Reliability Engineering and System Safety 153 (2016) 110–121

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

An efficient modularized sample-based method to estimate the first-order Sobol' index

Chenzhao Li, Sankaran Mahadevan *

Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 1 June 2015
Received in revised form
28 March 2016
Accepted 20 April 2016
Available online 29 April 2016

Keyword:
Global sensitivity analysis
Sobol' index
Sample-based
Correlated variables

ABSTRACT

Sobol' index is a prominent methodology in global sensitivity analysis. This paper aims to directly estimate the Sobol' index based only on available input–output samples, even if the underlying model is unavailable. For this purpose, a new method to calculate the first-order Sobol' index is proposed. The innovation is that the conditional variance and mean in the formula of the first-order index are calculated at an unknown but existing location of model inputs, instead of an explicit user-defined location. The proposed method is modularized in two aspects: 1) index calculations for different model inputs are separate and use the same set of samples; and 2) model input sampling, model evaluation, and index calculation are separate. Due to this modularization, the proposed method is capable to compute the first-order index if only input–output samples are available but the underlying model is unavailable, and its computational cost is not proportional to the dimension of the model inputs. In addition, the proposed method can also estimate the first-order index with correlated model inputs. Considering that the first-order index is a desired metric to rank model inputs but current methods can only handle independent model inputs, the proposed method contributes to fill this gap.

1. Introduction

Uncertainty propagation problems generally involve a computational model in the form of $y=f(\boldsymbol{x},\boldsymbol{d})$ where $\boldsymbol{x}=\{x_1,\ldots,x_k\}$ is the vector of stochastic model inputs and $\boldsymbol{d}$ is the vector of deterministic inputs. Global sensitivity analysis (GSA) studies how the uncertainty in the output y can be apportioned to the uncertainty in the stochastic model inputs $\boldsymbol{x}=\{x_1,\ldots,x_k\}$, so that the importance of each stochastic model input can be ranked. Based on the result of GSA, inputs with negligible contribution can be fixed at their mean values thus reducing the number of stochastic variables. Reviews on various GSA methods can be found in [1,2]. The Sobol' sensitivity indices method based on variance decomposition is a prominent one among these methods. Usage of the Sobol' indices in different engineering problems can be found in [3–7].

Assuming that $y=f(\boldsymbol{x})$ is a real integrable function and all the model inputs $\boldsymbol{x}=\{x_1,\ldots,x_k\}$ are mutually independent, Sobol' [8] proved the following formula to decompose the variance of y:

$$V(y)=\sum_{i}^{k}V_i+\sum_{i_1=1}^{k}\sum_{i_2=i_1+1}^{k}V_{i_1i_2}+\sum_{i_1=1}^{k}\sum_{i_2=i_1+1}^{k}\sum_{i_3=i_2+1}^{k}V_{i_1i_2i_3}+\ldots+V_{12\ldots k} \quad (1)$$

where V_i indicates the variance of y caused by x_i individually, and $V_{i_1\ldots i_s}(s\geq 2)$ indicates the variance of y caused by the interaction of $\{x_{i_1},\ldots,x_{i_s}\}$.

Dividing $V(y)$ at both sides of Eq. (1) for normalization, the Sobol' index is defined as:

$$1=\sum_{i}^{k}S_i+\sum_{i_1=1}^{k}\sum_{i_2=i_1+1}^{k}S_{i_1i_2}+\sum_{i_1=1}^{k}\sum_{i_2=i_1+1}^{k}\sum_{i_3=i_2+1}^{k}S_{i_1i_2i_3}+\ldots+S_{12\ldots k} \quad (2)$$

where the index S_i measures the contribution of x_i alone to the variance of y, without interacting with any other inputs. S_i is called first-order index or main effects index. Other indices $S_{i_1\ldots i_s}(s\geq 2)$ in Eq. (2) are higher-order indices, measuring the contribution of the interaction of $\{x_{i_1},\ldots,x_{i_s}\}$.

* Corresponding author. Tel.: +1 615 322 3040.
E-mail address: sankaran.mahadevan@vanderbilt.edu (S. Mahadevan).

http://dx.doi.org/10.1016/j.ress.2016.04.012
0951-8320/

This paper focuses on calculating the first-order index S_i, which is one of the important objectives in variance-based global sensitivity analysis. The calculation of S_i is based on the following formula:

$$S_i = \frac{V_i}{V(y)} = \frac{V_{x_i}(E_{\boldsymbol{x}_{-i}}(y|x_i))}{V(y)} \tag{3}$$

where $\boldsymbol{x}_{-i}$ means all the model inputs other than x_i.

