Reliability Engineering and System Safety 107 (2012) 149–156

Contents lists available at SciVerse ScienceDirect

Reliability Engineering and System Safety

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

An adaptive correlation ratio method using the cumulative sum of the reordered output

Elmar Plischke

Institute of Disposal Research, Clausthal University of Technology, 38678 Clausthal-Zellerfeld, Germany

ARTICLE INFO

Article history:
Received 14 October 2010
Received in revised form
11 May 2011
Accepted 2 December 2011
Available online 13 December 2011

Keywords:
Global sensitivity analysis
Sobol' index
Correlation ratio

ABSTRACT

We consider correlation ratios as estimators for first order sensitivity indices from given data. The computation is simplified by the introduction of the cumulative sum of the normalised reordered output. Ideas for the estimation using interpolation are also discussed.

1. Introduction

In 1905 Karl Pearson [10] introduced the correlation ratio $\hat{\eta}^2$ (CR) as a measure for the non-linear influence of a random vector $\mathbf{X}$ on a random variable Y especially for cases where linear regression produces only small R^2 values. Kolmogorov [5] later identified the CR as an estimator of $\eta^2 = \mathbb{V}[Y]^{-1}\mathbb{V}[\mathbb{E}[Y|\mathbf{X}]]$. In recent years, this quotient has received lots of attention in sensitivity analysis and keeps reappearing under many different names, e.g., first order sensitivity index, main effect, Sobol' index [15]. With this growing interest in variance-based sensitivity indicators we re-investigate the correlation ratio measure.

In the sensitivity analysis for model outputs, it is assumed that the *output* Y is given by a computationally demanding numerical simulation model $Y = f(\mathbf{X})$, depending only on the *input* vector $\mathbf{X}$ which has a known (multi-dimensional) probability distribution. For this paper, let us assume that the given data includes both the information about the input uncertainties and the input/output mapping so that we have no direct access to the simulation model or the sampling procedure. Therefore, the proposed algorithm is a post-processing method.

We develop a graphical representation of the data which is closely related to the contribution to the sample mean (CSM) plot [1] and derive methods of estimating the main effect η^2 from that graphical representation. This answers also the question of the relation between CSM and CR raised in [1].

2. Setup

Let Y be a random variable and $\mathbf{X}$ be a random vector of dimension ℓ. The sensitivity of Y on $\mathbf{X}$ can be expressed in the following index:

$$\eta^2 = \frac{\mathbb{V}[\mathbb{E}[Y|\mathbf{X}]]}{\mathbb{V}[Y]} \tag{1}$$

where $\mathbb{V}[Y]$ denotes the variance of Y and $\mathbb{E}[Y|\mathbf{X}]$ is the conditional expectation of Y given $\mathbf{X}$. The term $\mathbb{V}[\mathbb{E}[Y|\mathbf{X}]] = \mathbb{E}[(\mathbb{E}[Y] - \mathbb{E}[Y|\mathbf{X}])^2]$ is the variance of the conditional expectation of Y given $\mathbf{X}$.

The main effect η^2 is the fraction of the variance of the output Y attributed to a functional dependency on the input $\mathbf{X}$. In this note we study the one-dimensional case $\ell = 1$.

In order to compute η^2 we need to estimate $\mathbb{E}[Y|\mathbf{X}]$, the nonparametric regression curve for which there are many techniques available [22]. In sensitivity analysis, approaches that are discussed include piecewise constant functions (correlation ratios), piecewise linear functions or splines, regression models with orthogonal function spaces, e.g., harmonic functions (effective algorithm for sensitivity indices [11]), polynomials (high dimensional model representation [12]; polynomial chaos expansion [21,7]), and weighted moving averages [4]. More regression-based techniques are studied in [19,20].

