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

Elmar Plischke[a]

[a] *Institute of Disposal Research, Clausthal University of Technology, 38678 Clausthal-Zellerfeld, Germany*

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.

Key words: Global Sensitivity Analysis; Sobol' Index; Correlation Ratio.

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].

Email address: elmar.plischke@tu-clausthal.de (Elmar Plischke)

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 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 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

$$\begin{gathered} \bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j, \qquad \bar{y}_r = \frac{1}{n_r}\sum_{x_j \in \mathcal{X}_r} y_j, \qquad n_r = \sum_{x_j \in \mathcal{X}_r} 1, \\ \hat{\eta}^2 = \frac{\sum_{r=1}^{q} n_r (\bar{y}_r - \bar{y})^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2}. \end{gathered} \tag{2}$$

Here the value $\bar{y}$ denotes the mean, and the values $\bar{y}_r$ are the local means estimating $\mathbb{E}[Y|\mathbf{X} \in \mathcal{X}_r]$. An alternative formulation of (2) is available using

the the empirical local variances $s_r^2 = (n_r - 1)^{-1} \sum_{x_j \in \mathcal{X}_r} (y_j - \bar{y}_r)^2$ which then reads

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

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

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

to (2). In particular, rewriting $\sum_{j=1}^{n} (y_j - \bar{y})^2$ as $\left(\sum_{j=1}^{n} y_j^2\right) - n\bar{y}^2$, $\sum_{x_j \in \mathcal{X}_r} (y_j - \bar{y}_r)^2$ as $\left(\sum_{x_j \in \mathcal{X}_r} y_j^2\right) - n_r \bar{y}_r^2$, and $\sum_{r=1}^{q} n_r (\bar{y} - \bar{y}_r)^2$ as $\left(\sum_{r=1}^{q} n_r \bar{y}_r^2\right) - n\bar{y}^2$ 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 scatter-plot of $(\mathbf{X}, Y)$ data pairs and to draw the regression curve through the data. For example, in Figure 1 we used 200 simulations partitioned into 15 subsamples from the Ishigami test function [15]

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

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 $\bar{y}_r$ 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 $\bar{y}_r \approx \bar{y}$. 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.

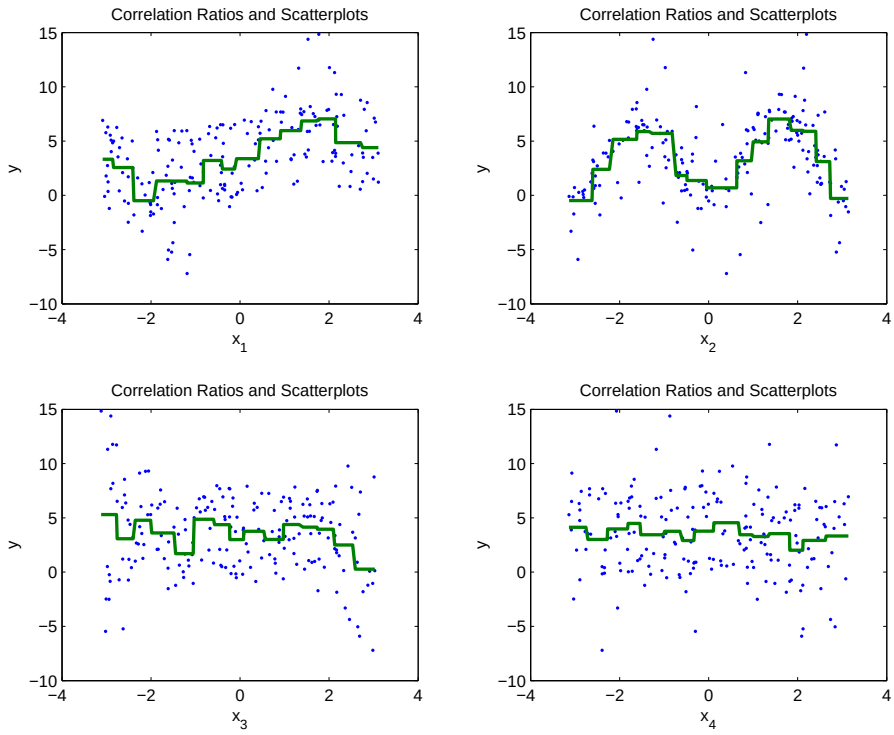

Figure 1: Visual inspection of the regression curves for the Ishigami function.

