability of GMM in approximating an arbitrary PDF based on data. $f_{new}(u_1, u_2, \cdots, u_{n_c}, u_Y; \boldsymbol{\theta})$ is approximated using a GMM as follows

$$f_{new}(u_1, u_2, \cdots, u_{n_c}, u_Y; \boldsymbol{\theta}) \approx \sum_{i=1}^{Q_c} \lambda_{i,w} \phi(\mathbf{w}, \mu_{i,w}, \boldsymbol{\Sigma}_{i,w}), \tag{46}$$

where $\mathbf{w} = [\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_{n_c}), \Phi^{-1}(u_Y)]$, Q_c is the number of Gaussian components, $\lambda_{i,w}$, $\mu_{i,w}$, and $\boldsymbol{\Sigma}_{i,w}$, which are the parameters $\boldsymbol{\theta}$ of the copula function, are respectively the weight, mean, and covariance of the i-th Gaussian component.

As shown in Eq. (46), in order to learn such an unknown PDF function from data using GMM, we first transform the data of $\mathbf{u}_c$ and u_Y into the standard normal space. This step is similar to that of Gaussian copula. Based on the GMM approximation of the PDF function, we have the PDF of the new copula function as

$$c_{new}(u_1, u_2, \cdots, u_{n_c}, u_y; \boldsymbol{\theta}) \approx \frac{1}{\phi(\Phi^{-1}(u_1))} \cdots \frac{1}{\phi(\Phi^{-1}(u_{n_c}))} \frac{1}{\phi(\Phi^{-1}(u_Y))} \sum_{i=1}^{Q_c} \lambda_{i,w} \phi(\mathbf{w}, \mu_{i,w}, \boldsymbol{\Sigma}_{i,w}). \tag{47}$$

The joint PDF of VoI and QoI, $f_{\mathbf{X}_c Y}(\mathbf{x}_c, y)$, is thus approximated by the new GMC model as

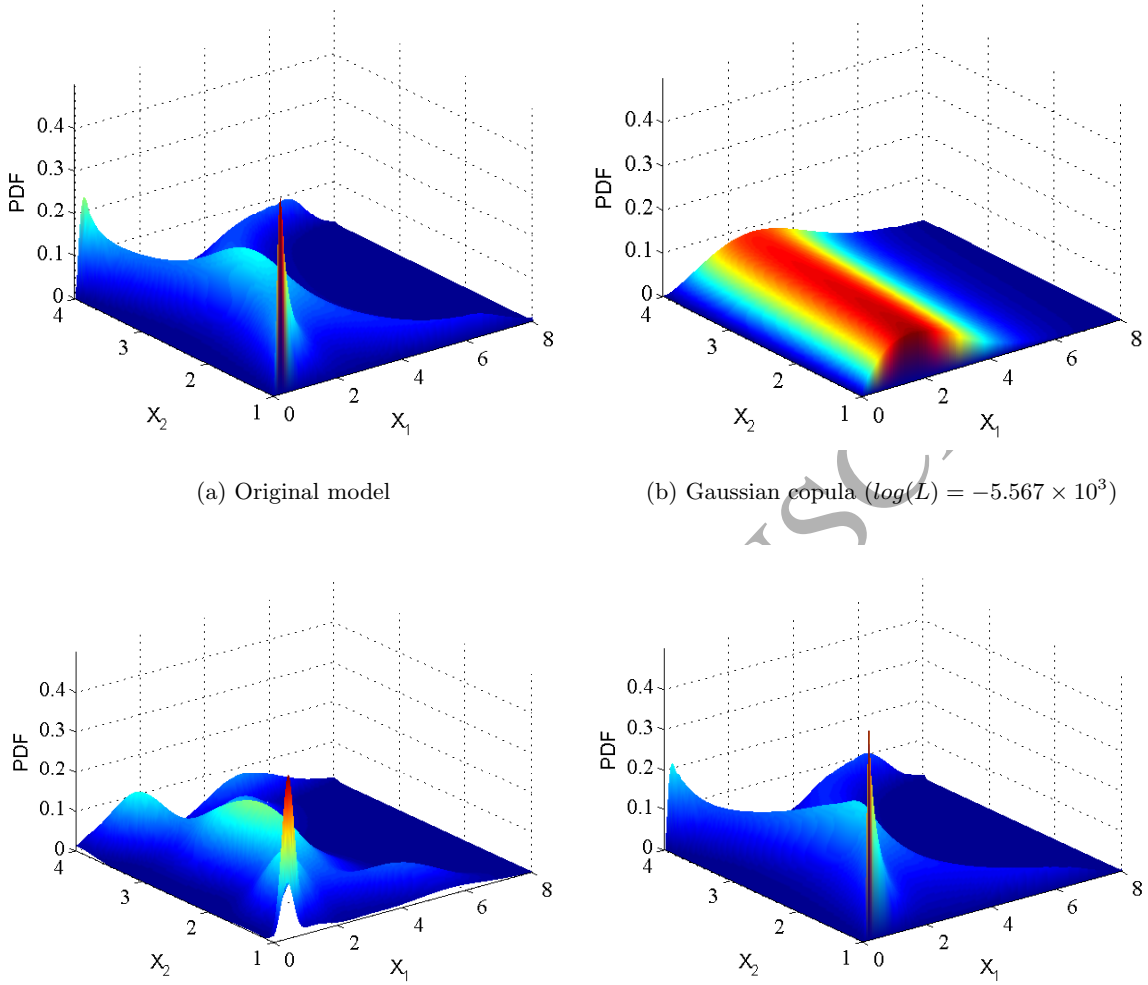
$$f_{\mathbf{X}_c Y}(\mathbf{x}_c, y) \approx \frac{1}{\phi(\Phi^{-1}(u_1))} \cdots \frac{1}{\phi(\Phi^{-1}(u_{n_c}))} \frac{1}{\phi(\Phi^{-1}(u_Y))} \sum_{i=1}^{Q_c} \lambda_{i,w} \phi(\mathbf{w}, \mu_{i,w}, \mathbf{\Sigma}_{i,w}) \times f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) f_Y(y). \tag{48}$$

Before applying the GMC model to GSA with given samples, we use a two-dimensional example to verify the effectiveness of the GMC model and to illustrate the advantage of the *non-parametric copula*. Suppose that we have two random variable X_1 and X_2 with marginal distributions given by $X_1 \sim \text{Weib}(x_1, 3, 2)$ and $X_2 \sim \text{Unif}(x_2, 1, 4)$, where $\text{Weib}(x_1, \cdot, \cdot)$ is the Weibull distribution and $\text{Unif}(x_2, \cdot, \cdot)$ is the uniform distribution. There is a complicated probabilistic function relationship between X_1 and X_2, which is modeled by a copula function $c(u_1, u_2)$ as follows

$$c(u_1, u_2) = 0.4 c_g(u_1, u_2, -0.85) + 0.6 c_t(u_1, u_2, 0.75, 3), \tag{49}$$

in which $c_g(u_1, u_2, -0.85)$ is a Gaussian copula with parameter -0.85 and $c_t(u_1, u_2, 0.75, 3)$ is a t copula with parameters 0.75 and 3.

We generate 2,000 data points of these two random variables based on this model and then learn the joint PDF between these two variables using different probability models (i.e. Gaussian copula, Gaussian mixture model, and Gaussian mixture copula). When we are learning the Gaussian mixture model and Gaussian mixture copula model, we use the same number of components to achieve a fair comparison. Fig. 2 gives the comparison of the joint PDF obtained from different probability models as well as the likelihood of observing the data for learning. The results presented in Fig. 2 indicate that the new Gaussian mixture copula model can represent the joint PDF of this example more accurately than the Gaussian mixture model and Gaussian copula model. This demonstrate the effectiveness and the advantage of the Gaussian mixture model in modeling the joint PDF.



(a) Original model

(b) Gaussian copula ($log(L) = -5.567 \times 10^3$)

(c) Gaussian mixture model ($log(L) = -5.160 \times 10^3$)

(d) Gaussian mixture copula ($log(L) = -5.055 \times 10^3$)

Figure 2: PDF comparison of different probability models

Next, we discuss how to compute $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ or $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using the new GMC model.