Furthermore, many algorithms compute η^2 directly, e.g., Fourier Amplitude Sensitivity Test [3,14], Sobol's Method [16,18], or Random Balance Design [23], by using special sampling schemes for $\mathbf{X}$. Hence these methods cannot be used directly as estimators working on given data. Instead, an intermediate meta-model is created from the data (Gaussian Emulator, [9]), and then the (emulated) output with respect to a specially designed sample

E-mail address: elmar.plischke@tu-clausthal.de

0951-8320/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ress.2011.12.007

can be evaluated at virtually no additional costs using this meta-model. The resulting input/output sample is then processed by the associated sensitivity algorithm.

In this paper we investigate the estimation of η^2 from given data without a meta-model layer. For this approach, a sample of n realisations of $\mathbf{X}$, $x=(x_i)_{i=1,\dots,n}$ is given. The corresponding realisations of Y, the output sample, are given by $y=(y_i)_{i=1,\dots,n}$. For the CR method with piecewise constant approximations we partition the input sample x into q disjoint subsample sets $\mathcal{X}_r$, $r=1,\dots,q$. The term $\mathbb{E}[Y|\mathbf{X}=x]$ used for evaluating (1) is then replaced by $\mathbb{E}[Y|\mathbf{X}\in\mathcal{X}_r]$. An estimate of the first order effect is obtained from

$$\overline{y}=\frac{1}{n}\sum_{j=1}^{n}y_j,\quad \overline{y}_r=\frac{1}{n_r}\sum_{x_j\in\mathcal{X}_r}y_j,\quad n_r=\sum_{x_j\in\mathcal{X}_r}1,$$

$$\hat{\eta}^2=\frac{\sum_{r=1}^{q}n_r(\overline{y}_r-\overline{y})^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}. \tag{2}$$

Here the value $\overline{y}$ denotes the mean, and the values $\overline{y}_r$ are the local means estimating $\mathbb{E}[Y|\mathbf{X}\in\mathcal{X}_r]$. An alternative formulation of (2) is available using the empirical local variances $s_r^2=(n_r-1)^{-1}\sum_{x_j\in\mathcal{X}_r}(y_j-\overline{y}_r)^2$ which then reads

$$\hat{\eta}^2=1-\frac{\sum_{r=1}^{q}(n_r-1)s_r^2}{\sum_{j=1}^{n}(y_j-\overline{y})^2}. \tag{3}$$

This follows from applying the sampled version of the variance decomposition formula $\mathbb{V}[Y]=\mathbb{E}[\mathbb{V}[Y|\mathbf{X}]]+\mathbb{V}[\mathbb{E}[Y|\mathbf{X}]]$,

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(y_i^2-\overline{y}^2)=\sum_{r=1}^{q}n_r(\overline{y}_r-\overline{y})^2+\sum_{r=1}^{q}(n_r-1)s_r^2, \tag{4}$$

to (2). In particular, rewriting $\sum_{j=1}^{n}(y_j-\overline{y})^2$ as $(\sum_{j=1}^{n}y_j^2)-n\overline{y}^2$, $\sum_{x_j\in\mathcal{X}_r}(y_j-\overline{y}_r)^2$ as $(\sum_{x_j\in\mathcal{X}_r}y_j^2)-n_r\overline{y}_r^2$, and $\sum_{r=1}^{q}n_r(\overline{y}-\overline{y}_r)^2$ as $(\sum_{r=1}^{q}n_r\overline{y}_r^2)-n\overline{y}^2$ and combining these results gives (4).

Unfortunately, formulas (2) and (3) give no clue on how to partition the data to produce optimal results. Some authors [6] suggest to use a partition size of $q=\lfloor\sqrt{n}\rfloor$, the integer part of the square root of n, so that each of the q subsamples contains roughly q realisations. It is not clear if this choice is optimal.

