Computer Physics Communications 185 (2014) 19–27

Contents lists available at ScienceDirect

Computer Physics Communications

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

A fast computational method for moment-independent uncertainty importance measure

Luo Xiaopeng, Lu Zhenzhou *, Xu Xin

School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Article history:
Received 22 November 2012
Received in revised form
30 March 2013
Accepted 11 August 2013
Available online 17 August 2013

Keywords:
Uncertainty importance analysis
Moment independence
Fast computational method

ABSTRACT

This work focuses on the fast computation of the moment-independent importance measure δ_i. We first analyse why δ_i is associated with a possible computational complexity problem. One of the reasons that we thought of is the use of two-loop Monte Carlo simulation, because its rate of convergence is $O(N^{-1/4})$, and another one is the computation of the norm of the difference between a density and a conditional density. We find that these problems are nonessential difficulties and try to give associated improvements. A kernel estimate is introduced to avoid the use of two-loop Monte Carlo simulation, and a moment expansion of the associated norm which is not simply obtained by using the Edgeworth series is proposed to avoid the density estimation. Then, a fast computational method is introduced for δ_i. In our method, all δ_i can be obtained by using a single sample set. From the comparison of the numerical error analyses, we believe that the proposed method is clearly helpful for improving computational efficiency.

1. Introduction

The main task of probabilistic importance analysis is to clarify and quantify the relationship between the distribution of the output Y and the distribution of an input X_i for a given model $Y = g(X_1, X_2, \ldots, X_d)$, where Y is a random variable and $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ is a d-dimensional random vector. Apostolakis [1] emphasized that the treatment of uncertainties can play a critical role in probabilistic risk assessment. And more and more works are demonstrating that uncertainty and sensitivity analyses have become an essential part of the modelling and risk assessment of complex systems [2–5].

Saltelli [2], Saltelli et al. [6], Rabitz et al. [7], Rabitz and Alis [8], Frey and Patil [9] and Sobol [10,11] have shown that the notion of conditional variance plays a very important role in probabilistic importance analysis. And the two most popular variance-based sensitivity measures are the main effect

$$S_i = \frac{\mathrm{Var}(E(Y|X_i))}{\mathrm{Var}(Y)}$$

and the total effect

$$ST_i = \frac{E(\mathrm{Var}(Y|\boldsymbol{X}_{-i}))}{\mathrm{Var}(Y)}$$

where $\boldsymbol{X}_{-i}$ indicates all input factors except X_i [6,12]. Actually, the main effect and total effect can be directly extended to the skewness and kurtosis forms. That is, the skewness-based and kurtosis-based main effects are defined by

$$S_i^s = \frac{\mathrm{Skew}(E(Y|X_i))}{\mathrm{Skew}(Y)}$$

and

$$S_i^k = \frac{\mathrm{Kurt}(E(Y|X_i))}{\mathrm{Kurt}(Y)}$$

while the associated total effects are defined by

$$ST_i^s = \frac{E(\mathrm{Skew}(Y|\boldsymbol{X}_{-i}))}{\mathrm{Skew}(Y)}$$

and

$$ST_i^k = \frac{E(\mathrm{Kurt}(Y|\boldsymbol{X}_{-i}))}{\mathrm{Kurt}(Y)}.$$

In practice, the two most popular sensitivity measures are the variance-based main effect and the variance-based total effect [6,13]. However, the use of the variance as a measure of uncertainty, as Borgonovo [2] mentioned, implicitly assumes that this moment is sufficient for describing output variability. And a distribution, as Saltelli [14] pointed out, is not uniquely determined by its variance; then "the total uncertainty should be represented by the entire output distribution and not by sole variance". To overcome this issue, he defined a new uncertainty importance measure δ_i by using the notion of conditional density [14] (or conditional distribution [15]).

