Reliability Engineering and System Safety 162 (2017) 28–39

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Extending Morris method for qualitative global sensitivity analysis of models with dependent inputs

Qiao Ge*, Monica Menendez

Institute for Transport Planning and Systems, ETH Zurich, HIL F 37.3, Stefano-Franscini-Platz 5, 8093 Zurich, Switzerland

ARTICLE INFO

Keywords:
Dependent inputs
Screening
Global sensitivity analysis
Dependent contributions
Independent contributions
Variance-based sensitivity indexes

ABSTRACT

Global Sensitivity Analysis (GSA) can help modelers to better understand the model and manage the uncertainty. However, when the model itself is rather sophisticated, especially when dependence exists among model inputs, it could be difficult or even unfeasible to perform quantitative GSA directly. In this paper, a non-parametric approach is proposed for screening model inputs. It extends the classic Elementary Effects (i.e., Morris) method, which is widely used for screening independent inputs, to enable the screening of dependent model inputs. The performance of the proposed method is tested with three numerical experiments, and the results are cross-compared with those from the variance-based GSA.

It is found that the proposed method can properly identify the influential and non-influential inputs from a complex model with several independent and dependent inputs. Furthermore, compared with the variance-based GSA, the proposed screening method only needs a few model runs, while the screening accuracy is well maintained. Therefore, it can be regarded as a practical tool for the initial GSA of high dimensional and computationally expensive models with dependent inputs.

1. Introduction

Along with the continuous development of computational techniques, and the increasing power of computers, simulation models are more and more advanced and powerful nowadays, and they have become one essential resource for scientific research and practical applications. However, the complexity of simulation models has also significantly increased, especially due to the fact that they now often include a large number of parameters. To aid modelers to better understand the model, and to manage the uncertainties in model computation, it is necessary to further investigate the relationship between model inputs and output(s), especially when the model is extremely sophisticated and/or treated like a black box. One commonly used tool for such task is Global Sensitivity Analysis (GSA) [1].

The sensitivity information can be obtained through performing GSA in the input space, and analyzing the impacts of variations of inputs on the variations of output(s). Such information can be qualitative (e.g., the sets of influential and non-influential inputs) or quantitative (e.g., the variance contributions of model inputs to the total variance of model output(s)). They can be used for e.g., identifying the key inputs, reducing the uncertainties of model response, setting priorities for model calibration. Due to its importance, several GSA approaches have been extensively developed in the past few decades. In general, these approaches can be classified according to the sensitivity indexes to be assessed: derivative-based sensitivity measure [2–5], regression-based sensitivity measure [6,7], qualitative (or screening) sensitivity measure [8–12], variance-based sensitivity measure [13–17], and moment independent sensitivity measure [18–20]. More information regarding the different GSA approaches can be found in [1,21].

All these sensitivity measures, however, are based on the independence assumption of the model inputs. In practice, due to certain constrains in the input space, and/or the intricate relations among the inputs obtained from empirical experiments (e.g., some inputs are the outcome of another model or experiment) [22], it is very likely that the model inputs are actually dependent inputs, or mixtures of both independent and dependent inputs. In these cases, simply assuming that all model inputs are independent, and directly applying the aforementioned GSA methods can be fallacious, and consequently can lead to incorrect inferences (see e.g., the numerical experiments in Section 4).

To the authors' knowledge, only a few recent studies that are able to perform GSA of models with dependent inputs can be found in the literature. For example, studies such as [23–27] discussed the variance decomposition with parametric or non-parametric methods for models with dependent inputs, while [28–30] extended the analytical formula-

* Corresponding author.
E-mail addresses: qiao.ge@ivt.baug.ethz.ch (Q. Ge), monica.menendez@ivt.baug.ethz.ch (M. Menendez).

http://dx.doi.org/10.1016/j.ress.2017.01.010
Received 22 December 2015; Received in revised form 8 December 2016; Accepted 13 January 2017
Available online 15 January 2017

tions and the corresponding numerical estimators to compute the extended Sobol' sensitivity indexes [13] for dependent inputs. All these methods are quantitative, i.e., they quantitatively decompose and analyze the variance contributions of dependent model inputs. Yet, for complex models in practice, the application of quantitative GSA might still be difficult or unfeasible when the model itself contains many inputs (i.e., high dimensional) and/or is computationally expensive.

Therefore, in this paper we recall the application of inputs screening before performing any quantitative GSA (see [21] for more details). The main goal is to identify the least influential inputs, so that they can be fixed at their nominal values without significantly influencing the model output. This approach is also known as the *Factor Fixing (FF) setting* [31,1]. It is found in [21,32,33] that for high dimensional and computationally expensive models with independent inputs, the inputs screening can reduce the complexity of the model, and hence enhance the efficiency of the quantitative GSA without losing accuracy. Therefore, it is expected that a proper screening method may also provide such benefits for models with dependent inputs.

The objective of this paper is thus to develop an efficient, non-parametric approach for screening dependent model inputs. The performance of the proposed screening method is evaluated through numerical experiments. The screening results from each experiment are cross-compared with the reference results obtained using the variance-based GSA [29,30]. This analysis shows that the proposed screening approach can efficiently identify the non-influential inputs with satisfactory accuracy at a low computational cost.

The paper is organized as follows. A brief review of the screening method, specifically, the Elementary Effects (EE) method and its recent extensions, as well as the variance decomposition for dependent inputs is given in Section 2. The methodology for the proposed screening method is presented in Section 3. The details about the numerical experiments, and the corresponding results are introduced and discussed in Section 4. Conclusions are given in Section 5.

2. Literature review

2.1. Elementary effects method

The Elementary Effects (EE) method, which is also known as the Morris method, was first introduced by Morris in [8]. Let f be a model with k independent inputs, i.e., $\mathbf{X} = \{X_1, X_2, \ldots, X_k\}$, which are defined in the k-dimensional input space Ω^k ($\Omega^k \subset \mathbb{R}^k$). Let Y be the model output, i.e., $Y = f(\mathbf{X})$. Moreover, let $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}$ be the values assigned to $\mathbf{X}$ in Ω^k, i.e., $\mathbf{x} \in \Omega^k$. If only x_i (i.e., the value of input X_i, $i \in [1, k]$) is varied by a given value Δ, while the values of all other inputs remain unchanged, the corresponding model output is $f(x_1, \ldots, x_{i-1}, x_i + \Delta, x_{i+1}, \ldots, x_k)$, with $\{x_1, \ldots, x_{i-1}, x_i + \Delta, x_{i+1}, \ldots, x_k\} \in \Omega^k$. The formula for computing the EE of X_i (i.e., EE_i) is given below:

$$EE_i = \frac{f(x_1, \ldots, x_{i-1}, x_i + \Delta, x_{i+1}, \ldots, x_k) - f(\mathbf{x})}{\Delta}. \tag{1}$$

The above definition employs the One-At-a-Time (OAT) design. To investigate the global sensitivity of X_i, the OAT experiment is repeated using N different values of $\mathbf{x}$ that are randomly sampled in Ω^k. Accordingly, N EEs can be obtained for input X_i. In [8], the mean (i.e., μ_i) and the standard deviation (i.e., σ_i) of these N EEs are used as the screening measures:

$$\mu_i = \frac{1}{N}\sum_{r=1}^{N} EE_{i,r}, \tag{2}$$

$$\sigma_i = \sqrt{\frac{1}{N-1}\sum_{r=1}^{N} (EE_{i,r} - \mu_i)^2}, \tag{3}$$

where $EE_{i,r}$ corresponds to the r-th EE of X_i.

In [9], the absolute mean (i.e., μ_i^*) is proposed to replace μ_i:

$$\mu_i^* = \frac{1}{N}\sum_{r=1}^{N} |EE_{i,r}|, \tag{4}$$

where $|\cdot|$ stands for the absolute value. As discussed in [9], when the model contains many interactive inputs and/or the model is not monotonic, using μ_i^* as a screening measure can significantly reduce the Type II error (i.e., considering an influential input as non-influential [1]). According to [8,9], an input X_i is identified as:

a) a non-influential input, if μ_i^* is close to zero;
b) an influential input with negligible non-linear effects, if μ_i^* is high but σ_i is low; or
c) an influential input with non-linear effects and/or strong interactions with other inputs, if both μ_i^* and σ_i are high.

It was found in [2,3] that for non-monotonic models, the EE method is less accurate than the Monte Carlo/Quasi-Monte Carlo integration method in estimating the derivative-based sensitivity measures [2,4]. However, the same studies [2,3] also emphasized that such high accuracy of estimates may not be required for inputs screening, where the aim is to identify influential and non-influential inputs with a low computational cost. Thus, the EE method can still be considered as a good compromise between accuracy and efficiency, especially for the SA of high dimensional and computationally expensive models.

It is worth mentioning that although the EE method was developed for screening purposes [8], this approach can also be used for ranking the inputs in order of importance [34,35,3]. In particular, Saltelli et al. [35] recommended to use μ^* for inputs ranking when the SA is FF setting. Moreover, the empirical study in [9] showed that using μ^* for ranking independent model inputs could achieve similar results as using the Sobol' total sensitivity index (see Section 2.2.1). Due to its high efficiency, the EE method has been successfully employed for ranking independent inputs by many researchers from different disciplines (e.g., [9,36–38,33,39,40]).

To enhance its computational efficiency, the classic Morris EE method has been extended by adopting different sampling designs, for example, the trajectory design[1] [8,9,37,11], the cell design [10], and the radial design [42]. The trajectory design is the most commonly used sampling design for computing EE, while the radial design shows the best performance (see the experiments performed in [42] for more details). These two designs are therefore implemented in the screening approach proposed in this paper, and will be introduced in Section 3.2.1.

When the model contains dependent inputs, the classic EE method's biggest drawback is that it does not account for the impacts from inputs dependence. For example, when an input is changed by Δ, any other correlated input should simultaneously show some variation. However, such variation is are not included in Eq. (1). Hence, when dependence exists among model inputs, the application of the classic EE method could yield incorrect screening results. To take the effects of inputs dependence into account, it is necessary to recall the variance decomposition approach from the variance-based GSA. The corresponding information is presented in the following section.

2.2. Variance decomposition and sensitivity indexes

2.2.1. Model with independent inputs

We consider a square integrable function f with k independent inputs $\mathbf{X} = \{X_1, X_2, \ldots, X_k\}$ defined in $\mathbb{R}^k$. According to [13], f can be

[1] The trajectory design is also known as winding stairs in [41]. The difference is the trajectory design always produces random trajectories separately, while the winding stairs design joins all random trajectories together.

expanded as:

$$f(\mathbf{X}) = f_0 + \sum_{i=1}^{k} f_i + \sum_{i=1}^{k}\sum_{j=i+1}^{k} f_{ij} + \cdots + f_{1,\ldots,k}. \tag{5}$$

where $f_i = f_i(X_i)$, $f_{ij} = f_{ij}(X_i, X_j)$, and so on for the higher dimensional terms. The above expansion is known as *High Dimensional Model Representation* (HDMR) [1].

According to Sobol' [13], the HDMR is unique when all model inputs are independent and all terms in Eq. (5) have zero mean. The decomposition in this unique case is called *ANalysis Of VAriances* (ANOVA) decomposition. Sobol' proved in [13] that the terms in the ANOVA decomposition are pairwise orthogonal, and the variance of the model output, i.e, $V(Y)$, can be decomposed as below:

$$V(Y) = \sum_{i=1}^{k} V_i + \sum_{i=1}^{k}\sum_{j=i+1}^{k} V_{ij} + \cdots + V_{1,\ldots,k}, \tag{6}$$

where

$$V_i = V(f_i) = V(E(Y|X_i)), \tag{7}$$

$$V_{ij} = V(f_{ij}) = V(E(Y|X_i, X_j)) - V_i - V_j, \tag{8}$$

and so on for the variances at higher orders. Here $E(\cdot)$ and $V(\cdot)$ stand respectively for the expectation and variance, and $E(\cdot|\cdot)$ is the conditional expectation. V_i is the first order or main variance (i.e., the variance contributed by X_i alone), and V_{ij} is the second order variance (i.e., the variance due to the interaction between X_i and X_j). The total variance [43] with respect to X_i is calculated as:

$$V_i^T = V_i + \sum_{j\neq i}^{k} V_{ij} + \cdots + V_{1,\ldots,k} = E(V(Y|\mathbf{X}_{\sim i})) = V(Y) - V(E(Y|\mathbf{X}_{\sim i})), \tag{9}$$

where $V(\cdot|\cdot)$ stands for the conditional variance. Obviously, V_i^T contains V_i and all higher order variances (i.e., the variance caused by inputs interaction) contributed by X_i.