3.3.1. GMC-GSA approach one

Since the new GMC model is also a copula model, we can compute $\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c)$ similar to Eq. (21) as below

$$\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) = \int_{\Omega_Y} y f_{Y|\mathbf{X}_c}(y|\mathbf{x}_c)dy = \int_{\Omega_Y} y\frac{f_{\mathbf{X}_c Y}(\mathbf{x}_c, y)}{f_{\mathbf{X}_c}(\mathbf{x}_c)}dy, \tag{50}$$

where $f_{\mathbf{X}_c}(\mathbf{x}_c)$ is given by

$$f_{\mathbf{X}_c}(\mathbf{x}_c) \approx \frac{1}{\phi(\Phi^{-1}(u_1))} \cdots \frac{1}{\phi(\Phi^{-1}(u_{n_c}))} f_{new}(u_1, u_2, \cdots, u_{n_c}; \boldsymbol{\theta}_c) f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n). \tag{51}$$

Combining Eqs. (48) and (51) with Eq. (50), we have

$$\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) \approx \int_{\Omega_Y} y f_Y(y) \frac{1}{\phi(\Phi^{-1}(u_Y))} \sum_{i=1}^{Q_c} \lambda_{i,w}(\Phi^{-1}(\mathbf{u}_c))\phi(\Phi^{-1}(u_Y),$$
$$\mu_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)}, \sigma^2_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)})dy, \tag{52}$$

in which $\mathbf{u}_c = [F_1(x_1), F_2(x_2), \cdots, F_{n_c}(x_{n_c})]$, $\Phi^{-1}(\mathbf{u}_c) = [\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_{n_c})]$, and $\mu_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)}$, $\sigma_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)}$, and $\lambda_{i,w}(\Phi^{-1}(\mathbf{u}_c)))$ are the conditional mean, standard deviation, and weights of Y computed using Eqs. (33), (34), and (35), respectively.

The above equation can be rewritten as

$$\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) \approx \int_0^1 F_Y^{-1}(u_Y) \frac{1}{\phi(\Phi^{-1}(u_Y))} \sum_{i=1}^{Q_c} \lambda_{i,w}(\Phi^{-1}(\mathbf{u}_c))\phi(\Phi^{-1}(u_Y),$$
$$\mu_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)}, \sigma^2_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)})du_Y. \tag{53}$$

Similar to Eq. (32), Eq. (53) is approximated using MCS integration method as follows

$$\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c) \approx \frac{1}{N_{MCS}} \sum_{k=1}^{N_{MCS}} F_Y^{-1}(u_Y^{(k)}) \frac{1}{\phi(\Phi^{-1}(u_Y^{(k)}))} \sum_{i=1}^{Q_c} \lambda_{i,w}(\Phi^{-1}(\mathbf{u}_c))\phi(\Phi^{-1}(u_Y^{(k)}),$$
$$\mu_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)}, \sigma^2_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)}). \tag{54}$$

Based on Eq. (54), we estimate $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using a sampling-based method. Since the implementation procedure is the same as that of Eq. (24), we do not provide details here.

3.3.2. GMC-GSA approach two

Similar to the GC-GSA approach two (i.e. Sec. 3.1.2), we can estimate $\mathbb{E}_{\mathbf{X}_r}(Y^2|\mathbf{X}_c = \mathbf{x}_c)$ as follows

$$\mathbb{E}_{\mathbf{X}_r}(Y^2|\mathbf{X}_c = \mathbf{x}_c) \approx \frac{1}{N_{MCS}} \sum_{k=1}^{N_{MCS}} \Big[\big(F_Y^{-1}(u_Y^{(k)})\big)^2 \frac{1}{\phi(\Phi^{-1}(u_Y^{(k)}))} \sum_{i=1}^{Q_c} \lambda_{i,w}(\Phi^{-1}(\mathbf{u}_c)) \phi(\Phi^{-1}(u_Y^{(k)}), \mu_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)}, \sigma^2_{i,\Phi^{-1}(u_Y)|\Phi^{-1}(\mathbf{u}_c)}) \Big]. \quad (55)$$

Combining Eqs. (54) and (55), $\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c)$ can be evaluated similar to Eq. (27). Once $\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c)$ is available, similar to Eq. (28), $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ is estimated using the random samples of $\mathbf{X}_c$ generated using the GMC model. The implementation procedure of GMC-GSA is similar to that of GC-GSA (i.e. Sec. 3.1) and can be summarized into the following five steps:

- Step 1: Compute $\mathrm{Var}(Y)$ using data of Y given in Eq. (6), convert the data matrix given in Eq. (6) into the data of $\mathbf{u}_X$ and u_Y using kernel density function [55] and into data of equivalent standard normal distribution $\Phi^{-1}(\mathbf{u}_X))$ and $\Phi^{-1}(u_Y)$.
- Step 2: Extract data of $\Phi^{-1}(\mathbf{u}_c))$ and $\Phi^{-1}(u_Y)$ from the data matrix obtained from Step 1.
- Step 3: Fit a GMM for the new PDF function $f_{new}(u_1, u_2, \cdots, u_{n_c}, u_Y; \boldsymbol{\theta})$ using the data of $\Phi^{-1}(\mathbf{u}_c))$ and $\Phi^{-1}(u_Y)$.
- Step 4: Compute $\mathrm{Var}_{\mathbf{X}_c}(\mathbb{E}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using Eq. (54) or $\mathbb{E}_{\mathbf{X}_c}(\mathrm{Var}_{\mathbf{X}_r}(Y|\mathbf{X}_c = \mathbf{x}_c))$ using Eqs. (54) and (55).
- Step 5: Compute the Sobol' indices (e.g. first-order indices using Eq. (3) or (12) and total-effect indices using Eq. (4)) or (5).

3.4. Summary

This section presented PM-GSA approaches for GSA with given samples based on three probability models with increasing complexity, namely Gaussian copula model, Gaussian mixture model, and a new Gaussian mixture copula model. Two ways of computing GSA are investigated for each probability model. Comparing to *surrogate model-based GSA approaches* [18, 36, 13], the PM-GSA methods have the following advantages:

1. The proposed methods do not have requirements or assumptions on the model form of the underlying model. This allows us to perform various GSA computations for problems with complicated input-output connections.
2. The dimension of the probability models in the PM-GSA approaches is independent from the number of dimensions in the original physical model, and is always $N + 1$, where N is the order of the Sobol' indices. For example,when the first-order Sobol' indices of single variables need to computed, the joint PDF, $f_{\mathbf{X}_c Y}(\mathbf{x}_c, y)$, is only two-dimensional. This allows the proposed method to be applied to high-dimensional problems (≥ 40), for which the surrogate model-based methods will suffer from the curse of dimensionality [67].
3. Since the method is developed based on probability models, it is robust to the noises (singular) and outliers in the data, which makes it attractive to many practical problems.

Comparing to current *GSA approaches for given samples* [38, 42], the PM-GSA methods have the following advantages:

1. It can be applied to compute various Sobol' indices, such as first-order, higher-order, and total-effect Sobol' indices and Sobol' indices of correlated input variables or a set of input random variables, purely based on a data matrix. This overcomes the limitations of current GSA approaches with given samples, which can only compute the first-order Sobol' indices of single variables .

2. It has less parameters to tune than current available GSA methods with given samples.

The new Gaussian mixture copula model is applied to compute Sobol' indices in this paper. The application of this probability model, however, is not limited to GSA. It can be applied to various problems where the joint PDF of variables need to be modeled based on data, such as uncertainty quantification based on data [68] and uncertainty modeling based on data in design under uncertainty [69]. Next, we use four numerical examples including three benchmark mathematical examples and an engineering application example to illustrate the effectiveness of different PM-GSA methods. Based on the results of the numerical examples, we analyze the advantages and disadvantages of different methods.

4. Numerical Examples

In this section, the Sobol' indices of each example are computed using seven approaches: Gaussian copula-based approach one (GC1), Gaussian copula-based approach two (GC2), Gaussian mixture model-based approach one (GMM1), Gaussian mixture model-based approach two (GMM2), Gaussian mixture copula-based approach one (GMC1), Gaussian mixture copula-based approach two (GMC2), and MGSA [38]. In the first three benchmark mathematical examples, the seven methods are compared with Sobol' indices which are analytically available in the literature. In the engineering application example, the seven methods are compared with the double-loop MCS method (DMCS) since the analytical expressions are not available. In the DMCS method, 2×10^4 samples are used in the inner and outer loops, respectively. In addition, the MGSA method is only adopted to compute the first-order Sobol' indices of single input variables due to the limitations discussed in Sec. 2.

4.1. Rastrigin test function

The Rastrigin test function [39] is employed as our first example to demonstrate the effectiveness of the proposed PM-GSA approaches comparing to current GSA approaches for given samples [38, 39]. The Rastrigin test function is given by

$$Y = 10d + \sum_{i=1}^{d} (X_i^2 - \cos(2\pi X_i))^2, \tag{56}$$

where $d = 10$ is the number of input variables and $X_i, i = 1, 2, \cdots, d$ are identical independent random variables follow uniform distributions over $[-3, 3]$.

