Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 2 (2010) 7722–7723

Procedia Social and Behavioral Sciences

Sixth International Conference on Sensitivity Analysis of Model Output

An adaptive correlation ratio method

Elmar Plischke[a]1*

[a]*Insitut für Endlagerforschung, Technische Universität Clausthal, 38678 Clausthal-Zellerfeld, Germany*

Abstract

Pearson (1905) introduced the correlation ratio η^2 as a measure for the non-linear influence of an input random vector X on an output random variable Y especially for cases in which linear regression produces only small R^2 values.
In this note we develop a graphical representation of the data which is closely related to the contribution to the sample mean plot (Bolado-Lavin et al. (2009)) and derive methods of estimating the correlation ratio from that graphical representation.
Keywords: Global Sensitivity Analysis; Sobol' index, Correlation Ratio, adaptive method

1. Main text

Let Y be a random variable and $\boldsymbol{X}$ be a random vector of dimension ℓ. The sensitivity of Y on $\boldsymbol{X}$ is expressed in the following index, named *correlation ratio*

$$\eta^2 = \frac{\mathrm{Var}\big[\mathrm{E}[Y\,|\boldsymbol{X}]\big]}{\mathrm{Var}[Y]}$$

where $\mathrm{Var}[Y]$ denotes the variance of Y and $\mathrm{E}[Y|\boldsymbol{X}]$ is the conditional expectation of Y given $\boldsymbol{X}$. The conditional variance of Y given $\boldsymbol{X}$ is given by the term $\mathrm{Var}[Y|\boldsymbol{X}] = \mathrm{E}[(Y - \mathrm{E}[Y|\boldsymbol{X}])^2|\boldsymbol{X}]$. Conditional expectation and conditional variance are random variables of $\boldsymbol{X}$.

In this note we study the one-dimensional case $\ell = 1$. One-dimensional correlation ratios are also termed *Sobol' indices, first order effects* or *main effects*. This main effect is the fraction of variance of Y attributed to a functional dependency on $\boldsymbol{X}$. In order to compute η^2 we need to estimate $\mathrm{E}[Y|\boldsymbol{X}]$ which is the nonparametric regression curve. This may be accomplished by using piecewise constant or linear functions, or using regressions with suitable function spaces (e.g., Fourier transformations, Legendre polynomials, wavelet decompositions).

In this note, piecewise constant approximations are used. Unfortunately, up to now there are no methods available which automatically partition the data to produce optimal results.

* Corresponding author. Tel.: +49-5232 72 4921; fax: +49-5232-72-2810.
E-mail address: elmar.plischke@tu-clausthal.de.

doi:10.1016/j.sbspro.2010.05.197

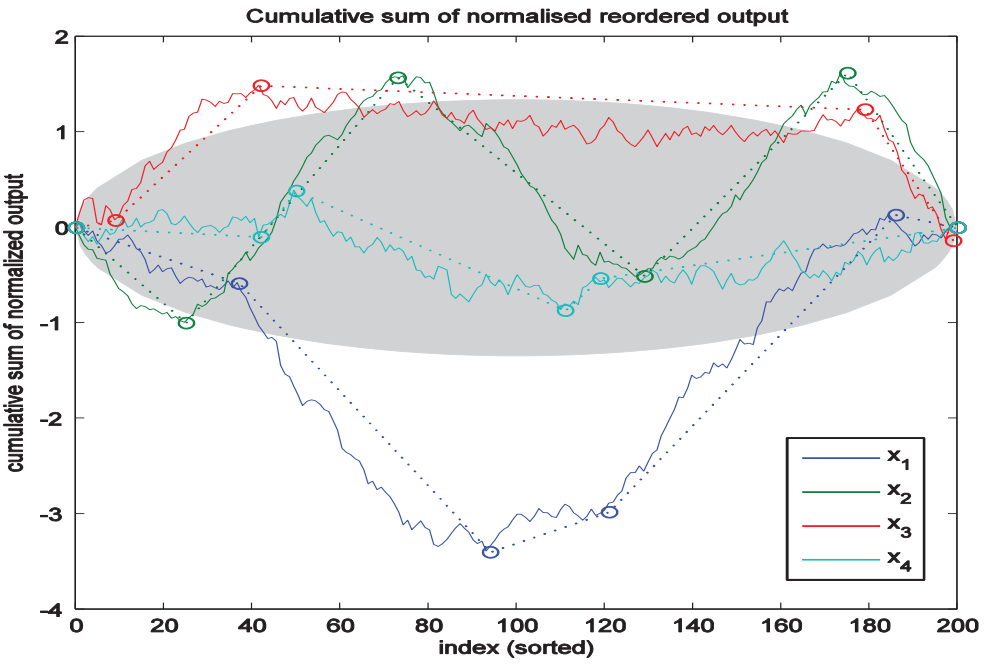

Figure 1: Ishigami test function, cumulative sums and significance region

The first approach of visualising the data is to use a scatter-plot of $(\boldsymbol{X},Y)$ data pairs and to draw a regression curve through these data. An estimate of η^2 is then given by the sum of squares of the centered predicted values divided by the sum of squares of the centered output data.

An alternative method of presenting the data leads more naturally to a formula for CR estimation. This method draws heavily from ideas of the contribution to the sample mean (CSM) plot. We use the cumulative sums of the normalised reordered output, as follows. For the samples x and y with n realisations of the random variables $\boldsymbol{X}$ and Y let $(z_i) = \left(\sqrt{n-1}s_y\right)^{-1}(y_{\pi(i)} - \bar{y})$ be the normalised output where the variance of Y is estimated by $s_y^2 = (n-1)^{-1}\sum_{i=1}^{n}(y-\bar{y})^2$and the permutation π is such that $(x_{\pi(i)}) = (x_{(i)})$ is the ordered statistics of the input of interest x. We define the cumulative sum $z^+(i) = \sum_{j=1}^{i} z_j$ (by convention, $z^+(0) = 0$). The ranks of the input x can now be plotted against the cumulative sum z^+, see Figure 1.

Given an ordered set of indices $\{j_r;\ r = 0, 1, \dots, q\}$ with $j_0 = 0$, $j_i < j_{i+1}$, $j_q = n$ we obtain an estimate of the first order effect by forming a weighted sum of squares of difference quotients of z^+,

$$\hat{\eta}^2 = \sum_{r=1}^{q}(j_r - j_{r-1})\left(\frac{z^+(j_r) - z^+(j_{r-1})}{j_r - j_{r-1}}\right)^2 = \sum_{r=1}^{q}\frac{\left(z^+(j_r) - z^+(j_{r-1})\right)^2}{j_r - j_{r-1}}.$$

When optimising the partition layout, minima and maxima of z+ are promising candidates since then the difference quotients will become large. To deal with multiple local extrema, we propose an algorithm of selecting suitable indices. In the paper, this will be discussed in detail.

As one of the further topics, a significance region approach is proposed, so that a curve outside this region yields a significant main effect.

2. References

Bolado-Lavin R., Castaings W., and Tarantola S., 2009:. Contribution to the sample mean plot for graphical and numerical sensitivity analysis. *Reliability Engineering&System Safety*, 94(6), 1041-1049.

Pearson K, 1905: On the General Theory of Skew Correlation and Non-linear Regression, Drapers' Company Research Memoirs. Dulau & Co., London.