It is assumed that each input X_i has a continuous Probability Density Function (PDF), i.e., $p(X_i)$. Let $p(\mathbf{X})$ be the joint PDF of $\mathbf{X}$, i.e., $p(\mathbf{X}) = p(X_1, \ldots, X_k)$. The expectation and variance of the model output Y are:

$$E(Y) = \int_{\mathbb{R}^k} f(\mathbf{X}) p(\mathbf{X}) \mathrm{d}\mathbf{X} = f_0, \tag{10}$$

$$V(Y) = \int_{\mathbb{R}^k} f^2(\mathbf{X}) p(\mathbf{X}) \mathrm{d}\mathbf{X} - f_0^2. \tag{11}$$

The analytical formulas for computing V_i and V_i^T are given below [43,29]:

$$V_i = \int_{\mathbb{R}^k} f(\mathbf{X}') p(\mathbf{X}') \mathrm{d}\mathbf{X}' \left[\int_{\mathbb{R}^{k-1}} f(X'_i, \mathbf{X}_{\sim i}) p(\mathbf{X}_{\sim i}) \mathrm{d}\mathbf{X}_{\sim i} - \int_{\mathbb{R}^k} f(\mathbf{X}) p(\mathbf{X}) \mathrm{d}\mathbf{X} \right], \tag{12}$$

$$V_i^T = \frac{1}{2} \int_{\mathbb{R}^{k+1}} [f(X_i, \mathbf{X}_{\sim i}) - f(X'_i, \mathbf{X}_{\sim i})]^2 p(\mathbf{X}) p(X'_i) \mathrm{d}\mathbf{X} \mathrm{d}X'_i. \tag{13}$$

In Eqs. (12) and (13), $\mathbf{X}' = \{X'_1, X'_2, \ldots, X'_k\}$. X'_i ($i \in [1, k]$) is an independent random input generated from the marginal PDF $p(X_i)$ (note that X'_i is different from X_i). $\mathbf{X}_{\sim i}$ contains all model inputs except X_i, i.e., $\mathbf{X} = \{X_i, X_{\sim i}\}$. $p(\mathbf{X}_{\sim i})$ is the joint PDF of $\mathbf{X}_{\sim i}$. As all inputs in $\mathbf{X}_{\sim i}$ are independent with each other, $p(\mathbf{X}_{\sim i}) = p(\mathbf{X})/p(X_i) = \prod_{j=1, j\neq i}^{k} p(X_j)$. By the same token, $p(\mathbf{X}) = \prod_{j=1}^{k} p(X_j)$ and $p(\mathbf{X}') = \prod_{j=1}^{k} p(X'_j)$.

Two sensitivity indexes can be obtained by normalizing V_i and V_i^T with $V(Y)$:

$$S_i = \frac{1}{V(Y)} \left[\int_{\mathbb{R}^k} f(\mathbf{X}') p(\mathbf{X}') \mathrm{d}\mathbf{X}' \left[\int_{\mathbb{R}^{k-1}} f(X'_i, \mathbf{X}_{\sim i}) p(\mathbf{X}_{\sim i}) \mathrm{d}\mathbf{X}_{\sim i} - \int_{\mathbb{R}^k} f(\mathbf{X}) p(\mathbf{X}) \mathrm{d}\mathbf{X} \right] \right], \tag{14}$$

$$S_i^T = \frac{1}{2V(Y)} \left[\int_{\mathbb{R}^{k+1}} [f(X_i, \mathbf{X}_{\sim i}) - f(X'_i, \mathbf{X}_{\sim i})]^2 p(\mathbf{X}) p(X'_i) \mathrm{d}\mathbf{X} \mathrm{d}X'_i \right]. \tag{15}$$

S_i is the first order sensitivity index, and S_i^T is the total sensitivity index. For a model with independent inputs, it is obvious that S_i and S_i^T should always be between 0 and 1. S_i, which measures the main variance contribution of X_i, should never exceed S_i^T (i.e., $S_i \leq S_i^T$, the equality only holds when X_i has no interaction with any other input). A high S_i means X_i must be an influential input. However, a low S_i does not necessarily mean that X_i is not important, as it may have high interaction effects. S_i^T measures the total variance contribution of X_i. Therefore, a low S_i^T (e.g., $S_i^T \cong 0$) indicates that X_i is a non-influential input, and it can be fixed without influencing the variance of the model output.

Moreover, for a model with independent inputs, the ANOVA decomposition and the two sensitivity indexes mentioned above can be used to determine the model structure [1,28]: 1) if $\sum_{i=1}^{k} S_i < 1$ and $\sum_{i=1}^{k} S_i^T > 1$, the model is a non-additive model and the interaction effects among model inputs are not trivial; and 2) if $\sum_{i=1}^{k} S_i = 1$ and $\sum_{i=1}^{k} S_i^T = 1$, the model is a purely additive model. Notice, however, that these two inferences do not necessarily hold when the model contains dependent inputs [6].

2.2.2. Model with dependent inputs

When the model contains dependent inputs, the ANOVA decomposition presented in Eq. (6) can not be simply applied without considering the joint distributions of model inputs [24]. In fact, according to [27,29,30], the total variance contributed by an input includes three parts: the main variance, the variance caused by interactions, and the variance caused by dependence. Several studies have discussed the variance decomposition and the corresponding sensitivity indexes for model with dependent inputs (see e.g., [23,25,29,30,27]). Here we recall the findings presented in [29,30], in which the non-parametric methods are used to decompose the variances and compute the sensitivity indexes.

Let $\mathbf{X}_s$ be a subset of $\mathbf{X}$ (i.e., $\mathbf{X}_s = \{X_{i_1}, \ldots, X_{i_s}\}$, $1 \leq i_1 < \cdots < i_s \leq k$), and $\mathbf{X}_{\sim s}$ be the complement set. In [29], Kucherenko et al. introduced two types of variances:

1. V_s^{full}: the dependent main variance of $\mathbf{X}_s$, which includes the main variance of $\mathbf{X}_s$, and the variance caused by the dependence of $\mathbf{X}_s$ with $\mathbf{X}_{\sim s}$, but excludes the variance due to the interaction of $\mathbf{X}_s$ with $\mathbf{X}_{\sim s}$; and
2. $V_s^{T,ind}$: the independent total variance of $\mathbf{X}_s$, which includes the main variance of $\mathbf{X}_s$, and the variance due to the interaction of $\mathbf{X}_s$ with $\mathbf{X}_{\sim s}$, but excludes the variance caused by the dependence of $\mathbf{X}_s$ with $\mathbf{X}_{\sim s}$.

In a recent study [30], Mara et al. proposed another two types of variances:

1. V_s^{ind}: the independent main variance of $\mathbf{X}_s$, which only includes the main variance of $\mathbf{X}_s$, but excludes the variance caused by the interaction and dependence of $\mathbf{X}_s$ with $\mathbf{X}_{\sim s}$; and
2. $V_s^{T,full}$: the dependent total variance of $\mathbf{X}_s$, which includes the main variance of $\mathbf{X}_s$, and the variance caused by the interaction and dependence of $\mathbf{X}_s$ with $\mathbf{X}_{\sim s}$.

According to the law of total variance, $V(Y) = V_s^{full} + V_{\sim s}^{T,ind} = V_s^{T,ind} + V_{\sim s}^{full} = V_s^{ind} + V_{\sim s}^{T,full} = V_s^{T,full} + V_{\sim s}^{ind}$. If we let $\mathbf{X}_s = \{X_i\}$, the equations below can then be used to compute

the above four variances of any model input X_i analytically [29,30]:

$$V_i^{ind} = \int_{\mathbb{R}^k} f(\mathbf{X}')p(\mathbf{X}')\mathrm{d}\mathbf{X}'\left[\int_{\mathbb{R}^{k-1}} f(\overline{X}'_i, \mathbf{X}_{\sim i})p(\mathbf{X}_{\sim i})\mathrm{d}\mathbf{X}_{\sim i} - \int_{\mathbb{R}^k} f(\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X}\right], \tag{16}$$

$$V_i^{full} = \int_{\mathbb{R}^k} f(\mathbf{X}')p(\mathbf{X}')\mathrm{d}\mathbf{X}'\left[\int_{\mathbb{R}^{k-1}} f(X'_i, \overline{\mathbf{X}}_{\sim i})p(\overline{\mathbf{X}}_{\sim i}|X'_i)\mathrm{d}\overline{\mathbf{X}}_{\sim i} - \int_{\mathbb{R}^k} f(\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X}\right], \tag{17}$$

$$V_i^{T,ind} = \frac{1}{2}\int_{\mathbb{R}^{k+1}} [f(X_i, \mathbf{X}_{\sim i}) - f(\overline{X}'_i, \mathbf{X}_{\sim i})]^2 p(\mathbf{X})p(\overline{X}'_i|\mathbf{X}_{\sim i})\mathrm{d}\mathbf{X}\mathrm{d}\overline{X}'_i, \tag{18}$$

$$V_i^{T,full} = \frac{1}{2}\int_{\mathbb{R}^{k+1}} [f(X_i, \mathbf{X}_{\sim i}) - f(X'_i, \overline{\mathbf{X}}_{\sim i})]^2 p(\mathbf{X})p(X'_i)\mathrm{d}\mathbf{X}\mathrm{d}X'_i. \tag{19}$$

In Eqs. (16) to (19), $\mathbf{X}' = \{X'_1, X'_2, \ldots, X'_k\}$ is a set of k dependent random inputs whose joint PDF is $p(\mathbf{X})$. The random inputs $\overline{X}'_{\sim i}$ and $\overline{\mathbf{X}}_{\sim i}$ follow the conditional PDFs $p(\overline{X}'_{\sim i}|\mathbf{X}_{\sim i})$ and $p(\overline{\mathbf{X}}_{\sim i}|X'_i)$, respectively (more details can be found in [29,30]).

These variances can be further normalized with $V(Y)$, and accordingly four sensitivity indexes can be obtained:

$$S_i^{ind} = \frac{1}{V(Y)}\left[\int_{\mathbb{R}^k} f(\mathbf{X}')p(\mathbf{X}')\mathrm{d}\mathbf{X}'\left[\int_{\mathbb{R}^{k-1}} f(\overline{X}'_i, \mathbf{X}_{\sim i})p(\mathbf{X}_{\sim i})\mathrm{d}\mathbf{X}_{\sim i} - \int_{\mathbb{R}^k} f(\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X}\right]\right], \tag{20}$$

$$S_i^{full} = \frac{1}{V(Y)}\left[\int_{\mathbb{R}^k} f(\mathbf{X}')p(\mathbf{X}')\mathrm{d}\mathbf{X}'\left[\int_{\mathbb{R}^{k-1}} f(X'_i, \overline{\mathbf{X}}_{\sim i})p(\overline{\mathbf{X}}_{\sim i}|X'_i)\mathrm{d}\overline{\mathbf{X}}_{\sim i} - \int_{\mathbb{R}^k} f(\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X}\right]\right], \tag{21}$$

$$S_i^{T,ind} = \frac{1}{2V(Y)}\left[\int_{\mathbb{R}^{k+1}} [f(X_i, \mathbf{X}_{\sim i}) - f(\overline{X}'_i, \mathbf{X}_{\sim i})]^2 p(\mathbf{X})p(\overline{X}'_i|\mathbf{X}_{\sim i})\mathrm{d}\mathbf{X}\mathrm{d}\overline{X}'_i\right], \tag{22}$$

$$S_i^{T,full} = \frac{1}{2V(Y)}\left[\int_{\mathbb{R}^{k+1}} [f(X_i, \mathbf{X}_{\sim i}) - f(X'_i, \overline{\mathbf{X}}_{\sim i})]^2 p(\mathbf{X})p(X'_i)\mathrm{d}\mathbf{X}\mathrm{d}X'_i\right]. \tag{23}$$

S_i^{ind} and $S_i^{T,ind}$ are called *independent sensitivity indexes* [30], and S_i^{full} and $S_i^{T,full}$ are know as *full sensitivity indexes*[2][28]. Similar to Sobol' sensitivity indexes for independent inputs, the inequalities $S_i^{ind} \le S_i^{T,ind}$ and $S_i^{full} \le S_i^{T,full}$ also hold here for dependent inputs. However, it is not necessary to have $S_i^{ind} \le S_i^{T,full}$, $S_i^{full} \le S_i^{T,ind}$ or $S_i^{T,ind} \le S_i^{T,full}$, because the effects of inputs dependence can be either positive or negative (see the example in [29]).