3. Visualisation

One approach of visualising input/output data is to use a scatterplot of $(\mathbf{X},Y)$ data pairs and to draw the regression curve through the data. For example, in Fig. 1 we used 200 simulations partitioned into 15 subsamples from the Ishigami test function [15]

$$Y=\sin X_1+7\sin^2 X_2+0.1X_3^4\sin X_1 \tag{5}$$

where $X_i\sim U(-\pi,\pi)$ are uniformly distributed in $[-\pi,\pi]$. This function has three input parameters, parameter 4 does not enter into the calculations and is used here as a dummy parameter. The curve $\mathbb{V}[\mathbb{E}[Y|\mathbf{X}=x]]$ is approximated by $\overline{y}_r$ for $x\in\mathcal{X}_r$ and an estimate of η^2 is then obtained by (2).

It is unclear if the chosen partition yields good results: the functional dependence of y on x_2 is resolved with the step-wise approximation of a period-two function while the influence of x_1 on y produces a not so impressive step-wise approximation of a period-one function. Here, one might need finer intervals to resolve fast changes in a better way. However, for this step more data are needed. For properly identifying the zero influences of x_3 and x_4 we actually should have used large intervals such that $\overline{y}_r\approx\overline{y}$. While the influence on x_4 is by choice purely random, x_3 gives a "structured zero" with large variation at the boundaries. This is an example of a non-functional influence on the output. A sensitivity measure which is able to detect such influences is discussed in [2]. The next section also offers a visual method for the output variance being influenced by input parameters.

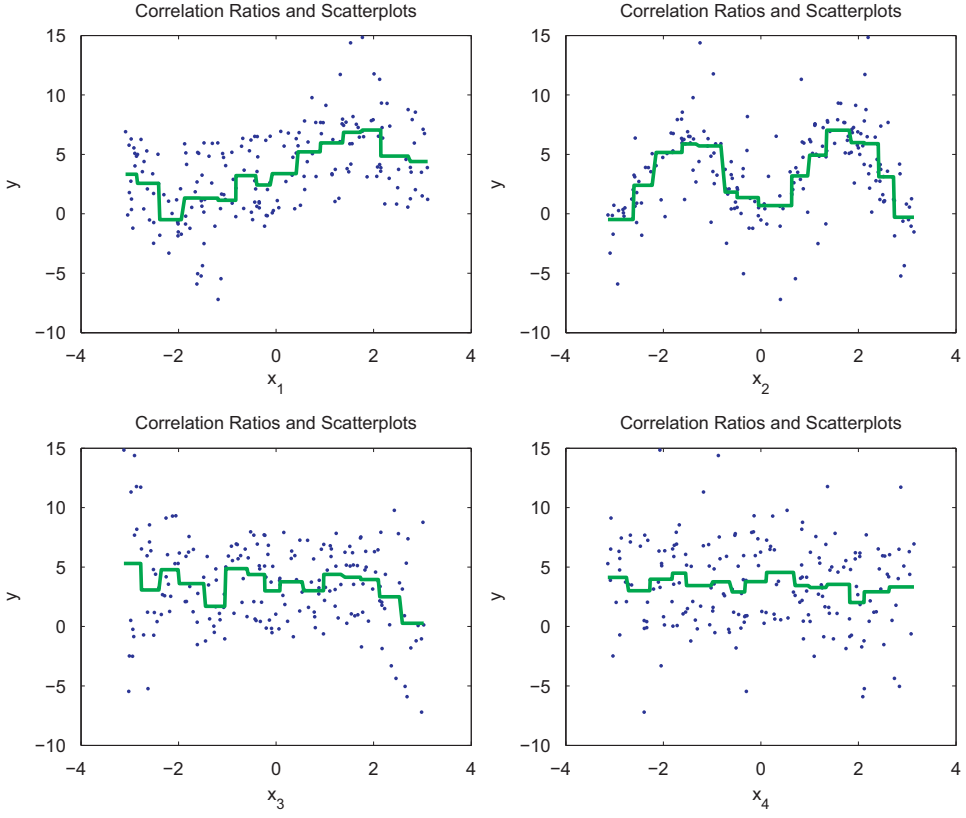

Fig. 1. Visual inspection of the regression curves for the Ishigami function.