* Corresponding author. Tel.: +86 29 88460480; fax: +86 29 88460480.
E-mail addresses: luo_works@163.com (X. Luo), zhenzhoulu@nwpu.edu.cn (Z. Lu), xuxin103@163.com (X. Xu).

0010-4655/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cpc.2013.08.006

However, we find that there is a possible computational complexity problem for index δ_i. Liu and Homma [16] introduced a computational method by using conditional distribution. And there are some other works on the computational methods for obtaining index δ_i [17–20]. But comparing against the efficiency of variance-based indexes, most of the improvements that we have seen are still not satisfactory. One of the advantages of variance-based indexes is that the computing efficiency is far better than that for δ_i.

Indeed, in some sense, the $\mathcal{L}^1$-norm, which is used in the moment-independent index, can be represented approximatively by finitely many moments; and the difference for all the moments between a density and a conditional density can also be represented by a norm. We think that, for the δ_i index, the so-called computational complexity is nonessential.

This problem usually arises for two reasons. The first is associated with the use of two-loop Monte Carlo simulation (MC) [14,16]; the second is associated with the computation of the norm of the density function [14] (or distribution function [15]). For the first one, the recent work [21] reported an approach for computing δ_i from a given sample without using a double loop. This study provides reliable evidence supporting our viewpoint that the use of two-loop MC is not due to the definition itself of δ_i. For the second one, it seems that this is totally due to the definition itself of δ_i, because there is an $\mathcal{L}^1$-norm used in its definition. But, as we noted above, the $\mathcal{L}^1$-norm can be represented approximatively by finitely many moments. So we try to give associated improvements in this paper, and we believe that the proposed method is clearly helpful for improving the computational efficiency, in view of the following analysis and comparison.

The remainder of this paper is organized as follows. Section 2 provides a mathematical analysis of both reasons noted above and associated improvements. Section 3 introduces a fast computational method for obtaining index δ_i, which can give all δ_i by using a single sample set. And we provide example analysis with a comparison of numerical error analyses in Section 4. Conclusions are offered in Section 5.

2. The problem statement and its improvement

2.1. The definition of δ_i

Let $Y = g(X_1, X_2, \ldots, X_d)$ be the mathematical model, where Y is a random variable with density f_y (if it exists), and X_i ($i = 1, 2, \ldots, d$) is a random variable with density f_i. Let $f_{y|x_i}$ be the conditional density of Y on X_i (if it exists), and

$$\mathfrak{d}(x_i) = \|f_y - f_{y|x_i}\|_1 \tag{1}$$

where $\|\cdot\|_1$ is the norm of $\mathcal{L}(-\infty, +\infty)$, which consists of all the integrable functions on $(-\infty, +\infty)$. Then $\mathfrak{d}(x_i)$ is a function of the random variable X_i, and the expectation of $\mathfrak{d}(x_i)$ is

$$\mathbb{E}_i(\mathfrak{d}(x_i)) = \int_{-\infty}^{+\infty} \|f_y - f_{y|x_i}\|_1 f_i(x_i)\mathrm{d}x_i. \tag{2}$$

Borgonovo's index δ_i ($i = 1, 2, \ldots, d$) [14] is defined as

$$\delta_i = \frac{1}{2}\mathbb{E}_i(\mathfrak{d}(x_i)). \tag{3}$$

2.2. The problem statement

We think that the possible computational complexity problem for index δ_i is associated with the use of two-loop MC and the computation of the norm of the density function (or distribution function).