Based on Eq. (3), computing S_i analytically is nontrivial, since $E_{\boldsymbol{x}_{-i}}(\bullet)$ in Eq. (3) indicates a multi-dimensional integral. Computing S_i by Monte Carlo simulation (MCS) directly is also expensive. The numerator in Eq. (3) leads to a double-loop MCS [1]: the inner loop $E_{\boldsymbol{x}_{-i}}(y|x_i)$ computes the mean value of y using n_1 random samples of $\boldsymbol{x}_{-i}$; and the outer loop computes $V_{x_i}(E_{\boldsymbol{x}_{-i}}(y|x_i))$ by iterating the inner loop n_2 times at different values of x_i. In addition, another n_3 MCS iterations are required to compute $V(y)$ in Eq. (3). The cost of this sample-based method, defined as the total number of model evaluations to compute all S_i ($i = 1$ to k), is $kn_{dl}^2 + n_{dl}$ if $n_1 = n_2 = n_3 = n_{dl}$.This cost increases with n_{dl} and k, and is unaffordable if a single model evaluation is time-consuming or economically expensive, since n_{dl} is often of the order greater than 1000 in many practical applications.

Various algorithms have been proposed to reduce the computational cost of the Sobol' indices. These algorithms can be categorized into analytical methods and sample-based methods. In the analytical methods, the original model $y = f(\boldsymbol{x})$ is generally approximated by some surrogate model of special form, so that the multi-dimensional integral can be converted into multiple univariate integrals, which can be easily calculated analytically or numerically. Zhang and Pandey [9] approximated the original model $y = f(\boldsymbol{x})$ by a multiplication of univariate functions; then the univariate integral was calculated by Gaussian quadrature. Sudret [10] proposed that if the original model is approximated by a polynomial chaos expansion (PCE), the Sobol' index can be calculated by post-processing the PCE coefficients. Chen et al. [11] proposed another analytical method for commonly used surrogate models such as the linear regression model, Gaussian process model [12], Gaussian radial basis function model, and MARS model [13]; and analytical solution of the Sobol' index is available if the model inputs are normally or uniformly distributed. Analytical methods reduce the number of model evaluations significantly, but may require: 1) extra approximations and assumptions, and 2) extra computational cost in building the surrogate model.

Compared to the analytical methods, sample-based methods are more widely used [14–18] in engineering due to their simplicity in implementation. The basic sample-based method for GSA is the double-loop MCS, which has been explained earlier and often has prohibitive computational cost. Various efficient sample-based methods have been developed in the literature to reduce this cost. A brief review of these sample-based methods is given in Section 2. To the authors' knowledge, the computational cost (number of model evaluations) of most sample-based methods is proportional to the model input dimension k. Therefore the first objective of this paper is to develop a more efficient sample-based method whose computational cost is not proportional to k, but much less.

A key assumption of the Sobol' index is the mutual independence of model inputs. With correlated model inputs, Eqs. (1) and (2) are no longer valid. However, Saltelli [19] pointed out that the first-order index S_i is still an informed choice to rank the importance of correlated model inputs, since S_i can be defined in another way where independent model inputs are not assumed:

1. The importance of x_i at a particular location x_i can be measured by $V_{\boldsymbol{x}_{-i}}(y|x_i = \tilde{x}_i)$, i.e., smaller $V_{\boldsymbol{x}_{-i}}(y|x_i = \tilde{x}_i)$ indicates greater importance of x_i;
2. The dependence of this measurement on the location of x_i is removed by taking the average of $V_{\boldsymbol{x}_{-i}}(y|x_i = \tilde{x}_i)$, i.e. $E_{x_i}(V_{\boldsymbol{x}_{-i}}(y|x_i))$;
3. By the law of total variance $V(y) = E_{x_i}(V_{\boldsymbol{x}_{-i}}(y|x_i)) + V_{x_i}(E_{\boldsymbol{x}_{-i}}(y|x_i))$, a larger $V_{x_i}(E_{\boldsymbol{x}_{-i}}(y|x_i))$ equally indicates a greater importance of x_i;
4. The first-order index is redefined by normalization, thus $S_i = V_{x_i}(E_{\boldsymbol{x}_{-i}}(y|x_i))/V(y)$.