The first-order Sobol' indices of all the random variables can be analytically derived as $1/d$ (i.e. 0.1 for this example). We assume that the Rastrigin test function is unknown and randomly generate 400 MCS samples of the inputs and output. Only 400 samples are used here to keep it consistent with the number of samples used in Ref. [39]. Based on the generated samples, we then compute the first-order indices of input variables using different methods. To account for the uncertainty in the generated random samples, the above procedure is implemented repeatedly for 50 times. Fig. 3 shows the first-order Sobol' indices computed from different methods (i.e. GC1, GC2, GMM1, GMM2, GMC1, GMC2, MGSA) and the associated one standard deviation error bar due to data uncertainty of the 400 samples and the model uncertainty of the probability models. In this figure, we also give the results of the local polynomial smoother method (referred as LP) [39], which are extracted from Ref. [39]. Note that Ref. [39] did not provide the error bar of the LP method. We therefore only present the corresponding deterministic results of the LP method in Fig. 3.

From the results given in Fig. 3, we obtain the following two major findings:

- For this particular example, the Gaussian copula model-based GSA methods (i.e. GC1 and GC2) are much less accurate than the other methods. Between GC1 and GC2, GC1 is more

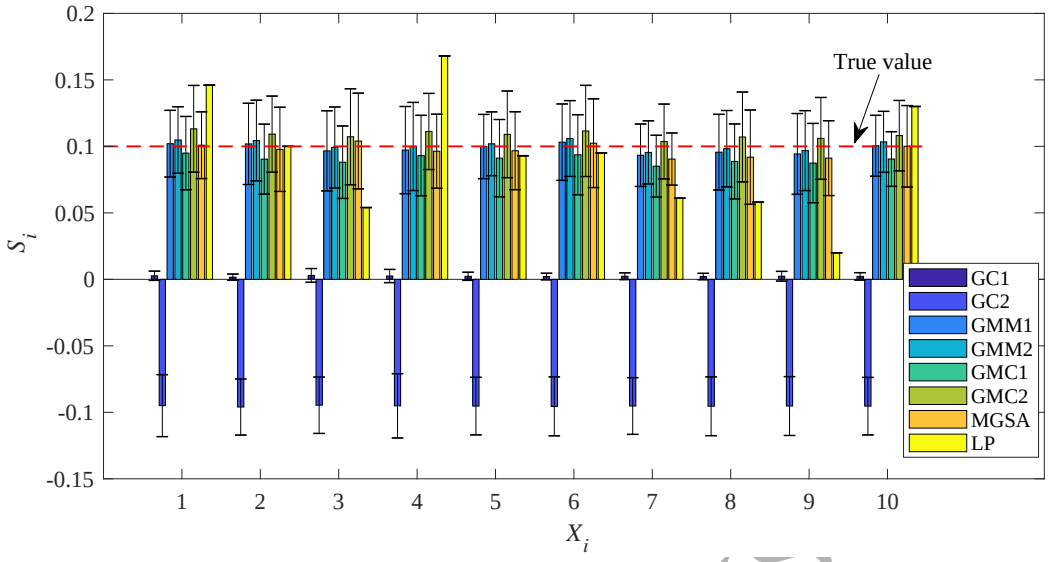


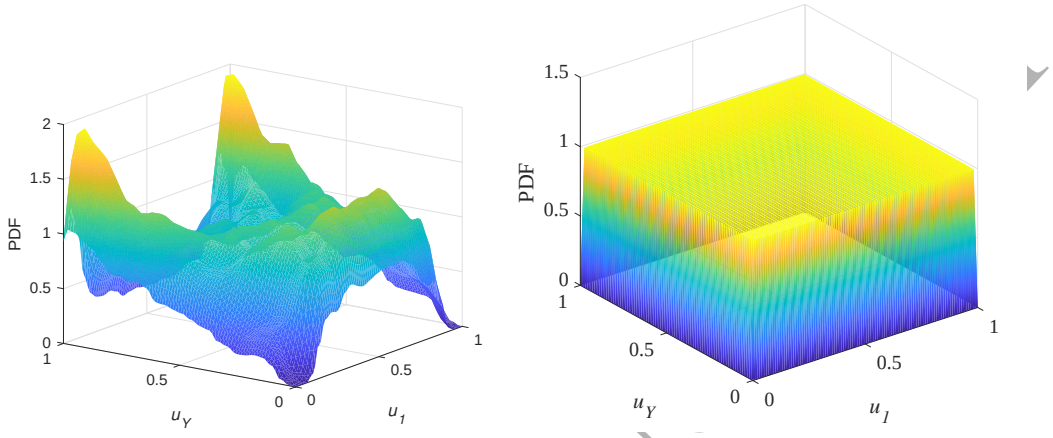
Figure 3: First-order Sobol indices of the Rastrigin test function

accurate than GC2.

- With the same number of samples (i.e. 400), the Gaussian mixture model-based GSA methods (i.e. GMM1 and GMM2) and the Gaussian mixture copula-based GSA methods (GMC1 and GMC2) can reach to the same accuracy level as that of the local polynomial smoother method [39] and MGSA method [38].

In order to explain the first finding, we take X_1 and Y as an example and plot the joint PDF of U_1 and U_Y. Fig. 4 depicts the true joint PDF of U_1 and U_Y obtained based on a large number of samples (i.e. 1×10^5) and the approximated joint PDF (i.e. $c(u_1, u_Y)$) from Gaussian copula. It shows that Gaussian copula is apparently not a good option to approximate the joint PDF of U_1 and U_Y. Due to the quadratic function shape given in Eq. (56) and the symmetric input domain of X_1 with respect to the origin, the correlation between X_1 and Y is almost zero. As shown in Fig. 4, the joint PDF in the CDF domain appears to be almost constant for the Gaussian coupla approximation. Using Gaussian copula can lead to large errors. This explains why the GC methods

give us bad results as shown in Fig. 3. The good results of GMC methods shown in Fig. 3 imply that the new non-parametric copula (i.e. GMC) does not have the above discussed limitation of Gaussian copula.



(a) True Joint PDF (b) Gaussian copula

Figure 4: Comparison of joint PDF in the CDF domain

4.2. Sobol' function

The Sobol' function [70], is adopted as our second example to show the capability of the PM-GSA methods for problems with high-dimensional variables. This function has been widely used as a benchmark problem for GSA in the literature [13] and is given by

$$Y = f(\mathbf{X}) = \prod_{i=1}^{K} \frac{|4X_i - 2| + a_i}{1 + a_i}, \tag{57}$$

in which $X_i, i = 1, 2, \cdots, K$ are random variables follow Uniform distributions [0, 1], $K = 60$ in this example, and $a_i, i = 1, 2, \cdots, K$ are coefficients given by

$$a_i = 25|\sin(0.5i) + \cos(0.75i + 2)|, i = 1, \cdots, 60. \tag{58}$$

The first-order indices of this function can be computed analytically and is given by [13].

$$S_i = \frac{1}{3\mathrm{Var}(Y)(a_i + 1)^2}, \forall i = 1, 2, \cdots, 60, \tag{59}$$

where

$$\mathrm{Var}(Y) = -1 + \prod_{i=1}^{K} \left[\frac{1}{3(a_i+1)^2} + 1 \right]. \tag{60}$$

We also assume that this function is unknown and generate synthetic data from the model to compute Sobol' indices using the proposed data-driven GSA methods. We generate only 2,000 MCS samples as our given samples for this 60 dimensional problem and repeat this multiple times to compute the variance in the computed Sobol' indices. Figs. 5, 6, and 7 show the first-order Sobol' indices obtained from different methods comparing to the true values. The results imply that the Gaussian copula fails to compute the Sobol' indices accordingly for this example. GMM1, GMM2, and GMC1 can accurately estimate the first-order Sobol' indices for variables $X_i, i = 1, 2, \cdots, 60$. GMC2 can estimate the first Sobol' indices. But the accuracy of GMC2 is not as good as the other three methods. Among GMM1, GMM2 and GMC1, the performance of GMC1 is the best. It indicates that the performance of GMM-based methods is more robust than the GMC-based methods even if the GMC-based methods sometime can give us better results. This phenomenon can be attributed to the transformation from the original data into the data of CDF values using a kernel density smoothing function in GMC. The kernel density function introduces an additional layer of uncertainty to the GMC model. When the CDF values can be computed analytically (i.e. the distributions of the input variables are known), this part of the uncertainty is expected to be reduced and the robustness of GMC-based methods can be improved.

