Environmental Modelling & Software 47 (2013) 55–63

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

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

Moment-independent regional sensitivity analysis: Application to an environmental model

Pengfei Wei, Zhenzhou Lu*, Danqing Wu, Changcong Zhou

School of Aeronautics, Northwestern Polytechnical University, Mail Box 120, Youyi West Road, Xi'an, Shaanxi 710072, China

ARTICLE INFO

Article history:
Received 18 August 2012
Received in revised form
25 April 2013
Accepted 7 May 2013
Available online 12 June 2013

Keywords:
Moment-independent
Contribution to delta indices plot
Delta ratio function
Uncertainty reduction
Level E model

ABSTRACT

Through several decades of development, global sensitivity analysis has been developed as a very useful guide tool for assessing scientific models and has gained pronounced attention in environmental science. However, standard global sensitivity analysis aims at measuring the contribution of input variables to model output uncertainty on average by investigating their full distribution ranges, but does not investigate the contribution of specific ranges. To deal with this problem, researchers have developed several regional sensitivity analysis techniques such as the contribution to sample mean and variance (CSM and CSV) plots. In this paper, a moment-independent regional sensitivity analysis technique called contribution to delta indices (CDI) plot is developed for assessing the effect of a specific range of an individual input to the uncertainty of model output. The CDI plot can be obtained with the same set of samples used for computing the CSM and CSV. Compared with the CSM and CSV, the CDI plot uses the probability density function shift of model output to describe the uncertainty instead of the mean and variance, thus it is moment-independent. An analytical linear model, the Ishigami function and an environmental model are employed to test the proposed RSA technique.

1. Introduction

Global sensitivity analysis (GSA) aims at measuring the contribution of input uncertainty to the model output uncertainty by exploring the whole distribution ranges of model inputs (Saltelli, 2002). It is a very useful tool for developing scientific models, thus has been widely used in environmental science (e.g., Draper et al., 1999; Pastres et al., 1999; Saltelli and Tarantola, 2002; Vezzaro and Mikkelsen, 2012).

The last few decades have witnessed a rapid development in GSA techniques (Sobol, 1993; Homma and Saltelli, 1996; Chun et al., 2000; Borgonovo, 2007; Sobol and Kucherenko, 2009; Liu and Homma, 2010; Wei et al., 2012). Among all the available GSA techniques, the variance-based one developed by Sobol (1993), Homma and Saltelli (1996) and the moment-independent one proposed by Borgonovo (2007) are the most widely used (e.g., Estrada and Diaz, 2010; Borgonovo et al., 2012; Dimov et al., 2012; Zhan et al., 2013). The variance-based GSA indices are “global, quantitative and model free” (Borgonovo, 2006), and various smart computational methods are available for quantitative analysis, thus have gained the most attention of analysts and practitioners. However, as has been pointed out by Borgonovo (2006), the premise of this technique—that the variance is sufficient to describe the variability of model output—is not always true. Comparably, the moment-independent indices (also called delta indices) are not only “global, quantitative and model free”, but also make no assumption on the independence among the input variables, thus have drawn growing attention in the past few years.

Standard GSA techniques (including the variance-based one and the moment-independent one) quantify the contributions of the input variables to the uncertainty of model output by exploring their whole distribution ranges, but do not tell which region of the distribution of an input variable (lower tail, central region or upper tail) contributes the most to the uncertainty of model output. In order to measure the regional importance of input variables, Bolado-Lavin et al. (2009) and Tarantola et al. (2012) proposed two regional sensitivity analysis (RSA) techniques called contribution to sample mean and variance (CSM and CSV, respectively) plots. These two techniques are able to measure the contribution of a specific region of an input variable to the mean and variance of model output. However, like the limit of the variance-based GSA technique, since both the mean and variance are not always sufficient to characterize the uncertainty of model output, the applications of CSM and CSV are limited.

In this paper, we propose a moment-independent RSA technique called contribution to delta indices (CDI) plot. After the

* Corresponding author.
E-mail addresses: wpf0414@163.com (P. Wei), zhenzhoulu@nwpu.edu.cn (Z. Lu), wudanqing86@163.com (D. Wu), changcongzhou@mail.nwpu.edu.cn (C. Zhou).

1364-8152/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2013.05.003

important input variables have been observed by using the delta indices, the CDI plots can be used for measuring the contributions of specific ranges of these important input variables to the uncertainty of model. The CDI plot has two definite advantages. First, it doesn't make any assumptions on model output moment, and use the shift of probability density function (PDF) to characterize the uncertainty of model output, thus it is moment-independent. Second, the CDI plots for all input variables can be computed by using the same set of samples as those for computing the delta indices, thus no extra computational cost is introduced.

2. Materials and methods

2.1. Moment-independent regional sensitivity analysis

2.1.1. Review of the delta indices

Consider a computational model represented by an input–output function $Y = g(\mathbf{X})$, where Y is the model output with PDF $f_Y(\boldsymbol{y})$, and $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ denotes the n-dimensional vector of random input variables with joint PDF $f_X(\boldsymbol{x})$. If the input variables are independent, then $f_X(\boldsymbol{x}) = \prod_{i=1}^{n} f_i(x_i)$, where $f_i(x_i)$ is the marginal PDF of the input variable X_i. In this paper, only the continuous random input variables are considered. Thus the support of each input is a range.