The analytical formulas in Eqs. (20) to (23) require the computation of multi-dimensional integrals, which can be rather difficult or even unfeasible when the model f and/or the conditional PDFs are complex functions. In practice, the following Monte Carlo (MC) estimators can be used [29,30]:

$$\widehat{V}(Y) = \frac{1}{N}\sum_{r=1}^{N} [f(x_{i,r}, \mathbf{x}_{\sim i,r})]^2 - \left[\frac{1}{N}\sum_{r=1}^{N} f(x_{i,r}, \mathbf{x}_{\sim i,r})\right]^2, \tag{24}$$

$$\widehat{S}_i^{ind} = \frac{1}{N\cdot\widehat{V}(Y)}\sum_{r=1}^{N} f(x_{i,r'}, \mathbf{x}'_{\sim i,r})\cdot\left[f(\overline{x}'_{i,r}, \mathbf{x}_{\sim i,r}) - f(x_{i,r}, \mathbf{x}_{\sim i,r})\right], \tag{25}$$

$$\widehat{S}_i^{full} = \frac{1}{N\cdot\widehat{V}(Y)}\sum_{r=1}^{N} f(x_{i,r'}, \mathbf{x}'_{\sim i,r})\cdot\left[f(x'_{i,r}, \overline{\mathbf{x}}_{\sim i,r}) - f(x_{i,r}, \mathbf{x}_{\sim i,r})\right], \tag{26}$$

$$\widehat{S}_i^{T,ind} = \frac{1}{2N\cdot\widehat{V}(Y)}\sum_{r=1}^{N} \left[f(x_{i,r}, \mathbf{x}_{\sim i,r}) - f(\overline{x}'_{i,r}, \mathbf{x}_{\sim i,r})\right]^2, \tag{27}$$

$$\widehat{S}_i^{T,full} = \frac{1}{2N\cdot\widehat{V}(Y)}\sum_{r=1}^{N} \left[f(x_{i,r}, \mathbf{x}_{\sim i,r}) - f(x'_{i,r}, \overline{\mathbf{x}}_{\sim i,r})\right]^2, \tag{28}$$

where N is the number of MC simulations. $x_{i,r}$, $\mathbf{x}_{\sim i,r}$, $x'_{i,r}$, $\mathbf{x}'_{\sim i,r}$, $\overline{x}'_{i,r}$, and $\overline{\mathbf{x}}_{\sim i,r}$ are, respectively, the r-th draw from the random samples of X_i, $\mathbf{X}_{\sim i}$, X'_i, $\mathbf{X}'_{\sim i}$, $\overline{X}'_i$ and $\overline{\mathbf{X}}_{\sim i}$.

Neither of the two main sensitivity indexes are proper measures for inputs screening under the FF setting. For instance, $S_i^{ind} \approx 0$ and $S_i^{full} \approx 0$ are necessary but not sufficient conditions for fixing X_i, because X_i could have higher order interactions with other inputs. Therefore, it is more appropriate to adopt the two total sensitivity indexes $S_i^{T,ind}$ and $S_i^{T,full}$ as the screening measures.

More specifically, when $S_i^{T,ind} \approx 0$ but $S_i^{T,full} \gg 0$, it can be inferred that the influence of X_i on the output variance is only due to its dependence with other model inputs. On the contrary, if $S_i^{T,ind} \gg 0$, it means that almost all variance contributions of X_i are from X_i alone and/or its interactions with other inputs. In both cases, X_i is identified as an influential input since it has contributions either through its dependence with other influential inputs (i.e., the former case), or by itself and/or the interactions with other inputs (i.e., the latter case). X_i can only be classified as a non-influential input when both $S_i^{T,ind}$ and $S_i^{T,full}$ are close to 0.

In Eqs. (27) and (28), the computation of $\widehat{S}_i^{T,ind}$ and $\widehat{S}_i^{T,full}$ requires N runs of $f(x_{i,r}, \mathbf{x}_{\sim i,r})$, $f(\overline{x}'_{i,r}, \mathbf{x}_{\sim i,r})$, and $f(x'_{i,r}, \overline{\mathbf{x}}_{\sim i,r})$, respectively. Thus, the total model runs required to compute the total sensitivity indexes for all k inputs is $3kN$. In general, N needs to a big number (e.g., from 10^4 to 10^5) for the MC estimator to get accurate results. Hence, when the model is computationally expensive and/or contains many inputs, the total computational cost can be rather high, and it may still be difficult to perform the MC estimation directly.

On the other hand, a very high accuracy in the initial inputs screening may not be necessary as suggested by [2,3], especially when the main purpose is to determine the non-influential inputs. Therefore, a qualitative analysis for inputs screening, which balances both efficiency and accuracy, can be very a valuable tool. In the next sections, a non-parametric method for screening dependent model inputs is proposed. The methodology is presented in Section 3, and it is tested with numerical functions in Section 4.

3. Methodology

3.1. Extended EE for dependent inputs

To screen dependent model inputs, we extend the classic EE method in a similar way as the variance decomposition approach introduced in Section 2.2.2. Two new measures for screening dependent inputs are proposed:

a) independent elementary effects EE_i^{ind}: elementary effects excluding the impacts due to the dependence between input X_i and other inputs;
b) full elementary effects EE_i^{full}: elementary effects including the impacts due to the dependence between input X_i and other inputs.

Let X_i, $\mathbf{X}_{\sim i}$, X'_i, $\overline{X}'_i$ and $\overline{\mathbf{X}}_{\sim i}$ be the inputs following the same definitions as those in Eqs. (18) and (19) (or equivalently in Eqs. (22) and (23)). Let x_i, $\mathbf{x}_{\sim i}$, x'_i, $\overline{x}'_i$, and $\overline{\mathbf{x}}_{\sim i}$ denote the values assigned to these inputs. The formulas for computing EE_i^{ind} and EE_i^{full} are given below:

[2] In [28] the full sensitivity indexes are denoted as S_i and S_i^T. In order to avoid the confusion with the sensitivity indexes for independent inputs (Eqs. (14) and (15)), S_i^{full}

$$EE_i^{ind} = \frac{f(\overline{x}'_i, \mathbf{x}_{\sim i}) - f(x_i, \mathbf{x}_{\sim i})}{\Delta}, \tag{29}$$

$$EE_i^{full} = \frac{f(x'_i, \overline{\mathbf{x}}_{\sim i}) - f(x_i, \mathbf{x}_{\sim i})}{\Delta}, \tag{30}$$

In order to evaluate the global sensitivity of X_i, the computation of EE_i^{ind} and EE_i^{full} is repeated in the entire input space using N random samples (the details about the sampling design are given in Section 3.2.1). The absolute means and standard deviations of EE_i^{ind} and EE_i^{full}, i.e., $(\mu_i^{*,\,ind}, \sigma_i^{ind})$ and $(\mu_i^{*,\,full}, \sigma_i^{full})$ can be computed and used as the screening measures. Specifically, when both $(\mu_i^{*,\,ind}, \sigma_i^{ind})$ and $(\mu_i^{*,\,full}, \sigma_i^{full})$ are very close to 0, X_i is expected to be a non-influential input. When $\mu^{*,\,ind} \gg 0$ and $\sigma_i^{ind} \approx 0$, X_i is influential because of its own contribution to the output variance. When $\mu^{*,\,ind} \approx 0$, $\sigma_i^{ind} \approx 0$, $\sigma_i^{full} \approx 0$, but $\mu^{*,\,full} \gg 0$, X_i is indirectly influential, i.e., its influence is mainly due to its dependence with other influential inputs. When $\sigma_i^{ind} \gg 0$ and/or $\sigma_i^{full} \gg 0$, X_i has strong interaction effects with other inputs.

It is worth mentioning that the formulation of the extended elementary effects in Eqs. (29) and (30) is independent of any specific dependence structure, or any specific design for producing dependent random samples. Thus, the extended EE approach is a general approach for screening both independent and dependent inputs. In other words, no matter how the inputs are associated (e.g., the inputs might be dependent but not correlated), Eqs. (29) and (30) and the above mentioned screening criteria should remain the same.

3.2. Sampling design

Dependent random samples are needed when computing EE_i^{ind} and EE_i^{full}, as well as the screening measures $(\mu_i^{*,\,ind}, \sigma_i^{ind})$. The sampling design used in this paper comprises two consecutive steps: first, generate independent random samples from uniform distributions; second, transform the independent random samples into dependent random samples. The details are given below.

3.2.1. Independent samples

As already introduced in Section 2.1, different sampling designs (e.g., trajectory design [8], radial design [17], and cell design [10]) can be used to enhance the computational efficiency of the classic EE method. The case studies in [9,11] have shown that the trajectory design, especially the Quasi-Optimized Trajectory (Quasi-OT, see [11]) design, is an effective sampling design for screening independent parameters. However, because the analysis becomes much more complex when the model contains dependent parameters, it is still unclear which sampling design has better performance in practice. Therefore, the trajectory design and the radial design are both implemented in this paper for generating independent random samples. The cell design is not considered here because it is less effective than the other two sampling designs according to the empirical study in [42]. Table 1 gives an example, i.e., the r-th draw of the independent random sample ($r \in [1, N]$, N is the sample size), from both designs.

In Table 1, a point represents a k-dimensional vector, and each element of this vector represents a possible value of the corresponding input. $\mathbf{a}_r$ (i.e., $\{a_{1,r},\ a_{2,r}, \ldots,\ a_{k,r}\}$) and $\mathbf{b}_r$ (i.e., $\{b_{1,r},\ b_{2,r}, \ldots,\ b_{k,r}\}$) are two different points taken from the k-dimensional unit hypercube $\mathbb{H}^k$, $\mathbb{H}^k = \mathcal{U}^k[0, 1]$. The r-th random sample is formed by $k + 1$ points in both designs. These points are derived by varying and combining the elements of $\mathbf{a}_r$ and $\mathbf{b}_r$.

Table 1
An Example of Trajectory Design and Radial Design.