4. Cumulative sums of the normalised reordered output

Let us investigate a different method of presenting the data which more naturally leads to a CR estimation and which also allows for an adaptive partition layout. This method draws heavily from ideas of the contribution to the sample mean (CSM) plot, however, circumvents some of the problems which CSM has with non-positive data or with an output mean of zero.

Let π denote the permutation so that $(x_{\pi(i)}) = (x_{(i)})$ is the order statistics of the input of interest x, i.e., $x_{(i)} \leq x_{(i+1)}$ for all $i = 1, 2, \ldots, n-1$. Now, using the square root of the sum of squares $s_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ as a scaling factor we define $(\tilde{y}_i) = s_{yy}^{-1/2}(y_{\pi(i)} - \bar{y})$ to be the normalised reordered output. Now consider the scaled cumulative sums of $\tilde{y}$

$$z(i) = \frac{1}{\sqrt{n}} \sum_{j=1}^{i} \tilde{y}_j = \frac{1}{\sqrt{n \cdot s_{yy}}} \sum_{j=1}^{i} \left(y_{\pi(j)} - \bar{y} \right). \tag{6}$$

The empirical cumulative distribution function of the input x, the ranks of the input x, or the reordered input $(x_{(i)})$ itself can now be plotted against the cumulative sums z, see Figure 2. For a suitable abbreviation of this type of

plot, we suggest the term CUSUNORO as a short-hand for "**cu**mulative **su**ms of **no**rmalised **r**eordered **o**utput".

By convention, the empty sum yields 0 so that $z(0) = 0$ holds. Due to the renormalisation process we also have $z(n) = 0$. Moreover, CUSUNORO is shift- and scale-invariant. The factor $n^{-1/2}$ adjusts the CUSUNORO curve for different sample sizes. Although z is a vector, we prefer to write it as a function. The rationale behind this choice will become clearer in Section 7.

Let us now study the relation between CSM [1] and the CUSUNORO z. For this let us recall the definition of the contribution to the sample mean. Given a set of paired data $(x_i, y_i)_{i=1,\dots,n}$ with $y_i \geq 0$ (replace y_i with $y_i - \min_i y_i$, if needed) and $\bar{y} > 0$ the CSM plot is the graph of

$$\mathrm{csm} : \{0, 1, \dots, n\} \to [0, 1], \quad k \mapsto \frac{\sum_{i \leq k} y_{\pi(i)}}{n\bar{y}} = \frac{k\bar{y}_k}{n\bar{y}}$$

where $\bar{y}_k = k^{-1} \sum_{i=1}^{k} y_{\pi(i)}$ is the k-left mean and, again, π is the permutation associated with the order statistics of x. As y is non-negative, csm is monotonously increasing, with $\mathrm{csm}(0) = 0$ and $\mathrm{csm}(n) = 1$. From

$$\sqrt{\frac{s_{yy}}{n}} z(k) = \frac{1}{n} \sum_{i \leq k} (y_{\pi(i)} - \bar{y}) = \frac{k}{n} (\bar{y}_k - \bar{y}) = \bar{y} \left(\mathrm{csm}(k) - \frac{k}{n} \right) \tag{7}$$

it is clear to see that both CSM and CUSUNORO curves are different ways of visualising the same information. The assumptions used for CSM are valid in case of squared data, so let us consider the contribution to the sample variance (CSV) [24],

$$\mathrm{csv} : i \mapsto v(i) = s_{yy}^{-1} \sum_{j=1}^{i} \left(y_{\pi(j)} - \bar{y} \right)^2 \tag{8}$$

which will allow us to estimate the conditional variances $\mathbb{V}[Y|\mathcal{X}_r]$. Figure 2 also shows the CSV for the Ishigami example. Note that parameter 3 heavily influences the sample variance of the output, a fact which is not visible when only considering first order effects.