4.3. A nonlinear model with nonlinear dependences

A nonlinear model with nonlinear dependences given in Ref. [47] is employed as our third example to illustrate the effectiveness of the proposed methods in handling dependent variables and a set of variables. The nonlinear function is given by

$$Y = f(\mathbf{X}) = X_1 X_2 + X_3 X_4, \tag{61}$$

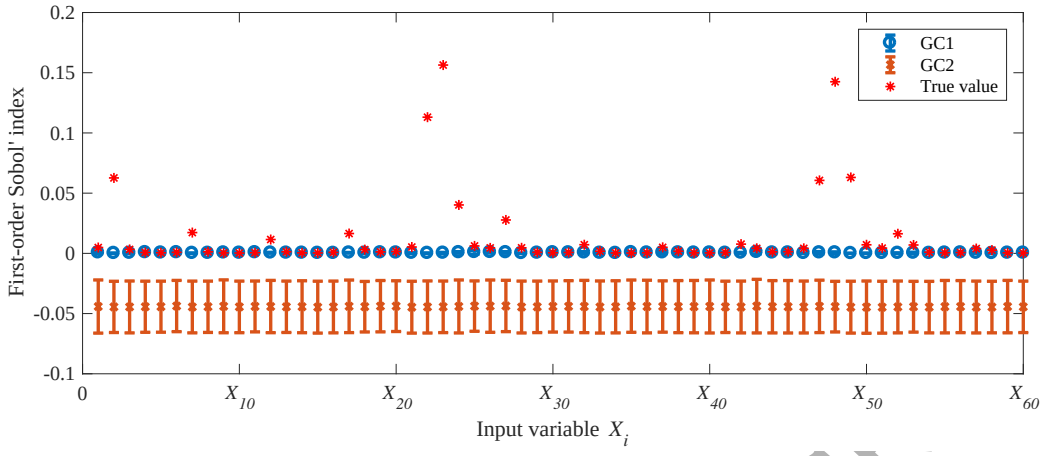


Figure 5: Comparison of Gaussian copula-based methods and the true values for the Sobol' function

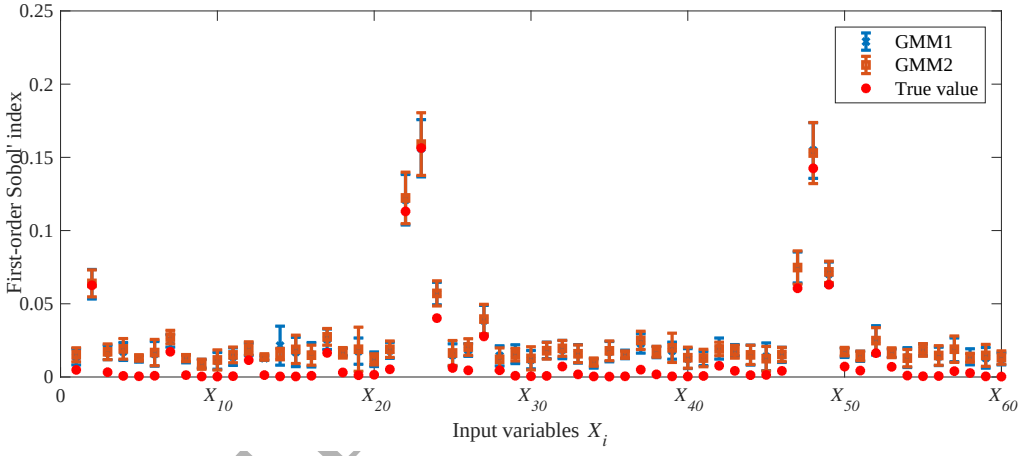


Figure 6: Comparison of Gaussian mixture model-based methods and the true values for the Sobol' function

where $(X_1, X_2) \in [0,1]^2$ is uniformly distributed within the triangle $X_1 + X_2 \leq 1$, $(X_3, X_4) \in [0,1]^2$ is uniformly distributed within the triangle $X_3 + X_4 \geq 1$, X_1 and X_2 are dependent due to the shared hidden variable L_1, and X_3 and X_4 are dependent due to the shared hidden variable L_3.

X_1 and X_2 are given by

$$\begin{aligned} X_1 &= 1 - \sqrt{1 - L_1}, \\ X_2 &= L_2 \sqrt{1 - L_1}, \end{aligned} \tag{62}$$

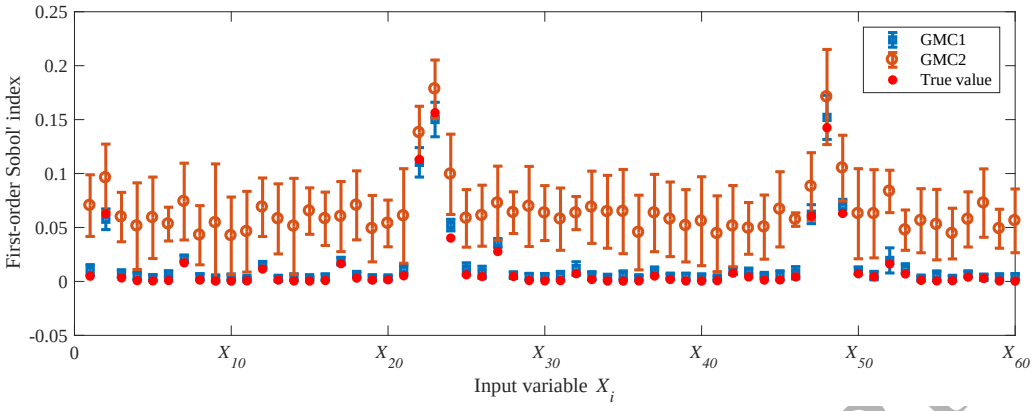


Figure 7: Comparison of Gaussian mixture copula-based methods and the true values for the Sobol' function

in which $(L_1 \neq 1, L_2) \in [0,1]^2$ are uniformly distributed hidden variables.

X_3 and X_4 are given by

$$\begin{aligned} X_3 &= \sqrt{L_3}, \\ X_4 &= (L_4 - 1)\sqrt{L_3} + 1, \end{aligned} \tag{63}$$

in which $(L_3, L_4) \in [0,1]^2$ are also uniformly distributed hidden variables.

We assume that the nonlinear function (i.e. Eq. (61)) and the nonlinear dependence (i.e. Eq. (62) and (63)) are unknown, and perform GSA purely based on given samples of $X_i, i = 1,2,3,4$ and Y. Similar to Ref. [47], we generate 1024 MCS samples of $X_i, i = 1,2,3,4$ and Y. Using the generated MCS samples as given samples, we then compute various Sobol' indices. Same as the previous two examples (Secs. 4.1 and 4.2), the procedure of generating random samples and computing Sobol' indices is implemented repeatedly for 50 times to assess the data uncertainty due to the limited number of samples.

4.3.1. First-order and total-effect Sobol' indices of dependent variables

We first compute the first-order and total-effect Sobol' indices of individual dependent random variables using Eqs. (3) and (5). Note that according to the definitions given in Ref. [2], the obtained first-order indices and total-effect indices are respectively the *full* first-order sensitivity

indices and the *independent* total-effect indices. The analytical values (i.e. true values) of the first-order and total-effect Sobol' indices of $X_i, i = 1, \cdots, 4$, are given in Ref. [47] as: $S_1 = S_2 = 1/30$, $S_3 = S_4 = 7/30$, $S_1^T = S_2^T = 1/15$, and $S_3^T = S_4^T = 2/3$. Fig. 8 gives the first-order indices obtained from different methods. Similar conclusions can be obtained as that from examples 1 and 2. The Gaussian copula-based GSA methods cannot accurately estimate the first-order Sobol' indices whereas the GMM and GMC-based methods can accurately estimate the first-order Sobol' indices for dependent variables. In addition, the uncertainty in the estimation of Sobol' indices is very small for this example.