For two-loop MC, it is clear that the order of convergence (or rate of convergence) is $O(N^{-1/4})$. By this we mean that its error tends to 0 as fast as $cN^{-1/4}$ if N is large enough, where c is a nonzero constant and N is the sample size. Roughly speaking, if $\hat{p}$ is an estimate of p defined by

$$\hat{p}_N = \frac{1}{N}\sum_{l=1}^{N}\left[\frac{1}{N}\sum_{s=1}^{N} p_{sl}\right], \quad \lim_{N\to\infty}\hat{p}_N = p \tag{4}$$

where p_{sl} is the value for fixed s and l, then, from the central limit theorem, we have

$$\lim_{N\to\infty} N^{1/4}(\hat{p}_N - p) = c < \infty. \tag{5}$$

If there is a sample size $N = n_1$ such that $|\hat{p}_{n_1} - p| \leqslant \epsilon_1$, then the error estimate decreases from ϵ_1 to $\epsilon_2 = \epsilon_1/10$ if and only if the sample size increases from n_1 to $n_2 = 10^4 n_1$. So the rate of convergence is quite slow.

By the way, we say that the rate is slower than $N^{-1/4}$ if

$$\lim_{N\to\infty} N^{1/4}(\hat{p}_N - p) = \infty$$

and is faster than $N^{-1/4}$ if

$$\lim_{N\to\infty} N^{1/4}(\hat{p}_N - p) = 0.$$

For the computation of the norm, we think that $\|f_y - f_{y|x_i}\|_1$ should be treated as a whole quantity but not an integral of the difference between f_y and $f_{y|x_i}$.

If $\|f_y - f_{y|x_i}\|_1$ is viewed as an integral of the difference of two density functions f_y and $f_{y|x_i}$, then we must first get two approximations of f_y and $f_{y|x_i}$ before computing the integral of the difference. Indeed, as a well-known result, for an approximation of density at a fixed point, the rate of convergence is always slower than $N^{-1/2}$ [22,23]; even if we can compute this integral by approximation of distribution, the rate of convergence could always be faster than $N^{-1/2}$ but still not satisfactory.

Hence, we should avoid the use of two-loop MC and the computation of an integral of the difference of two approximations. This starting point is also in agreement with the findings in [21].

2.3. Improvement 1. A kernel estimate

Before we consider the computation of $\mathfrak{d}(x_i)$ for every fixed x_i, we first discuss a kernel estimate of $\int e(y)[f_y(y) - f_{y|x_i=t}(y)]\mathrm{d}y$ for any function $e(y)$. In order to compute this integral which is defined in the range of Y, we need to transform it to another defined in the range of X. We shall show that the dimension of $D \subset \mathbb{R}^d$ is different from the dimension of $D_i \subset \mathbb{R}^{d-1}$, where D and D_i are the domains of f_x and $f_{x|x_i=t}$, respectively. And we shall extend $f_{y|x_i=t}$ defined in D_i to another form defined in D by a kernel function, which can help us to avoid the use of two-loop MC.

The conditional density $f_{y|x_i=t}$ is determined by the model g and

$$f_{x|x_i=t} = f(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_d) \tag{6}$$

defined on $\mathbb{R}^{d-1}$, that is, for any $S \subset \mathbb{R}$,

$$\int_S f_{y|x_i=t}(y)\mathrm{d}y = \int_{g^{-1}_{x_i=t}(S)} f_{x|x_i=t}(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}' \tag{7}$$

where $\boldsymbol{x}' = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$, $\mathrm{d}\boldsymbol{x}' = \mathrm{d}x_1 \ldots \mathrm{d}x_{i-1}\mathrm{d}x_{i+1} \ldots \mathrm{d}x_d$, and $g^{-1}_{x_i=t}(S) = \{\boldsymbol{x}' \in \mathbb{R}^{d-1} : g_{x_i=t}(\boldsymbol{x}') \in S\}$.

Similarly, density f_y is determined by the model g and

$$f_x = f(x_1, \ldots, x_d) \tag{8}$$

defined on $\mathbb{R}^d$, that is, for any $S \subset \mathbb{R}$,

$$\int_S f_y(y)\mathrm{d}y = \int_{g^{-1}(S)} f_x(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{9}$$

where $g^{-1}(S) = \{\boldsymbol{x} \in \mathbb{R}^d : g(\boldsymbol{x}) \in S\}$.