Point	Trajectory design	Radial design
$\mathbf{p}_{1,r}$	$a_{1,r},\ a_{2,r},\ a_{3,r}, \ldots, a_{k,r}$	$a_{1,r},\ a_{2,r},\ a_{3,r}, \ldots, a_{k,r}$
$\mathbf{p}_{2,r}$	$b_{1,r},\ a_{2,r},\ a_{3,r}, \ldots, a_{k,r}$	$b_{1,r},\ a_{2,r},\ a_{3,r}, \ldots, a_{k,r}$
$\mathbf{p}_{3,r}$	$b_{1,r},\ b_{2,r},\ a_{3,r}, \ldots, a_{k,r}$	$a_{1,r},\ b_{2,r},\ a_{3,r}, \ldots, a_{k,r}$
⋮	⋮	⋮
$\mathbf{p}_{i,r}$	$b_{1,r}, \ldots, b_{i-1,r}, a_{i,r}, \ldots, a_{k,r}$	$a_{1,r}, \ldots, a_{i-2,r}, b_{i-1,r}, a_{i,r}, \ldots, a_{k,r}$
$\mathbf{p}_{i+1,r}$	$b_{1,r}, \ldots, b_{i-1,r}, b_{i,r}, \ldots, a_{k,r}$	$a_{1,r}, \ldots, a_{i-2,r}, a_{i-1,r}, b_{i,r}, \ldots, a_{k,r}$
⋮	⋮	⋮
$\mathbf{p}_{k+1,r}$	$b_{1,r},\ b_{2,r},\ b_{3,r}, \ldots, b_{k,r}$	$a_{1,r},\ a_{2,r},\ a_{3,r}, \ldots, b_{k,r}$

Specifically, for the trajectory design, there is exactly one element difference between any two consecutive points,[3] i.e., the only difference between $\mathbf{p}_{i,r}$ and $\mathbf{p}_{i+1,r}$ is the i-th element. On the other hand, for the radial design, there is exactly one element difference between the first point $\mathbf{p}_{1,r}$ and any other point. For example, $\mathbf{p}_{1,r}$ and $\mathbf{p}_{i,r}$ just differ in the $(i-1)$-th element.

The Morris sampling strategy [8] is adopted for the trajectory design in this dissertation. This means $\mathbf{a}_r$ and $\mathbf{b}_r$ are sampled from fixed grid levels (see [8] for more details). Consequently, the coordinate difference Δ is a constant for all trajectories, i.e., $b_{i,r} = a_{i,r} + \Delta$, $\forall\ i \in [1, k]$, $\forall\ r \in [1, N]$. To make an efficient exploration of the input space, the design of Quasi-OT [11] is used to generate N random trajectories that have wide dispersion in the input space.

The radial design is implemented using Sobol's quasi-random sequence [44]. Following the procedure described in [17], a N-by-$2k$ matrix of Sobol' quasi-random numbers is produced. The first k elements of the r-th row in this matrix are taken to form $\mathbf{a}_r$, and the rest k elements of the same row are used to form $\mathbf{b}_r$. Unlike the Morris sampling strategy, the coordinate difference does not have a fixed value throughout the radial sampling, i.e., Δ is not required to be a constant under the radial design.

3.2.2. Dependent samples

Two transformations, i.e., the inverse Rosenblatt transformation [45] and the inverse Nataf transformation [46–48], are considered in this paper to transform independent random samples into dependent random samples. The inverse Rosenblatt transformation requires the knowledge of conditional distributions, while the inverse Nataf transformation can be used without such knowledge. The details are introduced below.

Inverse rosenblatt transformation

Let $\mathbf{U} = \{U_1, U_2, \ldots, U_k\}$ be a set of k independent random variables uniformly distributed over $\mathbb{H}^k$. Let $\mathbf{u} = \{u_1, u_2, \ldots, u_k\}$ be a set of possible values assigned to these variables. The values of the dependent random samples of $\mathbf{X}$ can be obtained through the following inverse Rosenblatt transformation [45]:

$$\begin{cases} x_1 = F_1^{-1}(u_1) \\ x_2 = F_{2|1}^{-1}(u_2|x_1) \\ \vdots \\ x_k = F_{k|1,\ldots,k-1}^{-1}(u_k|x_1, \ldots, x_{k-1}) \end{cases} \tag{31}$$

where $F_1^{-1}(x_1)$ is the inverse Cumulative Distribution Function (CDF) of X_1, $F_{j|1,\ldots,j-1}^{-1}(x_j|x_1, x_2, \ldots, x_{j-1})$ $(j \in [2, k])$ is the inverse conditional CDF of X_j given $X_1 = x_1, \cdots, X_{j-1} = x_{j-1}$.

Note that the inverse Rosenblatt transformation is unique if and only if $\mathbf{X}$ is a set of independent inputs. For dependent inputs, this transformation has $k!$ different possibilities depending on the order of inputs in $\mathbf{X}$. For instance, if the inputs are ordered as $\{X_{i+1}, \ldots, X_k, X_1, \ldots, X_i\}$, the transformation is derived as:

[3] For the ease of demonstration, in the trajectory design shown in Table 1, the position of the element that changes between two consecutive points is arbitrarily set to follow the order $1, 2, \ldots, k$. However, this setting is not mandatory in practice. For example, $\mathbf{p}_{1,r}$ and $\mathbf{p}_{2,r}$ could differ on the 3rd element, and $\mathbf{p}_{2,r}$ and $\mathbf{p}_{3,r}$ could differ on the 1st element. The only requirement is that the coordinate variation of any input only happens once in a given trajectory.

$$\begin{cases} x_{i+1} = F_{i+1}^{-1}(u_{i+1}) \\ x_{i+2} = F_{i+2|i+1}^{-1}(u_{i+2}|x_{i+1}) \\ \vdots \\ x_k = F_{k|i+1,\ldots,k-1}^{-1}(u_k|x_{i+1},\ldots,x_{k-1}) \\ x_1 = F_{1|i+1,\ldots,k}^{-1}(u_1|x_{i+1},\ldots,x_k) \\ \vdots \\ x_i = F_{i|i+1,\ldots,k,1,\ldots,i-1}^{-1}(u_i|x_{i+1},\ldots,x_k, x_1,\ldots,x_{i-1}) \end{cases} \quad (32)$$

The inverse Rosenblatt transformation is a general approach because it is valid for any kind of joint distribution and dependence structure, especially when the inputs have structural dependences imposed by non-rectangular domains (see e.g. [49,30]). To use this transformation, it is required that the conditional distributions should be available for all model inputs. However, for practical engineering problems, it is often rather difficult or even sometimes impossible to obtain the conditional CDFs for arbitrary distributions, and the CDFs of some well-known distributions (e.g., the normal distribution) do not have a close form. Moreover, when the inputs have different distribution types, it is also rather challenging to derive the inverse conditional CDFs with explicit mathematical formulations. In these cases, the inverse Rosenblatt transformation is not a feasible option for generating dependent random samples.

Inverse nataf transformation

In practice, it is very common to use the pairwise inputs correlation to define the dependence structure. Such information can be easily obtained from e.g., literature, relevant research, empirical data, expert opinions, or even reasonable hypotheses. When the marginal distributions and correlations are known, the inverse Nataf transformation [46–48,50] can be adopted to generate dependent random samples with given marginals and correlations.

Let $\mathbf{Z^c} = \{Z_1^c, Z_2^c, \ldots, Z_k^c\}$ be a set of correlated random variables with standard multivariate normal distribution $\mathcal{N}(0, \mathbf{\Sigma_Z})$ ($\mathbf{\Sigma_Z}$ is the covariance matrix). Let $\mathbf{R_Z}$ be the linear correlation matrix (i.e., Pearson correlation matrix) of $\mathbf{Z}^c$. Obviously, $\mathbf{\Sigma_Z} = \mathbf{R_Z}$ because $Z_i^c \sim \mathcal{N}(0, 1)$, $\forall\ i \in [1, k]$. Moreover, let $\mathbf{z}^c = \{z_1^c, \ldots, z_k^c\}$ be a set of possible values drawn from the random samples of $\mathbf{Z^c}$. The values of the dependent random samples of $\mathbf{X}$ can be obtained using the following component-wise transformation, i.e., the inverse Nataf transformation:

$$x_i = F_i^{-1}(\Phi(z_i^c)), \quad (33)$$

where $\Phi(\cdot)$ is the CDF of the univariate standard normal distribution, $F_i^{-1}(\cdot)$ is the inverse CDF of X_i. Eq. (33) assures that the distribution of any transformed random variable matches its desired marginal distribution. Thus, we just need to correctly define the correlation of $\mathbf{Z}^c$, so that the random samples of $\mathbf{X}$ generated by the inverse Nataf transformation will also have the desired correlation.

The inverse Nataf transformation $F_i^{-1}(\Phi(\cdot))$ is monotonic, but in most cases (e.g., X_i has non-normal marginal distribution) it is also non-linear. This makes the linear correlation coefficient of Z_i^c and Z_j^c (i.e., $\rho_{Z_i^c,Z_j^c}$) different from that of X_i and X_j (i.e., ρ_{X_i,X_j}). Thus, when the marginal distributions are given, and the desired correlation is defined by the *linear* correlation matrix of $\mathbf{X}$ (i.e., $\mathbf{R_X}$), $\mathbf{R_Z}$ must be carefully adapted in order to make the linear correlation matrix of the generated random samples match $\mathbf{R_X}$. When all inputs have standard distributions (e.g., normal, uniform, lognormal), $\mathbf{R_Z}$ can be computed using empirical estimators (e.g., [47,48]) or algorithms (e.g., [50]). However, when some inputs have non-standard distributions and/or non-linear correlations, the selection of a proper $\mathbf{R_Z}$ can still be very difficult or even unfeasible.

As an alternative, we propose to use the *rank* correlation matrix of $\mathbf{X}$ to describe the desired correlation. Specifically, Spearman's rank correlation [51] is adopted in this paper. Note that it is also possible to employ other rank correlations such as Kendall's Tau [51]. The formula for computing the Spearman's Rank Correlation Coefficient (SRCC) of inputs X_i and X_j is given below:

$$\rho_{X_i,X_j}^s = 1 - \frac{6\sum_{t=1}^n (d_i^t - d_j^t)^2}{n(n^2-1)}, \quad (34)$$

where d_i^t stands for the rank (in the ascending order) of the t-th observation of X_i, and n is the total number of observations. For example, if X_i has three observations {4.1, 0.1, 2.8}, then $\{d_i^1, d_i^2, d_i^3\} = \{3, 1, 2\}$.

The rank is an invariant measure under monotonic transformations, therefore $\rho_{Z_i^c,Z_j^c}^s = \rho_{X_i,X_j}^s$ for $\forall\ \{i, j\} \in [1, k]$. In other words, if we want the random samples of $\mathbf{X}$ to match the desired rank correlation matrix, we just need to make sure that the normally distributed variables $\mathbf{Z}^c$ have the same rank correlation matrix before performing the inverse Nataf transformation. Given the target SRCC ρ_{X_i,X_j}^s, to parameterize the distribution of $\mathbf{Z}^c$, the linear correlation matrix $\mathbf{R_Z}$ is derived (see [52,53]) by:

$$R_{ij} = \rho_{Z_i^c,Z_j^c} = 2\sin\left(\frac{\pi}{6}\rho_{Z_i^c,Z_j^c}^s\right) = 2\sin\left(\frac{\pi}{6}\rho_{X_i,X_j}^s\right), \quad (35)$$

where R_{ij} is the element of $\mathbf{R_Z}$ located at the i-th row and j-th column.

Let $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ be a set of uncorrelated standard normal variables. Let $\mathbf{C}$ be the correlation matrix of the sample of $\mathbf{Z}$. Because of random simulation error, in practice $\mathbf{C}$ is only approximately equal to a perfect identity matrix. The process for generating random samples with given marginal distributions F_i and Spearman's rank correlations ρ_{X_i,X_j}^s can be summarized as below:

(1) Let $\mathbf{z} = \{z_1, \ldots, z_k\}$ be a draw of the random samples of $\mathbf{Z}$. $\mathbf{z}$ can be derived from an independent uniform vector $\mathbf{u} = \{u_1, u_2, \ldots, u_k\}$ by using the transformation $z_i = \Phi^{-1}(u_i)$, where $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard normal distribution.
(2) Use Eq. (35) to compute $\mathbf{R_Z}$ with given ρ_{X_i,X_j}^s. Note that if $\mathbf{X}$ are normally distributed, this step can be skipped because $\mathbf{R_Z} = \mathbf{R_X}$.
(3) Perform Cholesky decomposition of $\mathbf{R_Z}$: $\mathbf{R_Z} = \mathbf{M}'\cdot\mathbf{M}$, where $\mathbf{M}$ is an upper triangular matrix.
(4) Perform Cholesky decomposition of $\mathbf{C}$: $\mathbf{C} = \mathbf{Q}'\cdot\mathbf{Q}$, where $\mathbf{Q}$ is an upper triangular matrix.
(5) Derive the dependent normally distributed vector $\mathbf{z}^c$: $\mathbf{z}^c = \mathbf{z}\cdot\mathbf{Q}^{-1}\cdot\mathbf{M}$.
(6) Use the inverse Nataf transformation to obtain the desired random sample: $x_i = F_i^{-1}(\Phi(z_i^c))$, $i = 1, \ldots, k$.