Fig. 9 gives the results comparison of the total-effect Sobol' indices obtained from different

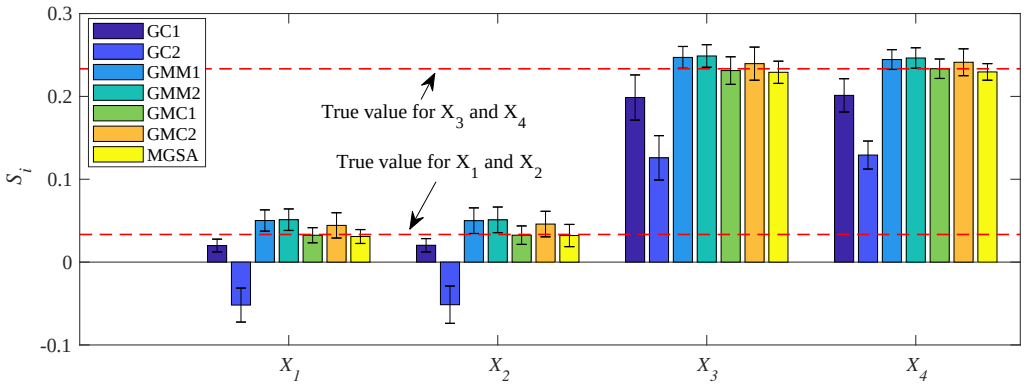


Figure 8: First-order Sobol' indices of dependent variables

methods. The results show that GMM1, GMM2, GMC1, and GMC2 can effectively estimate the total-effect Sobol' indices. Note that MGSA method is not presented in Fig. 9 since MGSA cannot be used to compute the total-effect indices.

4.3.2. First-order indices of sets of variables

We then investigate the capability of the proposed method in computing Sobol' indices of sets of variables. We compute the first-order Sobol' indices of X_1, (X_1, X_2), and (X_1, X_2, X_3). The analytical values of the first-order indices of these sets of variables are given by [47], $S_1 = 1/30$,

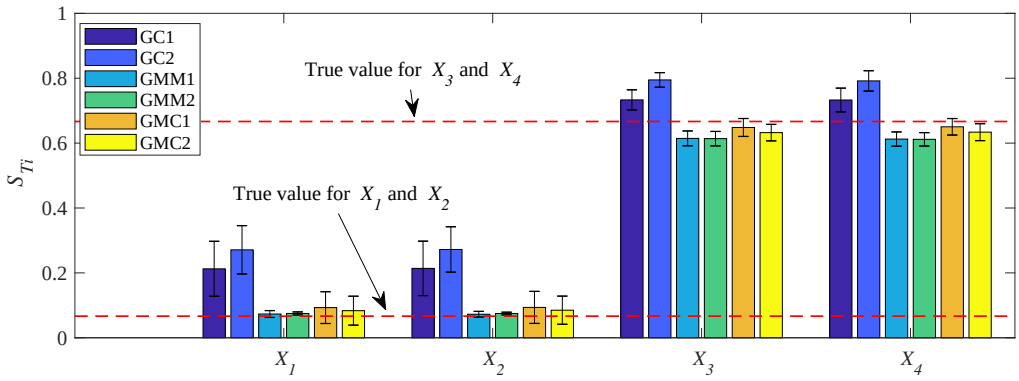


Figure 9: Total-effect indices of dependent variables

$S_{12} = 1/10$, $S_{123} = 1/3$. Fig. 10 gives the results comparison of the PM-GSA methods with the analytical values and the non-parametric method proposed in Ref. [47] (referred as NPM in the figure). MGSA method is not used for comparison since MGSA cannot be applied to GSA of sets of variables. The results show that the PM-GSA approaches (except GC1 and GC2) can accurately estimate the first-order Sobol' indices of set of variables. In addition, the variability of the results obtained from the PM-GSA methods is less than that of the NPM method. The above results demonstrate the effectiveness of the proposed PM-GSA methods for GSA with dependent random variables and set of variables.

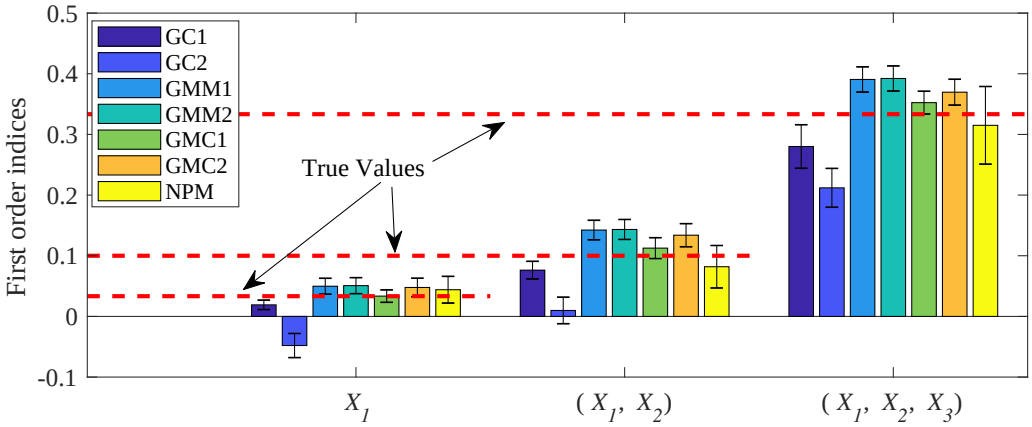


Figure 10: First-order Sobol indices of sets of variables

4.4. A cantilever beam

4.4.1. Problem description

A cantilever beam as depicted in Fig. 11 is modified from Ref. [38] as our fourth example, by changing the distribution of E to a multi-modal distribution. This example is used to show the effectiveness of the proposed method in computing higher-order indices and total-effect indices, and to investigate the effects of sample size on the accuracy of GSA results. The quantity of interest of this example is the tip deflection of the beam, which is given by the Timoshenko beam theory as follows [71]

$$Y = \frac{2p}{Ebh^3}\left[(4+5\upsilon)\frac{h^2L}{4} + 2L^3\right], \tag{64}$$

where $\mathbf{X} = [X_1, X_2, X_3, X_4, X_5, X_6] = [p, E, \upsilon, b, h, L]$.

The distribution of E (GPa) is represented as a multi-modal distribution as below

$$f_E(e) = 0.4N(e, 200, 1) + 0.6N(e, 190, 0.95^2), \tag{65}$$

in which $N(e, a, b^2)$ stands for normal distribution with mean of a and standard deviation of b. Table 1 gives the information of the other random variables of the beam example. Note that Y is an output variable in this example.

Table 1: Random variables of the cantilever beam example

Variable	p (kN)	υ	b (m)	h (m)	L (m)
Distribution	Normal	Normal	Lognormal	Lognormal	Lognormal
Mean	65	0.225	0.2	0.3	1.5
Standard deviation	0.5	0.03	0.02	0.02	0.05

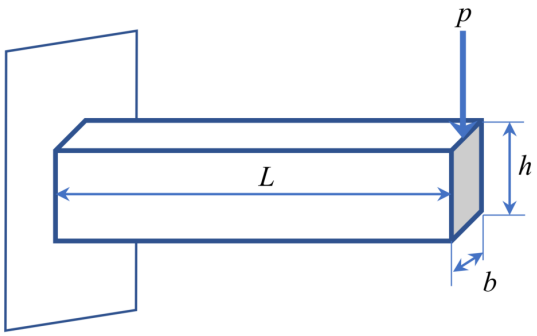


Figure 11: A cantilever beam

4.4.2. Closed indices [34]

Based on the above information of the cantilever beam, we generate MCS samples for the input random variables and response variable and use that as the data matrix given in Eq. (6). We then compute the closed indices of any two random variables given in Eq. (64) using different types of PM-GSA methods. The closed index of two random variables includes the main effects of individual random variables and the second-order effects [34]. In order to verify the effectiveness of different methods, we compare the results of these methods with those obtained from DMCS (i.e. 2×10^4 samples for both inner and outer loops). Fig. 12 shows the results comparison of the closed indices obtained from different methods with the number of samples (i.e. s in Eq. (6)) equals to 2000. The results show that the proposed PM-GSA methods can accurately estimate the closed indices. The GC-based methods also perform pretty well even if the GC1 and GC2 give us bad results in the previous three examples.