5. Correlation ratio estimates

We will see that the CR is related to the gradients of the CUSUNORO curves z (6). However, these curves are non-smooth and therefore estimates of the gradients are hard to obtain which capture the "deterministic" trends without over-emphasising the uncertainty in the data.

5.1. Partitions of size two

Given an index $\kappa \in \{1, \dots, n-1\}$ we consider a CR estimate based on a two-interval partition $q = 2$ of (2) with $n_1 = \kappa$ and $n_2 = n - \kappa$ given by

$$\hat{\eta}_\kappa^2 = \frac{1}{s_{yy}} \left(\kappa \left(\bar{y}_\kappa - \bar{y} \right)^2 + (n - \kappa) \left(\bar{y}_{\sim\kappa} - \bar{y} \right)^2 \right) \tag{9}$$

with the κ-left mean given by $\bar{y}_\kappa = \kappa^{-1}\sum_{i\le\kappa} y_{\pi(i)}$ and the κ-right mean given by $\bar{y}_{\sim\kappa} = (n-\kappa)^{-1}\sum_{i>\kappa} y_{\pi(i)}$. Since (7) yields $(ns_{yy})^{1/2} z(\kappa) = \kappa(\bar{y}_\kappa - \bar{y})$ and the definitions give $\kappa(\bar{y}_\kappa - \bar{y}) = (n-\kappa)(\bar{y} - \bar{y}_{\sim\kappa})$, (9) can be written using z,

$$\hat{\eta}^2_\kappa = \frac{1}{s_{yy}}\left(\frac{\kappa^2(\bar{y}_\kappa - \bar{y})^2}{\kappa} + \frac{(n-\kappa)^2(\bar{y}_{\sim\kappa} - \bar{y})^2}{n-\kappa}\right) = z(\kappa)^2\left(\frac{n}{\kappa} + \frac{n}{n-\kappa}\right). \quad (10)$$

The estimate (10) will become large if $|z|$ attains its global maximum in κ. But if $z(\kappa) = 0$ then the associated estimate $\hat{\eta}^2_\kappa$ vanishes. Hence a CUSUNORO curve with many (unstructured) zero-crossings is likely to be produced from a non-influential input.

5.2. Arbitrary partitions

Consider the index list $\mathcal{J} = \{j_0 = 0, \ldots, j_r, \ldots j_q = n\}$ and let the partiton of $\mathcal{X}$ be given by the half-open intervals $\mathcal{X}_1 = (-\infty, x_{(j_r)}]$ $\mathcal{X}_r = (x_{(j_r-1)}, x_{(j_r)}]$, $r = 2, \ldots, q-1$, $\mathcal{X}_q = (x_{(j_q-1)}, \infty)$. Then, using (6) and (8) the terms in (2) can be rewritten as

$$n_r = j_r - j_{r-1}, \qquad \bar{y}_r - \bar{y} = \frac{\sqrt{ns_{yy}}}{n_r}\left(z(j_r) - z(j_{r-1})\right),$$
$$s^2_r = \frac{s_{yy}}{n_r - 1}\left(v(j_r) - v(j_{r-1}) - \tfrac{n}{n_r}\left(z(j_r) - z(j_{r-1})\right)^2\right).$$

We obtain an estimate of the first order effect by forming a weighted sum of squares of difference quotients of z,

$$\hat{\eta}^2_{\mathcal{J}} = s^{-1}_{yy}\sum_{r=1}^{q} n_r(\bar{y}_r - \bar{y})^2 = \sum_{r=1}^{q}\frac{n}{n_r}\left(z(j_r) - z(j_{r-1})\right)^2$$
$$= \sum_{r=1}^{q}\frac{(z(j_r) - z(j_{r-1}))^2}{j_r/n - j_{r-1}/n} = \sum_{r=1}^{q}(j_r/n - j_{r-1}/n)\left(\frac{z(j_r) - z(j_{r-1})}{j_r/n - j_{r-1}/n}\right)^2. \quad (11)$$

Hence, the curve $z(\cdot)$ of (6) is replaced with an interpolating polygonal line and then the sum of the squared gradients weighted by the line segment length is computed. Here j/n is the empirical cumulative distribution function $\hat{F}_{\mathbf{X}}(x_{(j)})$.