The above process is also known as Iman Conover (IC) procedure [54]). It assumes that the inputs $\mathbf{X}$ can be defined by a Gaussian copula [29]. In practice, however, modelers may choose other families of copulas (e.g., elliptical, Archimedean, see [55]), especially when the inputs have non-symmetric dependence, or when extreme values need to be considered.

3.3. Computation of the extended EE

Let $\mathcal{T}(\mathbf{p}_{i,r}, \omega)$ be a transformation function. This function transforms an independent uniform random vector $\mathbf{p}_{i,r}$ (i.e., the i-th point in Table 1, $i \in [1, k+1]$, $r \in [1, N]$) into a dependent random vector under given marginal distributions and dependence structure. ω is the shift parameter, $\omega \in [0, k]$.

$\mathcal{T}(\mathbf{p}_{i,r}, \omega)$ contains 3 sequential transformations, i.e., $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$:

$$\mathcal{T}(\mathbf{p}_{i,r}, \omega) = \mathcal{T}_3(\mathcal{T}_2(\mathcal{T}_1(\mathbf{p}_{i,r}, \omega))) \quad (36)$$

The specific process of $\mathcal{T}(\mathbf{p}_{i,r}, \omega)$ is explained below:

(1) $\mathcal{T}_1(\mathbf{p}_{i,r}, \omega)$: shift the first ω elements of $\mathbf{p}_{i,r}$ to its end, and hence generate an independent vector $\mathbf{p}_{i,r}^1$, $\mathbf{p}_{i,r}^1 \sim \mathcal{U}^k[0, 1]$. Specifically, when ω is 0 or k, $\mathbf{p}_{i,r}^1 = \mathbf{p}_{i,r}$. For the trajectory design in Table 1, if $\omega = i$, then $\mathbf{p}_{i,r}^1 = \mathcal{T}_1(\mathbf{p}_{i,r}, i) = \{p_{i+1,r}, \ldots, p_{k,r}, p_{1,r}, \ldots, p_{i,r}\} = \{a_{i+1,r}, \ldots, a_{k,r}, b_{1,r}, \ldots, b_{i-1,r}, a_{i,r}\}$.

(2) $\mathcal{T}_2(\mathbf{p}^1_{i,r})$: use the inverse Rosenblatt transformation (see Eq. (32)) or the inverse Nataf transformation (see Eq. (33) and steps in the IC procedure) to transform $\mathbf{p}^1_{i,r}$ into a new vector $\mathbf{p}^2_{i,r}$ according to given marginal PDFs and dependence structure. As a result, $\mathbf{p}^2_{i,r} = \mathcal{T}_2(\mathbf{p}^1_{i,r}) = \{x_{i+1,r}, \ldots, x_{k,r}, x_{1,r}, \ldots, x_{i,r}\}$.

(3) $\mathcal{T}_3(\mathbf{p}^2_{i,r})$: shift the first $(k-\omega)$ elements of $\mathbf{p}^2_{i,r}$ back to the end, so that the elements of the new vector, i.e., $\mathbf{p}^3_{i,r}$, are in the same order as the elements of $\mathbf{p}_{i,r}$. In this step, $\mathbf{p}^3_{i,r} = \mathcal{T}_3(\mathbf{p}^2_{i,r}) = \{x_{1,r}, \ldots, x_{k,r}\}$.

When the independent random samples are generated using the trajectory design, we can perform the transformation $\mathcal{T}(\cdot,i)$ with $\mathbf{p}_{i,r}$ and $\mathbf{p}_{i+1,r}$ in Table 1, and obtain two dependent vectors, namely, $\mathcal{T}(\mathbf{p}_{i,r}, i)$ and $\mathcal{T}(\mathbf{p}_{i+1,r}, i)$. It is easy to prove that $\mathcal{T}(\mathbf{p}_{i,r}, i)$ and $\mathcal{T}(\mathbf{p}_{i+1,r}, i)$ follow the same joint distribution $p(\mathbf{X})$. In addition, all elements in these two vectors are the same except the i-th element. If we let $\{x_{i,r}, \mathbf{x}_{\sim i,r}\} = \mathcal{T}(\mathbf{p}_{i,r}, i)$, and $\{x'_{i,r}, \mathbf{x}_{\sim i,r}\} = \mathcal{T}(\mathbf{p}_{i+1,r}, i)$, according to Eq. (29), the r-th independent EE of X_i, i.e., $EE^{ind}_{i,r}$, is derived as:

$$EE^{ind}_{i,r} = \frac{f(\mathcal{T}(\mathbf{p}_{i+1,r}, i)) - f(\mathcal{T}(\mathbf{p}_{i,r}, i))}{\Delta}. \tag{37}$$

Similarly, the values assigned to the r-th conditional and unconditional random samples, i.e., $\{x'_{i,r}, \mathbf{x}_{\sim i,r}\}$ and $\{x_{i,r}, \mathbf{x}_{\sim i,r}\}$, can be derived by applying the transformation $\mathcal{T}(\cdot, i-1)$ on the points $\mathbf{p}_{i+1,r}$ and $\mathbf{p}_{i,r}$, respectively. Accordingly, the r-th full EE of X_i, i.e., $EE^{full}_{i,r}$, can be derived as:

$$EE^{full}_{i,r} = \frac{f(\mathcal{T}(\mathbf{p}_{i+1,r}, i-1)) - f(\mathcal{T}(\mathbf{p}_{i,r}, i-1))}{\Delta}. \tag{38}$$

The computation of the independent and full EEs can be similarly performed when the radial design is used. Let $\mathbf{q}_{1,r}$ and $\mathbf{q}_{i+1,r}$ denote two vectors that represent the first and $(i+1)$-th points in the r-th radial sample, respectively. The independent and full EEs computed by using the random samples from the radial design are:

$$EE^{ind}_{i,r} = \frac{f(\mathcal{T}(\mathbf{q}_{i+1,r}, i)) - f(\mathcal{T}(\mathbf{q}_{1,r}, i))}{b_{i,r} - a_{i,r}}, \tag{39}$$

$$EE^{full}_{i,r} = \frac{f(\mathcal{T}(\mathbf{q}_{i+1,r}, i-1)) - f(\mathcal{T}(\mathbf{q}_{1,r}, i-1))}{b_{i,r} - a_{i,r}}. \tag{40}$$

The simulation runs required by the samples from the trajectory design and the radial design are $N(3k+1)$ and $3kN$, respectively.

4. Numerical experiment

This section includes three numerical experiments to test and demonstrate the performance of the extended EE method. The extended EE method is used to identify influential and non-influential inputs from models containing both independent and dependent inputs. We employ the aforementioned variance-based GSA [29,30] to calculate the independent and full sensitivity indexes, i.e., $S^{T,ind}$ and $S^{T,full}$, based on analytical formulations (i.e., Eqs. (22) and (23)) or Quasi-Monte Carlo (QMC) estimations (i.e., Eqs. (27) and (28) with Sobol's quasi-random sequence [44]). $S^{T,ind}$ and $S^{T,full}$ are used thereafter as the reference for assessing the screening ability of the extended EE method.

As mentioned in Section 2.1, the classic Morris method and its revisions are able to rank independent inputs (see e.g. [9,36–38,33,39,40]). It is, therefore, meaningful to know if the extended EE method can also provide correct inputs ranking when the model contains both independent and dependent inputs. Thus, in the experiments the inputs ranking obtained via $\mu^{*,ind}$ and $\mu^{*,full}$ is cross-compared with the ranking obtained via $S^{T,ind}$ and $S^{T,full}$. The results are reported below.

4.1. Test case 1: linear function

Consider a simple linear function with 3 inputs, i.e., $f(X_1, X_2, X_3) = X_1 + X_2 + X_3$. This test function has been used in [29,30] for testing the performance of the extended variance-based GSA for dependent inputs. It is assumed that all inputs have standard normal marginal distributions, i.e., $X_i \sim \mathcal{N}(0,1)$. The dependence structure is defined by the following linear correlation matrix $\mathbf{R}$:

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}.$$

Moreover, this test case includes two test scenarios using different sets of correlation coefficients. In the first scenario, positive correlations are considered, i.e., $[\rho_{12}, \rho_{13}, \rho_{23}] = [0.9, 0.4, 0.01]$. In the second scenario, it is assumed that negative correlations exist for inputs pairs $X_1 - X_2$ and $X_1 - X_3$, i.e., $[\rho_{12}, \rho_{13}, \rho_{23}] = [-0.9, -0.4, 0.01]$.

To compute $(\mu^{*,ind}, \sigma^{ind})$ and $(\mu^{*,full}, \sigma^{full})$, in this experiment we employ 90 random samples generated by the trajectory design (i.e., $N = 90$), and 100 random samples generated by the radial design (i.e., $N = 100$). Hence, the two designs require the same number of model runs, i.e., 900 runs. Moreover, because all inputs are normally distributed, the inverse Nataf transformation is linear in this case. Thus, Eq. (35) is not needed as we can directly apply the IC procedure by letting $\mathbf{R}_\mathbf{Z} = \mathbf{R}$. $S^{T,ind}$ and $S^{T,full}$ are computed using the analytical formulas Eqs. (22) and (23). The results of the two test scenarios are presented in Figs. 1 and 2, respectively.

Fig. 1 shows that the extended EE method based on the trajectory design and the radial design yield the same results in the first test scenario. Specifically, if only considering the independent variance contribution (see Figs. 1a and 1c), X_3 is a highly influential input, while X_1 and X_2 are inputs with limited influence. On the other hand, when we consider the full variance contributions including the dependent contribution (see Figs. 1b and 1d), because of the positive correlation, X_1 and X_2 are even more influential than X_3. As none of the three inputs have both $(\mu^{*,ind}, \sigma^{ind})$ and $(\mu^{*,full}, \sigma^{full})$ close to 0, all inputs are considered as influential inputs. The screening results are consistent with the results obtained with the variance-based GSA (Fig. 1e), which also suggests all inputs are influential inputs. Moreover, the inputs rankings based on $\mu^{*,ind}$ and $\mu^{*,full}$ are identical with that based on $S^{T,ind}$ and $S^{T,full}$, i.e., $S^{T,ind}_1 < S^{T,ind}_2 < S^{T,ind}_3$, and $S^{T,full}_3 < S^{T,full}_2 < S^{T,full}_1$.

The second test scenario introduces negative correlations among the inputs. Fig. 2 shows that in this scenario, the analysis based on the trajectory design again yields the same results as the analysis based on the radial design. X_3 is surely an influential input, with or without considering its dependent contribution. X_1 has rather low $\mu^{*,ind}$ but relatively high $\mu^{*,full}$, which means that its full variance contribution including the dependent contribution is not negligible. Therefore, to avoid making a Type II error in the input screening, X_1 should also be considered as an influential input. In this scenario, the non-influential input is X_2, as both $(\mu^{*,ind}, \sigma^{ind})$ and $(\mu^{*,full}, \sigma^{full})$ are very small. Consequently, X_2 can be fixed without influencing the results in the subsequent analysis. When making a cross-comparison between the results obtained with the extended EE method and the variance-based GSA, we can see that both methods produce the same inputs ranking in this test scenario.

It is worth mentioning that in the first scenario, EE^{ind}_i is always smaller than EE^{full}_i for $\forall\, i \in [1, 3]$. This means that under this dependence structure, the impact due to dependence always has a positive contribution to the total variance. On the other hand, in the second scenario, the impact of dependence is negative for X_2, which results in $EE^{ind}_2 > EE^{full}_2$. These findings highlight the importance of using both independent and full EEs for screening inputs, as the judgment on influential or non-influential inputs can be biased if only one measure is used.