4.4.3. Total-effect indices

Based on the same samples used to compute the closed indices, we compute the total-effect Sobol' indices of different variables. Fig. 13 presents the results comparison of the total-effect Sobol' indices obtained from different methods. Note that only the total-effect Sobol' indices of the three dominant variables (i.e. $X_4, X_5, \text{and} X_6$) are depicted in the figure. The results show that the

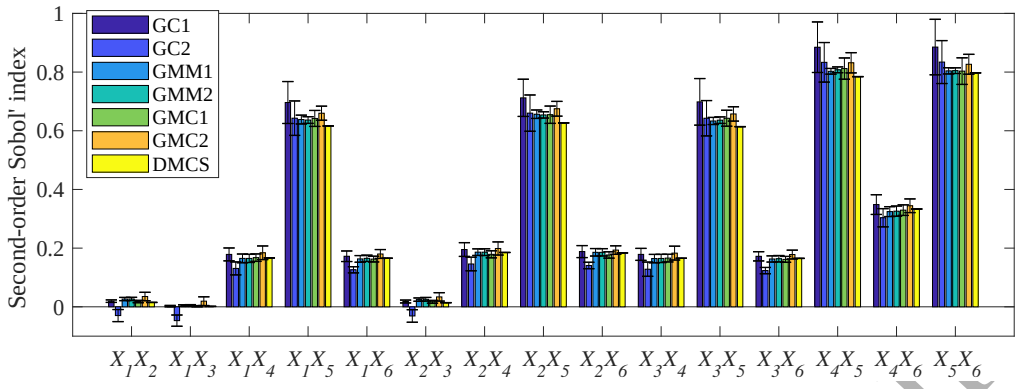


Figure 12: Closed indices of two random variables of the cantilever beam

proposed PM-GSA methods can accurately estimate the total-effect indices of the beam example.

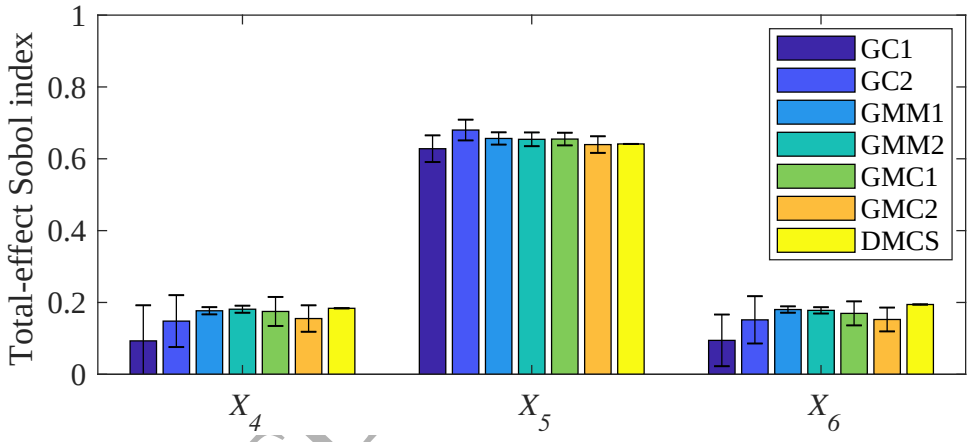


Figure 13: Total-effect Sobol' indices of the cantilever beam

4.4.4. Effect of sample size

In order to study the effect of sample size on the results of PM-GSA approaches, we compute the first-order and total-effect Sobol' indices of the dominant variables ($X_4, X_5,$ andX_6) with different number of samples (the numbers of samples equal to 200, 500, 1000, and 1500, respectively). Fig. 14 shows the results comparison of the first-order Sobol' indices obtained from different methods. Fig. 15 gives the results comparison of the total-effect Sobol' indices obtained from different methods. The results show that increasing the number of samples can increase the accuracy of

different methods. GMM-based methods tends to be more robust to the number of samples than the GMC-based methods.