Saltelli's paper [20] in 2002 mentioned that there is no alternative to the expensive double-loop MCS to compute S_i with correlated model inputs. The authors have not found any efficient algorithm in more recent studies, either. Thus the second objective of this paper is to develop an efficient algorithm that can handle correlated model inputs.

The outline of this paper is as follows. Section 2 briefly reviews existing sample-based methods for GSA, and discusses their computational cost. Section 3 illustrates the proposed modularized sample-based method that reduces the computational cost and handles correlated model inputs. Section 4 uses three numerical examples to compare the proposed method with existing methods.

2. Background: sample-based methods to estimate the first-order Sobol' index

2.1. Sobol' scheme

Consider a real integrable function $y = f(\boldsymbol{x})$ where $\boldsymbol{x} = \{x_1, \ldots, x_k\}$ is the vector of independent model inputs. Denote $\boldsymbol{z} = \{z_1, \ldots, z_k\}$ as the vector of the same independent model inputs, i.e., $z_i (i = 1 \text{to} k)$ and x_i are independently and identically distributed (i.i.d.). Sobol' [8] developed the following formula to compute the first-order index:

$$V_i = \int f(\boldsymbol{x}) f(x_i, \boldsymbol{z}_{-i}) p(\boldsymbol{x}) p(\boldsymbol{z}_{-i}) d\boldsymbol{x} d\boldsymbol{z}_{-i} - E^2(y) \tag{4}$$

where $p(\bullet)$ denotes the joint probability density function (PDF) of all the arguments, and it is the product of the PDFs of individual arguments under the assumption of independent model inputs. $\boldsymbol{z}_{-i}$ are all the variables in $\boldsymbol{z}$ other than z_i.

Eq. (4) leads to the following estimator of V_i:

$$V_i = \frac{1}{n}\sum_{j=1}^{n} f\left(\boldsymbol{x}^j\right) f\left(x_i^j, \boldsymbol{z}_{-i}^j\right) - \left[\frac{1}{n}\sum_{j=1}^{n} f\left(\boldsymbol{x}^j\right)\right]^2 \tag{5}$$

Eq. (5) requires n_s samples of $\boldsymbol{x}$ and n_s samples of $\boldsymbol{z}$, which are sampled independently from the distributions of the model inputs. In Eq. (5), the superscript j is the index of the samples and the subscript i is the index of model inputs. For example, $\boldsymbol{x}^j$ means the j-th sample of $\boldsymbol{x}$, and $\boldsymbol{z}_{-i}^j$ means the j-th sample of $\boldsymbol{z}$ except z_i. In Eq. (5), $f(\boldsymbol{x}^j)$ implies n_s model evaluations; $f\left(x_i^j, \boldsymbol{z}_{-i}^j\right)$ implies n_s model evaluations for each model input, i.e., kn_s evaluations for all the model inputs. To improve the accuracy, generally another n_s model evaluations are needed over the samples in $\boldsymbol{z}$, and the results are used to estimate $V(y)$ together with earlier evaluations over $\boldsymbol{x}$. The first-order index is calculated as $S_i = V_i/V(y)$. The overall cost for all the first-order indices is $kn_s + 2n_s$.

Eq. (5) is the first efficient sample-based method to compute the first-order Sobol' index. Several methods have been proposed to improve its accuracy or reduce computational cost. Homma and Saltelli [21] suggested a more accurate estimator of V_i by using $\frac{1}{n}\sum_{j=1}^{n} f(\boldsymbol{x}^j) f(\boldsymbol{z}^j)$ to calculate $E^2(Y)$ instead of $\left[\frac{1}{n}\sum_{j=1}^{n} f(\boldsymbol{x}^j)\right]^2$. Thus Eq. (5) becomes [22]:

$$V_i = \frac{1}{n}\sum_{j=1}^{n} f\left(\boldsymbol{x}^j\right)\left[f\left(x_i^j, \boldsymbol{z}_{-i}^j\right) - f\left(\boldsymbol{z}^j\right)\right] \tag{6}$$

Compared to Eq. (5), Eq. (6) brings no extra model evaluation.