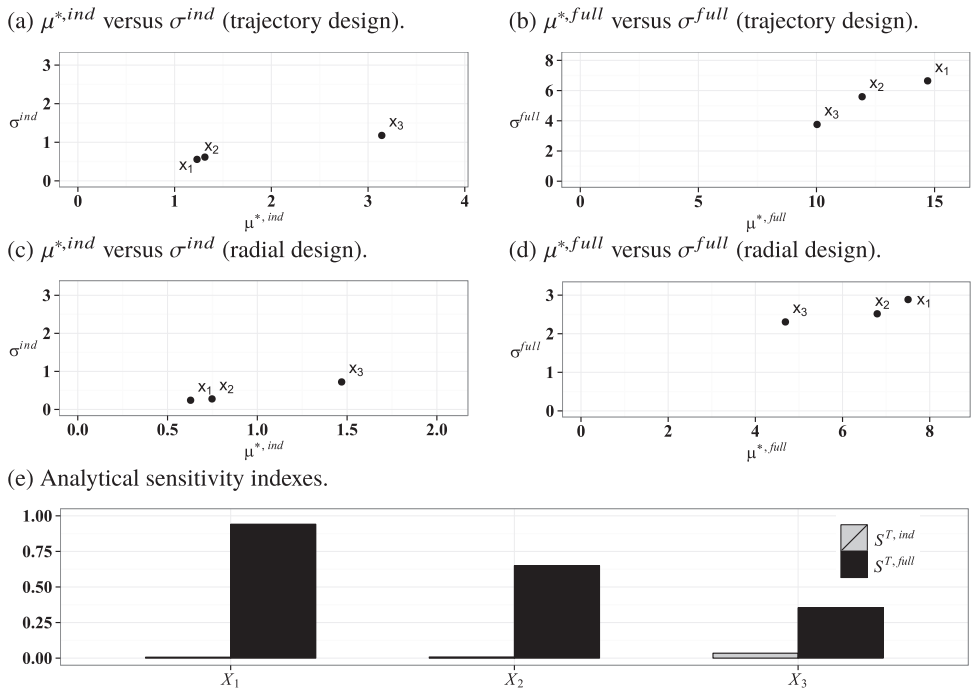

Fig. 1. Results of test case 1 scenario 1: $f(X_1,\ X_2,\ X_3) = X_1 + X_2 + X_3$, $X_i \sim \mathcal{N}(0,\ 1)$ for $\forall\ i \in [1,\ 3]$. $[\rho_{12}, \rho_{13}, \rho_{23}] = [0.9,\ 0.4,\ 0.01]$.

To sum up, this experiment shows that the extended EE method has the same accuracy as the variance-based GSA in terms of identifying influential and non-influential inputs. In addition, the inputs ranking derived based on $\mu^{*,ind}$ and $\mu^{*,full}$ match well with that based on $S^{T,ind}$ and $S^{T,full}$. This demonstrates the accuracy of the extended EE method in both screening and ranking the dependent inputs, while only a small number of model runs are required by both sampling designs.

4.2. Test case 2: non-linear function

A simple non-linear function with four inputs, i.e., $f(X_1, X_2, X_3, X_4) = X_1 \cdot X_3 + X_2 \cdot X_4$, is employed in the second test case. This test function has been used in [29] for testing the performance of the extended variance-based GSA.

For demonstration purposes, this test considers a more complex case, in which the four parameters are assumed to have four different marginal distributions: X_1 has a normal distribution $\mathcal{N}(0,\ 1)$, X_2 has a gamma distribution $\Gamma(2,\ 1)$, X_3 has a uniform distribution $\mathcal{U}[0,\ 1]$, and X_4 has a log-normal distribution $\ln\mathcal{N}(0,\ 1)$. In addition, it is assumed that Spearman's rank correlation between X_1 and X_4 is $\rho^s_{14} = 0.8$, and that between X_3 and X_4 is $\rho^s_{34} = 0.3$. The SRCC of other inputs pairs are arbitrarily set to 0.

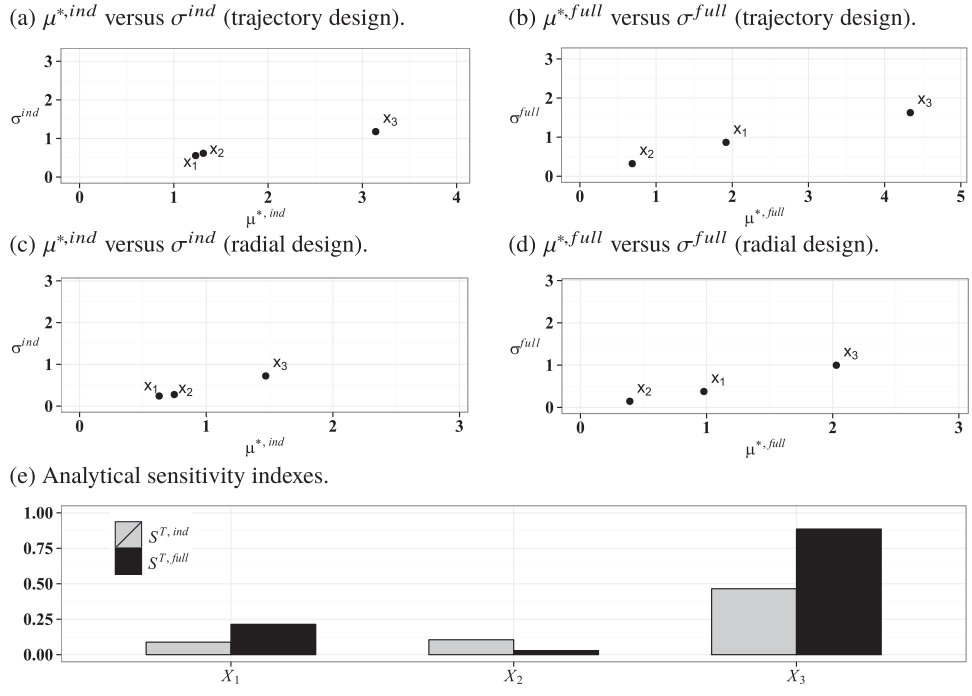

Fig. 2. Results of test case 1 scenario 2: $f(X_1,\ X_2,\ X_3) = X_1 + X_2 + X_3$, $X_i \sim \mathcal{N}(0,\ 1)$ for $\forall\ i \in [1,\ 3]$. $[\rho_{12}, \rho_{13}, \rho_{23}] = [-0.9,\ -0.4,\ 0.01]$.

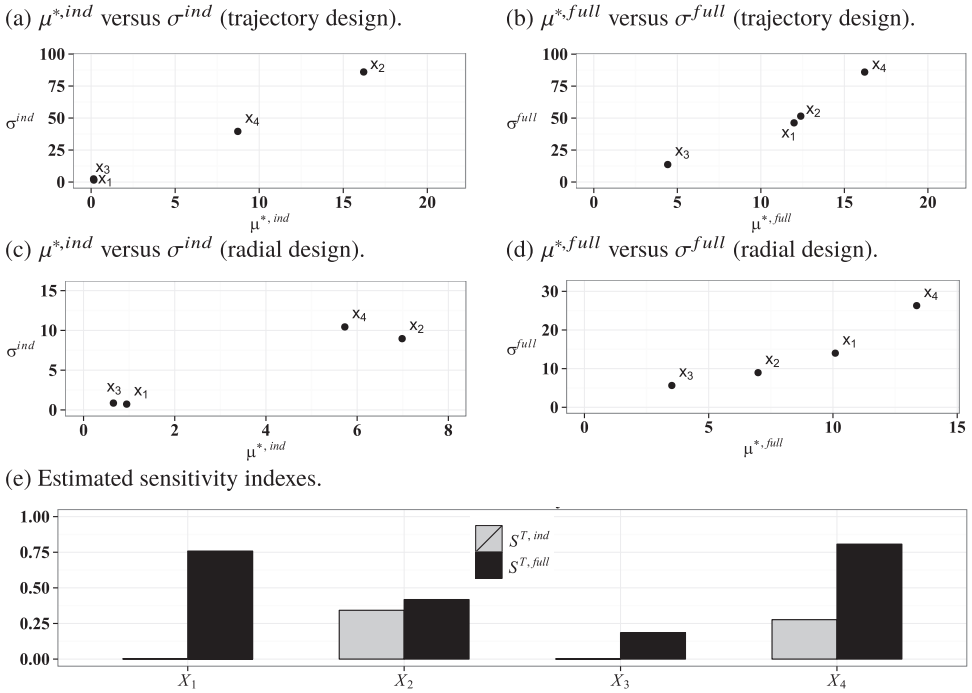

Fig. 3. Results of test case 2: $f(X_1, X_2, X_3, X_4) = X_1 \cdot X_3 + X_2 \cdot X_4$. $X_1 \sim \mathcal{N}(0,\ 1)$, $X_2 \sim \Gamma(2,\ 1)$, $X_3 \sim \mathcal{U}[0,\ 1]$, and $X_4 \sim \ln\mathcal{N}(0,\ 1)$. $\rho^s_{14} = 0.8$, $\rho^s_{34} = 0.3$.

The computation of $(\mu^{*,\,ind}, \sigma^{ind})$ and $(\mu^{*,\,full}, \sigma^{full})$ is based on 240 random samples generated by the trajectory design (i.e., $N = 240$), and 260 random samples generated by the radial design (i.e., $N = 260$). In total, both designs run the model 3120 times. As the four inputs have different marginal distributions, deriving the conditional CDFs becomes very complex in this case. Therefore, the inverse Nataf transformation is preferred here for generating dependent random samples. Because X_2, X_3, and X_4 have non-normal marginal distributions, the inverse Nataf transformation is non-linear for these inputs. Thus, Eq. (35) is required to transform the corresponding SRCCs to $\mathbf{R_Z}$ in the IC procedure. Moreover, due to the complexity of the analytical formulation of $S^{T,ind}$ and $S^{T,full}$, we employ the MC estimators in Eqs. (27) and (28) to compute the total sensitivity indexes. The estimation uses a size of 10,000 Sobol' quasi-random numbers, hence the total computational cost of the variance-based GSA is $3kN = 120,000$. The results are illustrated in Fig. 3.

In general, the screen results obtained with the trajectory design are consistent with those obtained with the radial design. Both Figs. 3a and 3c show that if not considering the dependent contribution, X_2 and X_4 are very influential, while X_1 and X_3 are the least influential inputs. If the dependent contribution is considered (see Figs. 3b and 3d), X_4 becomes the most influential input due to the high $(\mu^{*,\,ind}, \sigma^{ind})$ and $(\mu^{*,\,full}, \sigma^{full})$, while X_3 is still the input with the least influence. The results hence suggest that with such marginal distributions and dependence structure, X_3 can be considered as a non-influential input, and it can be fixed in the subsequent analysis. Moreover, although X_1 is one of the least influential inputs in Figs. 3a and 3c, since it has a strong correlation with the most influential input X_4, its dependent contribution should not be ignored. In other words, X_1 is also considered as an influential parameter in this case.

As for the inputs ranking, the findings from the extended EE method are generally consistent with those from the variance-based GSA (see Fig. 3e). Notice that there is a minor difference in inputs ranking with the two sampling designs. When considering the dependent contribution, X_2 seems to be more influential than X_1 if the trajectory design is used for sampling (Fig. 3b), while the analysis based on the radial design suggests that X_1 is more influential than X_2 (Fig. 3d). According to the reference results obtained with the QMC estimation, it is clear that the ranking obtained with the radial design is correct. This may be explained by the fact that the radial design does not require a fixed variation (i.e., Δ in Eqs. (29) and (30)) when generating the random samples, and this could result in a better exploration of the input space than the trajectory design. Such feature is more desirable for the case when both the input space and the model are non-linear, and hence the extended EE method with the radial design is expected to yield more accurate results in terms of inputs ranking. Nonetheless, since the inputs ranking obtained with the trajectory design does not influence the screening results (i.e., there are no Type I nor Type II errors in this case), the trajectory design can still be adopted if the major purpose is to identify and fix non-influential inputs.