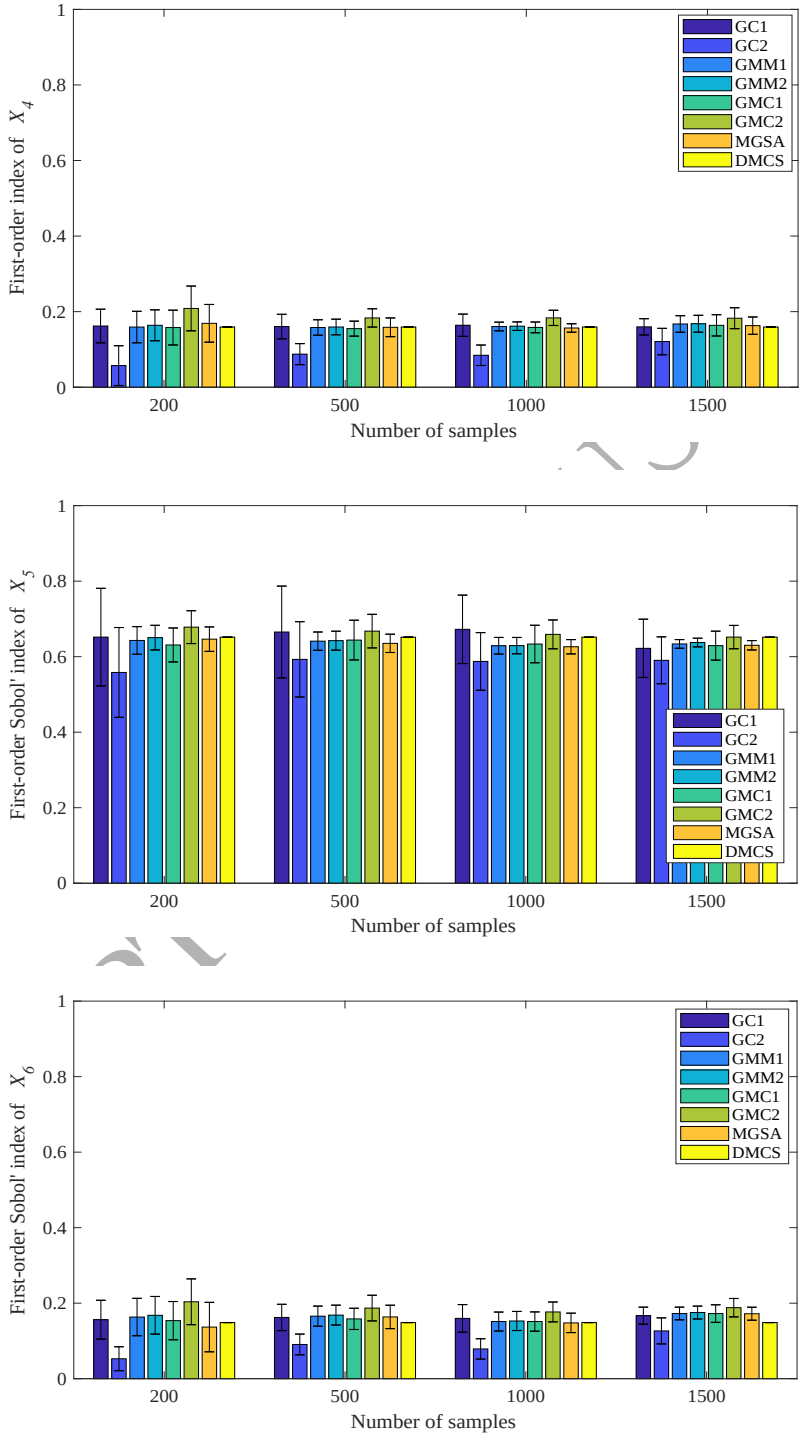


Figure 14: First-order Sobol' indices with different number of samples

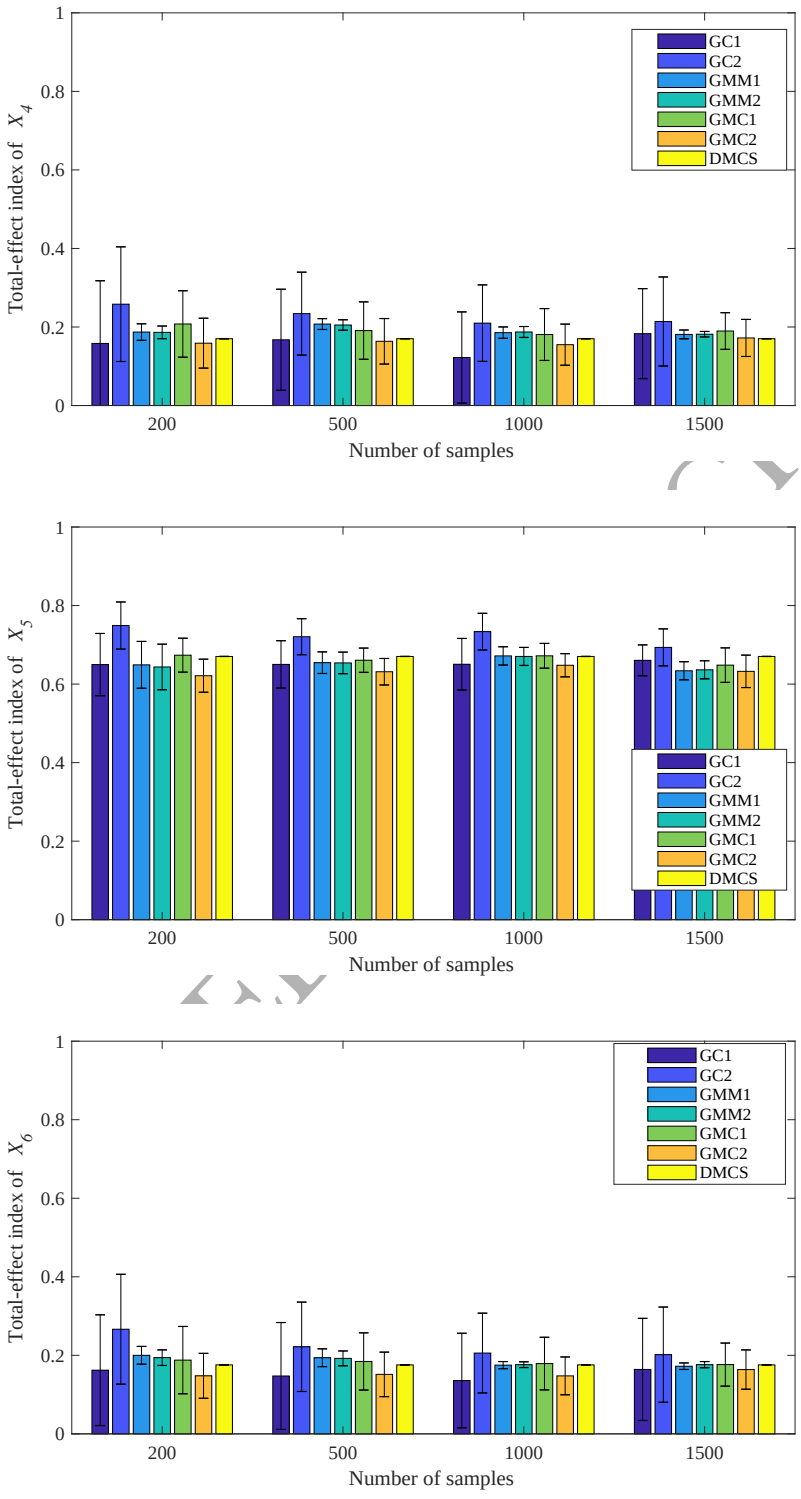


Figure 15: Total-effect Sobol' indices with different number of samples

4.5. Discussion

After analyzing the results of the above four examples, we obtain the following major findings of the proposed PM-GSA approaches:

1. GMM and GMC-based GSA methods are in general much more accurate than GC-based methods.
2. With the same number of samples, the PM-GSA approaches can reach to similar accuracy level as that of current GSA approaches with given samples [39, 38].
3. The PM-GSA approaches are applicable to problems with high-dimensional variables.
4. The PM-GSA approaches can effectively compute higher-order Sobol' indices, total-effect indices, and indices of sets of variables. The estimation of these indices cannot be accomplished by current GSA approaches with given samples [39, 38].
5. For the GMM and GMC based methods, sometimes GMC1 and GMC2 are more accurate than GMM1 and GMM2. Sometimes, it is the opposite. In practical applications, the decision maker needs to decide which model to use based on certain criteria, such as the likelihood, BIC, or AIC score.
6. The GMM-based methods tends to be more robust than the GMC-based methods even if the GMC-based methods can sometimes give us better results. This is caused by the additional layer of kernel density function for the data transformation in the GMC model. When the CDF functions of input variables are analytically available, the GMC-based method will be as robust as the GMM-based methods.
7. The accuracy of the PM-GSA approaches will increase with the number of available samples.

The above findings demonstrate the advantages of the proposed PM-GSA approaches as discussed in Sec.3.4.

5. Conclusions

This paper presents methods to compute various Sobol' indices such as the first-order, second-order, and total-effect Sobol' indices, and Sobol' indices of a set of variables, purely based on available input-output data. The data may be available from physical experiments, field observations, etc. Noises and unknown uncertainty variables are also allowed to be presented in the data. In the proposed methods, data of the variables of interest are extracted first from the available data matrix. Based on the extracted data, probability models are built to approximate the joint PDF of the variables of interest and QoI. With the probability models learned from the data, various types of Sobol' indices are computed. The proposed framework is first investigated using the Gaussian copula and Gaussian mixture models. Then a new Gaussian mixture copula model is proposed to build a robust probability model to compute the Sobol' indices by taking advantages of both Gaussian copula and Gaussian mixture model. Four numerical examples are studied to analyze the advantages and disadvantages of different PM-GSA methods. The results show that the Gaussian mixture and Gaussian mixture copula models-based GSA methods are able to accurately compute various Sobol' indices only based on an available data matrix. This allows us to compute various Sobol' indices for problems, where we can only collect limited number of data points due to constraints of either computational or experimental resources.

In the proposed method, joint distribution models are built using Gaussian copula, Gaussian mixture model, or Gaussian mixture copula. Similar to surrogate model-based GSA, building these joint distribution models require computational effort; however, this is negligible compared to data collection (i.e. Eq. (6)), such as function evaluations with a physics model or experiments/field data collection with the actual system. In addition, the computational effort is even more negligible in first-order Sobol' index computation since only the dimension of the model is only two.

Four numerical examples are used to demonstrate the effectiveness of the proposed method,

using synthetic data. Applications of the proposed method to data collected from practical field observations will be investigated in our future work. In addition, the following three topics are also worth pursuing in future: (1) Integration of the proposed PM-GSA methods with Bayesian methods to effectively quantify the uncertainty in the analysis results due to the limited number of data as well as errors in the probability model parameters; (2) Ensemble of various probability models, which will lead to a generic probability model with better accuracy than all the member probability models; and (3) Extension of the proposed approaches to other types of GSA indices, such as the measures proposed in Ref. [12].

Acknowledgement

The research reported in this paper was supported by the Air Force Office of Scientific Research (Grant No. FA9550-15-1- 0018, Technical Monitor: Dr. Jaimie Tiley). The support is gratefully acknowledged.

References

References

[1] A. Saltelli, S. Tarantola, K.-S. Chan, A quantitative model-independent method for global sensitivity analysis of model output, Technometrics 41 (1) (1999) 39–56.

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

[3] E. Borgonovo, E. Plischke, Sensitivity analysis: a review of recent advances, European Journal of Operational Research 248 (3) (2016) 869–887.

[4] T. Homma, A. Saltelli, Importance measures in global sensitivity analysis of nonlinear models, Reliability Engineering & System Safety 52 (1) (1996) 1–17.

[5] Z. Hu, S. Mahadevan, Global sensitivity analysis-enhanced surrogate (gsas) modeling for reliability analysis, Structural and Multidisciplinary Optimization 53 (3) (2016) 501–521.

[6] S. Sankararaman, K. McLemore, S. Mahadevan, S. C. Bradford, L. D. Peterson, Test resource allocation in hierarchical systems using bayesian networks, AIAA journal 51 (3) (2013) 537–550.

[7] A. Saltelli, R. Bolado, An alternative way to compute Fourier amplitude sensitivity test (FAST), Computational Statistics & Data Analysis 26 (4) (1998) 445–460.

[8] G. J. McRae, J. W. Tilden, J. H. Seinfeld, Global sensitivity analysisa computational implementation of the Fourier amplitude sensitivity test (FAST), Computers & Chemical Engineering 6 (1) (1982) 15–25.

[9] C. Xu, G. Gertner, Extending a global sensitivity analysis technique to models with correlated parameters, Computational Statistics & Data Analysis 51 (12) (2007) 5579–5590.

[10] D. Lewandowski, R. M. Cooke, R. J. D. Tebbens, Sample-based estimation of correlation ratio with polynomial approximation, ACM Transactions on Modeling and Computer Simulation (TOMACS) 18 (1) (2007) 3.

[11] H. Liu, W. Chen, A. Sudjianto, Probabilistic sensitivity analysis methods for design under uncertainty, in: 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, 2004, p. 4589.

[12] S. Da Veiga, Global sensitivity analysis with dependence measures, Journal of Statistical Computation and Simulation 85 (7) (2015) 1283–1305.

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

[14] J. Nossent, P. Elsen, W. Bauwens, Sobol sensitivity analysis of a complex environmental model, Environmental Modelling & Software 26 (12) (2011) 1515–1525.

[15] C. Zhang, J. Chu, G. Fu, Sobol sensitivity analysis for a distributed hydrological model of yichun river basin, china, Journal of hydrology 480 (2013) 58–68.

[16] Z. Hu, S. Mahadevan, Uncertainty quantification in prediction of material properties during additive manufacturing, Scripta Materialia 135 (2017) 135–140.