To investigate the effect of the uncertainty of each input variable X_i on the PDF of model output, Borgonovo (2007) proposed the following delta index for X_i,

$$\delta_i = \frac{1}{2} E_{X_i}(s(X_i)) = \frac{1}{2} \int s(X_i) f_{X_i}(x_i) \mathrm{d}x_i \tag{1}$$

where $s(X_i)$ is the area difference between the unconditional PDF $f_Y(\mathbf{y})$ and the conditional one $f_{Y|X_i}(y)$, i.e.,

$$s(X_i) = \int_{-\infty}^{+\infty} \left| f_Y(y) - f_{Y|X_i}(y) \right| \mathrm{d}y \tag{2}$$

The delta index for a set of input variables, say $\boldsymbol{R} = (x_{i_1}, x_{i_2}, \ldots, x_{i_r})$, is defined as

$$\begin{aligned}\delta_{i_1, i_2, \ldots, i_r} &= \frac{1}{2} E_{\mathbf{R}}(s(\mathbf{R})) = \frac{1}{2} \int f_{X_{i_1}, X_{i_2}, \ldots, X_{i_r}}(x_{i_1}, x_{i_2}, \ldots, x_{i_r}) \\ &\quad \times \left(\int \left| f_Y(y) - f_{Y|X_{i_1}, X_{i_2}, \ldots, X_{i_r}}(y) \right| \mathrm{d}y \right) \mathrm{d}x_{i_1} \mathrm{d}x_{i_2} \ldots \mathrm{d}x_{i_r}\end{aligned} \tag{3}$$

where $f_{Y|X_{i_1}, X_{i_2}, \ldots, X_{i_r}}(y)$ is the PDF of model output conditioned on $\mathbf{R}$, and $s(\mathbf{R})$ is the area difference between $f_Y(\mathbf{y})$ and $f_{Y|X_{i_1}, X_{i_2}, \ldots, X_{i_r}}(y)$.

Note that $f_{Y,X_i}(y, x_i) = f_{X_i}(x_i) f_{Y|X_i}(y)$, then Eq. (1) can be rewritten as (Plischke et al., 2013; Wei et al., 2013):

$$\delta_i = \frac{1}{2} \int \int \left| f_Y(y) f_{X_i}(x_i) - f_{Y,X_i}(y, x_i) \right| \mathrm{d}y \mathrm{d}x_i \tag{4}$$

Further, Eq. (4) can be expressed in terms of copula. Let $F_i(x_i)$ and $F_Y(y)$ denote the marginal CDF of X_i and Y, respectively, and $F_{Y,xi}(y,x_i)$ denote the joint CDF of X_i and Y_i. By Sklar's theory (Nelsen, 2006; Genest and Favre, 2007), if both $F_i(x_i)$ and $F_Y(y)$ are continuous, then there exists a unique copula C such that

$$F_{Y,X_i}(y, x_i) = C(F_Y(y), F_i(x_i)) \tag{5}$$

If $F_i(x_i)$ and $F_Y(y)$ are discontinuous marginal CDFs, then C is uniquely determined in the space $\mathrm{Ran}F_Y(y) \times \mathrm{Ran}F_{X_i}(x_i)$, where $\mathrm{Ran}F_Y(y)$ and $\mathrm{Ran}F_{X_i}(x_i)$ are the range of $F_Y(y)$ and $F_i(x_i)$, respectively, and $\times$ is the Cartesian product.

Let $u = F_Y(y)$ and $v_i = F_i(x_i)$. Both u and v_i are uniformly distributed in [0,1]. Then the copula $C(u,v_i)$ can be regarded as a joint CDF with uniform marginal distribution in [0,1]. The copula density $c(u,v_i)$ is given by

$$c(u, v_i) = \frac{\partial^2 C(u, v_i)}{\partial u \partial v_i} \tag{6}$$

Then Eq. (4) can be expressed by (see Appendix for proof)

$$\delta_i = \frac{1}{2} \int \int_{I^2} |c(u, v_i) - 1| \mathrm{d}u \mathrm{d}v_i \tag{7}$$

Eq. (7) indicates that δ_i equals the half of the volume difference between the copula density $z = c(u,v_i)$ and the plane $z = 1$ in the space $[0,1]^2$.

By Eq. (7), as copula density $c(u,v_i)$ has been estimated to be $\widehat{c}(u, v_i)$, δ_i can be estimated by numerically integrating the bivariate function $|\widehat{c}(u, v_i) - 1|/2$.