4.3. Test case 3: G-function

The third test case employs a G-function [56]. It is a strongly non-linear and non-monotonic function, and it is commonly employed as benchmark for evaluating the SA approaches for models with independent inputs (see examples in [3,9,17,42]). It is defined as below:

$$G = \prod_{i=1}^{n} g_i(X_i) \quad \text{with } g_i(X_i) = \frac{|4X_i - 2| + a_i}{1 + a_i}, \tag{41}$$

where X_i $(i \in [1,\ n])$ has uniform marginal distribution $\mathcal{U}[0,\ 1]$. The coefficient a_i, which defines the variations of the function $g_i(X_i)$, is always non-negative.

In this experiment we consider a G-function with 12 inputs. For the ease of demonstration, we employ the same coefficients a_i $(i \in [1,\ 12])$ as those provided in [9]: $[a_1, \ldots, a_{12}] = [0.001, 89.9, 5.54, 42.1, 0.78, 1.26, 0.04, 0.79, 74.51, 4.32, 82.51, 41.62]$. In addition, it is assumed that X_1 and X_2 have positive rank correlation $\rho^s_{12} = 0.8$, and X_7 and X_9 have negative rank correlation $\rho^s_{79} = -0.8$. The SRCCs for other inputs pairs are arbitrarily set to 0.

We use 390 random samples generated by the trajectory design (i.e., $N = 390$), and 400 random samples generated by the radial design (i.e., $N = 400$) to compute $(\mu^{*,\,ind}, \sigma^{ind})$ and $(\mu^{*,\,full}, \sigma^{full})$. Due to the complexity of the G-function and its high dimensionality, we again use the inverse Nataf transformation to generate dependent random samples. The total computational cost required by the trajectory design

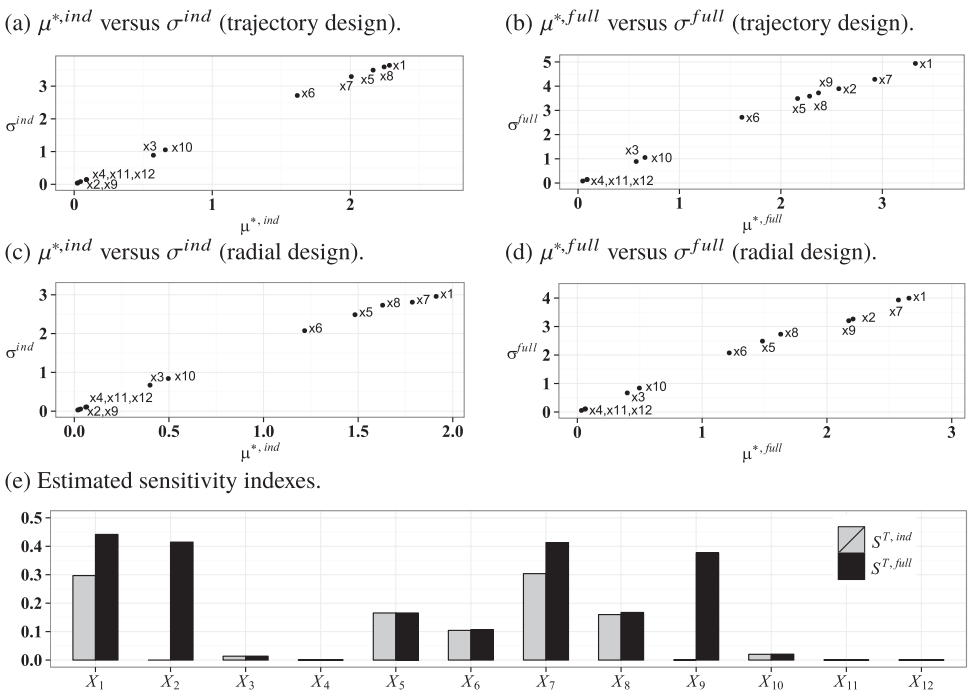

Fig. 4. Results of test case 3: G-function with 12 inputs. $[a_1, \ldots, a_{12}]$ = [0.001, 89.9, 5.54, 42.1, 0.78, 1.26, 0.04, 0.79, 74.51, 4.32, 82.51, 41.62], $\rho^s_{12} = 0.8$, and $\rho^s_{79} = -0.8$.

and the radial design are 14,430 and 14,400, respectively. In addition, since the sensitivity indexes are difficult to compute analytically, the MC estimator is used again to produce the reference results. Specifically, we use a size of 10,000 Sobol' quasi-random numbers in the QMC estimation, which results in a total computational cost of $3kN = 360,000$ model runs. The results are presented in Fig. 4.

Fig. 4 shows that independent of the sampling approach used, X_1, X_5, X_6, X_7, and X_8 always have high ($\mu^{*,ind}$, σ^{ind}) and ($\mu^{*,full}$, σ^{full}). Hence, they are the influential inputs, with or without considering the dependent contribution. In Figs. 4a and 4c, X_2 and X_9 have the lowest ($\mu^{*,ind}$, σ^{ind}), thus they are considered as non-influential inputs if the dependent contribution is ignored. However, both Figs. 4b and 4d show that these two inputs have rather high ($\mu^{*,full}$, σ^{full}). This indicates that the dependent contributions of these two inputs are not trivial. As a result, they are also identified as influential inputs in this case. X_4, X_{11}, and X_{12} have very low ($\mu^{*,ind}$, σ^{ind}) and ($\mu^{*,full}$, σ^{full}), therefore they are selected as non-influential inputs, and they can be fixed in the subsequent analysis.

The above findings can be explained by the fact that in the G-function, an input with a small coefficient a_i will be an influential input, and vice versa. In this case, the coefficients a_1, a_5, a_6, a_7, and a_8 are much smaller than the other coefficients, thus it is certain that the corresponding inputs X_1, X_5, X_6, X_7, and X_8 must be influential inputs. On the contrary, X_4, X_{11}, and X_{12} are independent of other influential inputs, and their coefficients a_4, a_{11}, and a_{12} are rather high, thus they should be non-influential inputs. The coefficients a_2 and a_9 are also high, which suggests that the corresponding inputs X_2 and X_9 would be non-influential inputs if their dependent contributions were not considered. However, in the given dependence structure, X_2 and X_9 have strong correlations with the two most influential inputs X_1 and X_7, respectively. As a result, the variations of X_2 and X_9 will also bring variations to X_1 and X_7, i.e., X_2 and X_9 have indirect variance contributions to the model output. Therefore, X_2 and X_9 are also identified as influential inputs in this case.

When taking the sensitivity indexes computed by the QMC estimation (Fig. 4e) as references, and cross comparing the results obtained with the two sampling designs, it is again found that the results obtained with the trajectory design are slightly less accurate than the results obtained with the radial design. For instance, X_7 is less influential than X_5 and X_8 in Fig. 4a, while Fig. 4c shows X_7 is more influential than the other two parameters. According to the inputs ranking by $S^{T,ind}$ in Fig. 4e, the results obtained with the radial design in Fig. 4c are correct. Nonetheless, if we consider that the inconsistency of the inputs ranking does not produce any Type I nor Type II error in selecting the influential and non-influential inputs, we can still agree that the screening results from the trajectory design are satisfactory.

4.4. Summary of the results

To sum up, the three test cases show that the extended EE method has satisfactory accuracy in screening both dependent and independent inputs. In particular, the screening results (i.e., the sets of influential and non-influential inputs) obtained with this method are consistent with the reference results obtained with the variance-based GSA. In addition, the extended EE method with the radial design always yields the same inputs ranking as the variance-based GSA for all test cases. The extended EE with the trajectory design also performs well in all tests, although there are some errors in inputs ranking in test cases 2 and 3. However, as in these two test cases the errors in inputs ranking do not substantially influence the judgment about whether an input is influential or non-influential (i.e., there is no Type I nor Type II error), these errors could be considered trivial for the preliminary screening analysis.

Moreover, if compared with the variance-based GSA, which is in general more accurate in inputs ranking but also requires a large number of model runs (e.g., 120,000 for test cases 2 and 3, respectively), the extended EE method has a comparable accuracy in input screening, but it has much less computational cost (e.g., 3120 and 14,400 runs for numerical experiments 2 and 3, respectively). Therefore, because of its high efficiency in computation, this method is especially useful for the initial input screening of high dimensional and/or computationally expensive models which contain both independent and dependent parameters.

Last but not least, the proposed method is distribution free, and it can work even without knowing the conditional distributions of the inputs. This feature is very attractive for the inputs screening of many

commercial simulation models, which often contain many inputs from different arbitrary distributions.

5. Conclusion

In this paper, a non-parametric approach is proposed for screening dependent inputs. It is developed based on the classic EE method (i.e., the Morris method), which is commonly used for screening independent inputs. Following the concept of variance decomposition for dependent inputs, which were initially introduced in [29,30], two extended elementary effects are therefore proposed: the independent elementary effects EE^{ind} and the full elementary effects EE^{full}. A qualitative analysis of these two measures can thus distinguish the corresponding input as a non-influential input when the absolute mean (i.e., μ^*) and standard deviation (i.e., σ) of both EE^{ind} and EE^{full} are close to 0, or an influential input in all other cases. Moreover, when $\mu^{*,ind}$ is close to 0 but $\mu^{*,full}$ is greater than 0, it means that the corresponding input is indirectly influential due to its dependence with other influential inputs.

Two sampling designs, namely, trajectory design and radial design, are implemented for generating independent random samples. The dependent random samples are generated by the inverse Rosenblatt transformation or the inverse Nataf transformation of the independent samples. The inverse Rosenblatt transformation is a general sampling approach, but it requires the knowledge of joint and conditional distributions of all inputs, which is sometimes more difficult or even unfeasible to obtain in many practical engineering applications. On the other hand, the inverse Nataf transformation only requires information about the marginal distributions and the inputs correlation (especially the rank correlation). As such information is commonly available from many sources (e.g., empirical experiments, literature, expert opinions, reasonable assumptions), the inverse Nataf transformation is more practical and appealing in industrial applications.

The performance of the proposed screening method is empirically evaluated by three numerical experiments. In these tests the results of the extended EE method are cross-compared with those obtained with the variance-based GSA. All test results have shown that the proposed method can properly distinguish the influential and non-influential inputs, even when the inputs have different marginal distributions (see test case 2), or the model contains many dependent and independent inputs (see test case 3). The inputs rankings based on $\mu^{*,ind}$ and $\mu^{*,full}$ are generally consistent with those based on $S^{T,ind}$ and $S^{T,full}$, yet the proposed approach requires much less model runs than the variance-based GSA using a QMC estimator. For example, in test case 3, the proposed screening method requires about 25 times less model runs than the QMC estimator.

Due to its high computational efficiency, the screening method proposed in this paper is very attractive for the purpose of preliminary screening of both independent and dependent inputs, especially when the model is high dimensional and/or computationally expensive. In such circumstances, the direct application of the variance-based GSA can be rather time consuming or even unfeasible. Hence, according to the FF setting, it is recommended to first use the proposed screening approach to find and fix the non-influential inputs, then conduct a quantitative analysis (e.g., variance-based GSA) sequentially to refine the SA results.

Acknowledgment

Research contained within this paper benefited from participation in the EU COST Action TU0903 Methods and tools for supporting the Use caLibration and validaTIon of Traffic simUlation moDEls (MULTITUDE) [57].