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

[18] B. Iooss, P. Lemaître, A review on global sensitivity analysis methods, in: Uncertainty management in simulation-optimization of complex systems, Springer, 2015, pp. 101–122.

[19] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, S. Tarantola, Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index, Computer Physics Communications 181 (2) (2010) 259–270.

[20] C. Prieur, S. Tarantola, Variance-based sensitivity analysis: Theory and estimation algorithms, Handbook of Uncertainty Quantification (2016) 1–23.

[21] M. J. Jansen, Analysis of variance designs for model output, Computer Physics Communications 117 (1-2) (1999) 35–43.

[22] A. Janon, T. Klein, A. Lagnoux, M. Nodet, C. Prieur, Asymptotic normality and efficiency of two Sobol index estimators, ESAIM: Probability and Statistics 18 (2014) 342–364.

[23] A. Saltelli, I. M. Sobol', Sensitivity analysis for nonlinear mathematical models: numerical experience, Matematicheskoe Modelirovanie 7 (11) (1995) 16–28.

[24] I. M. Sobol, Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Mathematics and computers in simulation 55 (1) (2001) 271–280.

[25] G. Glen, K. Isaacs, Estimating Sobol sensitivity indices using correlations, Environmental Modelling & Software 37 (2012) 157–166.

[26] J.-Y. Tissot, C. Prieur, Variance-based sensitivity analysis using harmonic analysis, working paper or preprint (Mar. 2012).
URL https://hal.archives-ouvertes.fr/hal-00680725

[27] X. Wang, K.-T. Fang, The effective dimension and quasi-Monte Carlo integration, Journal of Complexity 19 (2) (2003) 101–124.

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

[29] 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.

[30] S. Tarantola, D. Gatelli, T. A. Mara, Random balance designs for the estimation of first order global sensitivity indices, Reliability Engineering & System Safety 91 (6) (2006) 717–727.

[31] L. L. Gratiet, S. Marelli, B. Sudret, Metamodel-based sensitivity analysis: polynomial chaos expansions and Gaussian processes, Handbook of Uncertainty Quantification (2016) 1–37.

[32] A. Marrel, B. Iooss, S. Da Veiga, M. Ribatet, Global sensitivity analysis of stochastic computer models with joint metamodels, Statistics and Computing 22 (3) (2012) 833–847.

[33] D. Xiu, G. E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM journal on scientific computing 24 (2) (2002) 619–644.

[34] A. Janon, M. Nodet, C. Prieur, Uncertainties assessment in global sensitivity indices estimation from metamodels, International Journal for Uncertainty Quantification 4 (1).

[35] Z. Hu, X. Du, Mixed efficient global optimization for time-dependent reliability analysis, Journal of Mechanical Design 137 (5) (2015) 051401.

[36] W. Chen, R. Jin, A. Sudjianto, Analytical variance-based global sensitivity analysis in simulation-based design under uncertainty, Journal of mechanical design 127 (5) (2005) 875–886.

[37] L. Le Gratiet, C. Cannamela, B. Iooss, A bayesian approach for global sensitivity analysis of (multifidelity) computer codes, SIAM/ASA Journal on Uncertainty Quantification 2 (1) (2014) 336–363.

[38] C. Li, S. Mahadevan, An efficient modularized sample-based method to estimate the first-order Sobol index, Reliability Engineering & System Safety 153 (2016) 110–121.

[39] S. Da Veiga, F. Wahl, F. Gamboa, Local polynomial estimation for sensitivity analysis on models with correlated inputs, Technometrics 51 (4) (2009) 452–463.

[40] M. Eldred, C. Webster, P. Constantine, Evaluation of non-intrusive approaches for Wiener-Askey generalized polynomial chaos, in: 49th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 16th AIAA/ASME/AHS Adaptive Structures Conference, 10th AIAA Non-Deterministic Approaches Conference, 9th AIAA Gossamer Spacecraft Forum, 4th AIAA Multidisciplinary Design Optimization Specialists Conference, 2008, p. 1892.

[41] G. Jia, A. A. Taflanidis, Efficient evaluation of Sobol indices utilizing samples from an auxiliary probability density function, Journal of Engineering Mechanics 142 (5) (2016) 04016012.

[42] D. Sparkman, H. R. Millwater, J. Garza, B. P. Smarslok, Importance sampling-based post-processing method for global sensitivity analysis, in: 18th AIAA Non-Deterministic Approaches Conference, 2016, p. 1440.

[43] C. B. Storlie, J. C. Helton, Multiple predictor smoothing methods for sensitivity analysis: Description of techniques, Reliability Engineering & System Safety 93 (1) (2008) 28–54.

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

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

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

[47] 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.

[48] Z. Hu, S. Mahadevan, Time-dependent reliability analysis using a vine-arma load model, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering 3 (1) (2017) 011007.

[49] T. Bedford, R. M. Cooke, Probability density decomposition for conditionally dependent ran-

dom variables modeled by vines, Annals of Mathematics and Artificial intelligence 32 (1) (2001) 245–268.

[50] T. Bedford, R. M. Cooke, Vines: A new graphical model for dependent random variables, Annals of Statistics (2002) 1031–1068.

[51] P. Embrechts, F. Lindskog, A. McNeil, Modelling dependence with copulas, Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich.

[52] J. C. Rodriguez, Measuring financial contagion: A copula approach, Journal of empirical finance 14 (3) (2007) 401–423.

[53] P. Xue-Kun Song, Multivariate dispersion models generated from Gaussian copula, Scandinavian Journal of Statistics 27 (2) (2000) 305–320.

[54] MATLAB user's guide, The mathworks, Inc., Natick, MA 5 (1998) 333.

[55] S. J. Sheather, M. C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, Journal of the Royal Statistical Society. Series B (Methodological) (1991) 683–690.

[56] Z. Zivkovic, Improved adaptive Gaussian mixture model for background subtraction, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, Vol. 2, IEEE, 2004, pp. 28–31.

[57] C. E. Rasmussen, The infinite Gaussian mixture model, in: Advances in neural information processing systems, 2000, pp. 554–560.

[58] H. Greenspan, A. Ruf, J. Goldberger, Constrained Gaussian mixture model framework for automatic segmentation of MR brain images, IEEE transactions on medical imaging 25 (9) (2006) 1233–1245.

[59] T. K. Moon, The expectation-maximization algorithm, IEEE Signal processing magazine 13 (6) (1996) 47–60.

[60] L. Li, R. J. Hansman, R. Palacios, R. Welsch, Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring, Transportation Research Part C: Emerging Technologies 64 (2016) 45–57.

[61] J. Yang, X. Liao, X. Yuan, P. Llull, D. J. Brady, G. Sapiro, L. Carin, Compressive sensing by learning a Gaussian mixture model from measurements, IEEE Transactions on Image Processing 24 (1) (2015) 106–119.

[62] S. Nanty, C. Helbert, A. Marrel, N. Pérot, C. Prieur, Sampling, metamodeling, and sensitivity analysis of numerical simulators with functional stochastic inputs, SIAM/ASA Journal on Uncertainty Quantification 4 (1) (2016) 636–659.

[63] S. Nanty, C. Helbert, A. Marrel, N. Pérot, C. Prieur, Uncertainty quantification for functional dependent random variables, Computational Statistics 32 (2) (2017) 559–583.

[64] K. Yamaoka, T. Nakagawa, T. Uno, Application of akaike's information criterion (aic) in the evaluation of linear pharmacokinetic equations, Journal of Pharmacokinetics and Pharmacodynamics 6 (2) (1978) 165–175.

[65] S. Mahadevan, Probability, reliability, and statistical methods in engineering design, Wiley, 2000.

[66] A. Sklar, Random variables, joint distribution functions, and copulas, Kybernetika 9 (6) (1973) 449–460.

[67] E. Keogh, A. Mueen, Curse of dimensionality, in: Encyclopedia of machine learning, Springer, 2011, pp. 257–258.

[68] Z. Hu, S. Mahadevan, Bayesian network learning for data-driven design, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering 4 (4) (2018) 041002.

[69] Z. Hu, S. Mahadevan, X. Du, Uncertainty quantification of time-dependent reliability analysis in the presence of parametric uncertainty, ASCE-ASME Journal of risk and uncertainty in Engineering systems, Part B: Mechanical Engineering 2 (3) (2016) 031005.

[70] I. M. Sobol, Theorems and examples on high dimensional model representation, Reliability Engineering & System Safety 79 (2) (2003) 187–193.

[71] J. Hutchinson, Shear coefficients for timoshenko beam theory, ASME Journal of Applied Mechanics 68 (1) (2001) 87–92.