2.1.2. The contribution to delta indices plot

It is noticed that the delta index for X_i is defined by integrating $s(X_i)$ from $-\infty$ to $+\infty$. Similar to the definitions of CSM and CSV, by integrating $s(X_i)$ from $-\infty$ to $F_i^{-1}(q)$ with $q \in [0, 1]$, the contribution of the distribution range $(-\infty, F_i^{-1}(q)]$ of X_i to the model output uncertainty can be investigated. Therefore, the CDI plot for the input variable X_i is defined as:

$$\mathrm{CDI}_i(q) = \frac{1}{2\delta_i} \int_{-\infty}^{F_i^{-1}(q)} \left(\int_{-\infty}^{+\infty} \left| f_Y(y) - f_{Y|X_i}(y) \right| \mathrm{d}y \right) f_i(x_i) \mathrm{d}x_i \tag{8}$$

$\mathrm{CDI}_i(q)$ is plotted on the space $[0,1]^2$. After the important input variables are observed by the delta indices δ_i, the CDI plot can be used to investigate the effect of specific ranges of those important input variables on the uncertainty of model output.

By the copula density, we can rewrite Eq. (8) as

$$\mathrm{CDI}_i(q) = \frac{1}{2\delta_i} \int_0^q \int_0^1 |c(u, v_i) - 1| \mathrm{d}u \mathrm{d}v_i \tag{9}$$

If the copula density $c(u,v_i)$ has been properly identified by using a set of samples, then $\mathrm{CDI}_i(q)$ can be estimated using the same set of samples.

Compared with the mean and variance, the density shift is more proper for characterizing the uncertainty of model output since it contains the complete information of each moment. Similar to CSV, the following two properties are true for CDI: a) $\mathrm{CDI}_i(0) = 0$, $\mathrm{CDI}_i(1) = 1$; b) CDI is strictly non-decreasing function of q.

Given a set of N samples $(x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)})$ $(j = 1,2,\ldots,N)$ and the corresponding model output value $y^{(j)}$, the following procedure can be used for computing $\mathrm{CDI}_i(q)$:

Step 1: Convert the sample pair $(y^{(j)}, x_i^{(j)})$ to $(u^{(j)}, v_i^{(j)})$ $(j = 1,2,\ldots,N)$ by

$$u^{(j)} = \frac{1}{N} \sum_{k=1}^{N} I\left(y^{(k)} \le y^{(j)}\right), \quad v_i^{(j)} = F_i\left(x_i^{(j)}\right) \tag{10}$$

where $I(\cdot)$ is the indictor function, i.e., $I = 1$ if $y^{(k)} \le y^{(j)}$, else $I = 0$.

Step 2: Estimate the copula density $c(u,v_i)$ by the maximum penalized likelihood estimation (MPLE) procedure (Qu and Yin, 2012) or the kernel density estimation (KDE) procedure (Botev et al., 2010) or other density estimation methods. Denote the estimate as $\widehat{c}(u, v_i)$.

Step 3: Compute the delta indices $\widehat{\delta}_i$ by numerically integrating the function $|\widehat{c}(u, v_i) - 1|/2$ over the rectangular region [0,1]×[0,1].

Step 4: Compute $\mathrm{CDI}_i(q)$ by numerically integrating the function $|\widehat{c}(u, v_i) - 1|/2\widehat{\delta}_i$ over the rectangular region [0,1]×[0,q], where $0 < q < 1$.

2.1.3. Interpretation of CDI plot

CDI is a useful tool for investigating the contribution of specific distribution ranges of input variables to the model output uncertainty, and CDI can be interpreted in an analogous way to the CSM and CSV. If the CDI plot is close to the diagonal line, the contribution of the distribution range of the input variable to the model output is uniform. If the plot is convex downward, then this distribution range contributes less than the average level to the model output uncertainty. Otherwise, if the plot is convex upward, then the distribution range contributes more than the average level.

Next, we develop the relationship between the distribution range reduction of input variable and the mean shift of model output PDF. Suppose we have reduced the distribution range of the input X_i from $[-\infty, +\infty]$ to $[F_i^{-1}(s), F_i^{-1}(t)]$, then the PDF of the input X_i can be updated as

$$f_i^*(x_i) = \frac{f_i(x_i)}{\int_{F_i^{-1}(s)}^{F_i^{-1}(t)} f_i(x_i) \mathrm{d}x_i} \tag{11}$$

and the delta index for X_i with the updated PDF $f_i^*(x_i)$ is given by

$$\begin{aligned}\delta_i^{*[s,t]} &= \frac{1}{2} \int_{F_i^{-1}(s)}^{F_i^{-1}(t)} \left(\int_{-\infty}^{+\infty} \left| f_Y(y) - f_{Y|X_i}(y) \right| \mathrm{d}y \right) f_i^*(x_i) \mathrm{d}x_i \\ &= \frac{1}{2 \int_{F_i^{-1}(s)}^{F_i^{-1}(t)} f_i(x_i) \mathrm{d}x_i} \int_{F_i^{-1}(s)}^{F_i^{-1}(t)} \left(\int_{-\infty}^{+\infty} f_Y(y) - f_{Y|X_i}(y) \mathrm{d}y \right) f_i^*(x_i) \mathrm{d}x_i\end{aligned} \tag{12}$$

Eq. (12) defines the delta index for X_i in the reduced distribution range $[F_i^{-1}(s), F_i^{-1}(t)]$, but with the unchanged PDF $f_Y(y)$ of model output. By the copula density, Eq. (12) can be written as