References

[1] Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global Sensitivity Analysis. The Primer, John Wiley & Sons, Ltd, Chichester, UK, 2007. ⟨http://dx.doi.org/10.1002/9780470725184⟩.
[2] Sobol IM, Kucherenko S. Derivative based global sensitivity measures and their link with global sensitivity indices. Math Comput Simul 2009;79(10):3009–17. http://dx.doi.org/10.1016/j.matcom.2009.01.023.
[3] Kucherenko S, Rodriguez-Fernandez M, Pantelides C, Shah N. Monte carlo evaluation of derivative-based global sensitivity measures. Reliab Eng Syst Saf 2009;94(7):1135–48. http://dx.doi.org/10.1016/j.ress.2008.05.006.
[4] Sobol IM, Kucherenko S. A new derivative based importance criterion for groups of variables and its link with the global sensitivity indices. Comput Phys Commun 2010;181(7):1212–7. http://dx.doi.org/10.1016/j.cpc.2010.03.006.
[5] Lamboni M, Iooss B, Popelin A-L, Gamboa F. Derivative-based global sensitivity measures: general links with sobol indices and numerical tests. Math Comput Simul 2013;87:45–54. http://dx.doi.org/10.1016/j.matcom.2013.02.002.
[6] Oakley JE, O'Hagan A. Probabilistic sensitivity analysis of complex models: a bayesian approach. J R Stat Soc: Ser B (Stat Methodol) 2004;66(3):751–69. http://dx.doi.org/10.1111/j.1467-9868.2004.05304.x.
[7] Doksum K, Samarov A. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. Ann Stat 1995;23(5):1443–73.
[8] Morris MD. Factorial sampling plans for preliminary computational experiments. Technometrics 1991;33(2):161–74. http://dx.doi.org/10.2307/1269043.
[9] Campolongo F, Cariboni J, Saltelli A. An effective screening design for sensitivity analysis of large models. Environ Model Softw 2007;22(10):1509–18. http://dx.doi.org/10.1016/j.envsoft.2006.10.004.
[10] Saltelli A, Campolongo F, Cariboni J. Screening important inputs in models with strong interaction properties. Reliab Eng Syst Saf 2009;94(7):1149–55. http://dx.doi.org/10.1016/j.ress.2008.10.007.
[11] Ge Q, Menendez M. An efficient sensitivity analysis approach for computationally expensive microscopic traffic simulation models. Int J Transp 2014;2(2):49–64. http://dx.doi.org/10.14257/ijt.2014.2.2.04.
[12] Wang X, Yang M, Zhang Y, Kiang M. A screening approach for non-parametric global sensitivity analysis. J Stat Comput Simul 2015(April):1–20. http://dx.doi.org/10.1080/00949655.2015.1027895.
[13] Sobol IM. Sensitivity estimates for nonlinear mathematical models. Math Model Comput Exp 1993;1(4):407–14.
[14] Saltelli A, Sobol IM. About the use of rank transformation in sensitivity analysis of model output. Reliab Eng Syst Saf 1995;50(3):225–39. http://dx.doi.org/10.1016/0951-8320(95)00099-2.
[15] Homma T, Saltelli A. Importance measures in global sensitivity analysis of nonlinear models. Reliab Eng Syst Saf 1996;52(1):1–17. http://dx.doi.org/10.1016/0951-8320(96)00002-6.
[16] Saltelli A, Tarantola S, Chan K-S. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 1999;41(1):39–56. http://dx.doi.org/10.1080/00401706.1999.10485594.
[17] Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Comput Phys Commun 2010;181(2):259–70. http://dx.doi.org/10.1016/j.cpc.2009.09.018.
[18] Chun M-H, Han S-J, Tak N-I. An uncertainty importance measure using a distance metric for the change in a cumulative distribution function. Reliab Eng Syst Saf 2000;70(3):313–21. http://dx.doi.org/10.1016/S0951-8320(00)00068-5.
[19] Borgonovo E. A new uncertainty importance measure. Reliab Eng Syst Saf 2007;92(6):771–84. http://dx.doi.org/10.1016/j.ress.2006.04.015.
[20] Liu Q, Homma T. A new importance measure for sensitivity analysis. J Nucl Sci Technol 2010;47(1):53–61. http://dx.doi.org/10.1080/18811248.2010.9711927.
[21] Ge Q, Ciuffo B, Menendez M. Combining screening and metamodel-based methods: an efficient sequential approach for the sensitivity analysis of model outputs. Reliab Eng Syst Saf 2014;134:334–44. http://dx.doi.org/10.1016/j.ress.2014.08.009.
[22] Brell G, Li G, Rabitz H. An efficient algorithm to accelerate the discovery of complex material formulations. J Chem Phys 2010;132(17):174103. http://dx.doi.org/10.1063/1.3407440.
[23] Xu C, Gertner GZ. Uncertainty and sensitivity analysis for models with correlated parameters. Reliab Eng Syst Saf 2008;93(10):1563–73. http://dx.doi.org/10.1016/j.ress.2007.06.003.
[24] Da Veiga S, Wahl F, Gamboa F. Local polynomial estimation for sensitivity analysis on models with correlated inputs. Technometrics 2009;51(4):452–63. http://dx.doi.org/10.1198/TECH.2009.08124.
[25] Li GY, Rabitz H, Yelvington PE, Oluwole OO, Bacon F, Kolb CE, et al. Global sensitivity analysis for systems with independent and/or correlated inputs. J Phys Chem A 2010;114(19):6022–32. http://dx.doi.org/10.1021/jp9096919.
[26] Xu C. Decoupling correlated and uncorrelated parametric uncertainty contributions for nonlinear models. Appl Math Model 2013;37(24):9950–69. http://dx.doi.org/10.1016/j.apm.2013.05.036.
[27] Zhang K, Lu Z, Cheng L, Xu F. A new framework of variance based global sensitivity analysis for models with correlated inputs. Struct Saf 2015;55(3):1–9. http://dx.doi.org/10.1016/j.strusafe.2014.12.005.
[28] Mara Ta, Tarantola S. Variance-based sensitivity indices for models with dependent inputs. Reliab Eng Syst Saf 2012;107:115–21. http://dx.doi.org/10.1016/j.ress.2011.08.008.
[29] Kucherenko S, Tarantola S, Annoni P. Estimation of global sensitivity indices for models with dependent variables. Comput Phys Commun 2012;183(4):937–46.

http://dx.doi.org/10.1016/j.cpc.2011.12.020.
[30] Mara Ta, Annoni P, Tarantola S. Non-parametric methods for global sensitivity analysis of model output with dependent inputs. Environ Model Softw 2015;72:173–83. http://dx.doi.org/10.1016/j.envsoft.2015.07.010.
[31] Saltelli A, Tarantola S. On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. J Am Stat Assoc 2002;97(459):702–9. http://dx.doi.org/10.1198/016214502388618447.
[32] Ge Q, Ciuffo B, Menendez M. An exploratory study of two efficient approaches for the sensitivity analysis of computationally expensive traffic simulation models. IEEE Trans Intell Transp Syst 2014;15(3):1288–97. http://dx.doi.org/10.1109/TITS.2014.2311161.
[33] Ge Q, Ciuffo B, Menendez M. Comprehensive approach for the sensitivity analysis of high-dimensional and computationally expensive traffic simulation models. Transp Res Rec: J Transp Res Board 2014;2422:121–30. http://dx.doi.org/10.3141/2422-14.
[34] Campolongo F, Saltelli A. Sensitivity analysis of an environmental model: an application of different analysis methods. Reliab Eng Syst Saf 1997;57(1):49–69. http://dx.doi.org/10.1016/S0951-8320(97)00021-5.
[35] Saltelli A, Tarantola S, Campolongo F, Ratto M. Sensitivity analysis in practice: a guide to assessing scientific model. John Wiley & Sons; 2004.
[36] Sin G, Gernaey KV, Lantz AE. Good modeling practice for PAT applications: propagation of input uncertainty and sensitivity analysis. Biotechnol Prog 2009;25(4):1043–53. http://dx.doi.org/10.1002/btpr.166.
[37] Ruano M, Ribes J, Seco a, Ferrer J. An improved sampling strategy based on trajectory design for application of the Morris method to systems with many input factors. Environ Model Softw 2012;37:103–9. http://dx.doi.org/10.1016/j.envsoft.2012.03.008.
[38] Nossent J, Bauwens W. Multi-variable sensitivity and identifiability analysis for a complex environmental model in view of integrated water quantity and water quality modeling. Water Sci Technol : a J Int Assoc Water Pollut Res 2012;65(3):539–49. http://dx.doi.org/10.2166/wst.2012.884.
[39] Morris DJ, Speirs DC, Cameron AI, Heath MR. Global sensitivity analysis of an end-to-end marine ecosystem model of the north sea: factors affecting the biomass of fish and benthos. Ecol Model 2014;273:251–63. http://dx.doi.org/10.1016/j.ecolmodel.2016.01.005.
[40] Branger F, Giraudet L-G, Guivarch C, Quirion P. Global sensitivity analysis of an energy-economy model of the residential building sector. Environ Model Softw 2015;70:45–54. http://dx.doi.org/10.1016/j.envsoft.2015.03.021.
[41] Chan K, Saltelli A, Tarantola S. Winding stairs: a sampling tool to compute sensitivity indices. Stat Comput 2000;10(3):187–96. http://dx.doi.org/10.1023/A:1008950625967.
[42] Campolongo F, Saltelli A, Cariboni J. From screening to quantitative sensitivity analysis. A Unifi- Approach, Comput Phys Commun 2011;182(4):978–88. http://dx.doi.org/10.1016/j.cpc.2010.12.039.
[43] Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math Comput Simul 2001;55:271–80. http://dx.doi.org/10.1016/S0378-4754(00)00270-6.
[44] Sobol IM. Uniformly distributed sequences with an additional uniform property. USSR Comput Math Math Phys 1976;16(5):236–42.
[45] Rosenblatt M. Remarks on a multivariate transformation. Ann Math Stat 23 1952;3:470–2. http://dx.doi.org/10.1214/aoms/1177729394.
[46] Nataf A. Determination des distributions de probabilites dont les marges sont donnees. Comptes Rendus Hebd Des Sances De lAcadmie Des Sci 1962;255(1):42–3.
[47] Liu P-L, Der Kiureghian A. Multivariate distribution models with prescribed marginals and covariances. Probabilistic Eng Mech 1986;1(2):105–12. http://dx.doi.org/10.1016/0266-8920(86)90033-0.
[48] Der Kiureghian A, Liu P. Structural Reliability under Incomplete Probability Information. J Eng Mech 1986;112(1):85–104. http://dx.doi.org/10.1061/(ASCE)0733-9399(1986)112:1(85).
[49] Kucherenko S, Klymenko O, Shah N. Sobol' indices for problems defined in non-rectangular domains, arXiv preprint arXiv:1605.05069arXiv:1605.05069.
[50] Li H, LU Z, Yuan X. Nataf transformation based point estimate method. Chin Sci Bull 2008;53(17):2586–92. http://dx.doi.org/10.1007/s11434-008-0351-0.
[51] Kendall MG. Rank correlation methods, 4th edition. London: Charles Griffin & Company Limited; 1970.
[52] Hotelling H, Pabst MR. Rank correlation and tests of significance involving no assumption of normality. Ann Math Stat 1936;7(1):29–43.
[53] Embrechts P, McNeil A, Straumann D. Correlation and dependence in risk management: properties and pitfalls. In: Dempster M, editor. Risk management: value at risk and beyond. Cambridge, UK: Cambridge University Press; 2002. p. 176–223.
[54] Iman RL, Conover WJ. A distribution-free approach to inducing rank correlation among input variables. Commun Stat - Simul Comput 1982;11(3):311–34. http://dx.doi.org/10.1080/03610918208812265.
[55] Nelsen RB. An introduction to copulas, 139. Springer Science & Business Media; 1999.
[56] Archer GEB, Saltelli a, Sobol IM. Sensitivity measures,anova-like Techniques and the use of bootstrap. J Stat Comput Simul 1997;58(2):99–120. http://dx.doi.org/10.1080/00949659708811825.
[57] MULTITUDE, Methods and tools for supporting the use, calibration and validation of traffic simulation models (2015). ⟨http://www.multitude-project.eu⟩