To see that (9) and (11) yield the same result, consider $\mathcal{J} = \{0, j, n\}$. Then by (11),

$$\hat{\eta}^2_{\mathcal{J}} = \frac{j}{n}\left(\frac{z(j) - 0}{j/n - 0}\right)^2 + \frac{n-j}{n}\left(\frac{0 - z(j)}{n/n - j/n}\right)^2 = z(j)^2\left(\frac{n}{j} + \frac{n}{n-j}\right)$$

which gives (9).

5.3. An adaptive partition layout

When optimising the partition layout, the indices corresponding to minima and maxima of z are promising candidates as then the difference quotients in (11) are enlargened. As there is no need to fully reconstruct the CUSUNORO curve, we suggest the following algorithm which selects the locations of local extrema as suitable indices.

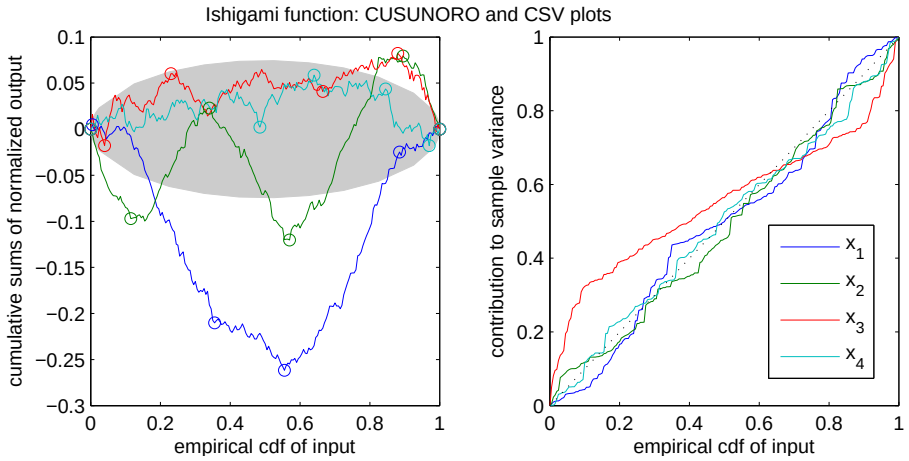

Figure 2: Compactly visualising the sensitivity of the data.

1. Create a working copy w of the CUSUNORO z (6).
2. Find the global extrema of w. Add the indices $j_1 < j_2$ where these extrema are attained to the list of indices $\mathcal{J}$ for the partition layout.
3. Subtract the piecewise linear trend obtained from the four points $(j_0, w_0) = (0, 0)$, (j_1, w_{j_1}), (j_2, w_{j_2}), $(j_3, w_n) = (n, 0)$ from the (j, w) data, i.e., replace the vector $(w_j)_{j=1,\ldots,n}$ with

$$w_j \leftarrow w_j - \begin{cases} \frac{j}{j_1} w_{j_1}, & \text{if } j \le j_1, \\ \frac{j-j_1}{j_2-j_1}(w_{j_2} - w_{j_1}) + w_{j_1}, & \text{if } j_1 < j \le j_2, \\ \frac{n-j}{n-j_2} w_{j_2}, & \text{if } j_2 < j. \end{cases} \tag{12}$$

4. Repeat from Step 2 on until a prescribed number of indices are obtained.

The estimate of η^2 is computed by (11), after adding 0 and n to the list of indices $\mathcal{J}$ for the partition layout. In Figure 2 the points obtained by this process are marked. Clearly, more sophisticated exit-criteria may be developed by taking the change of the η^2 estimate into account when the partition list is updated. A minimum-distance criterion between selected indices might also be of use.

6. Non-significant parameters

Let us discuss the detection of un-influential inputs. Using CSM curves, random permutations of the output sample are suggested to derive suitable confidence bands [1]. For polynomial fits, the Wilcoxon rank sum test for preventing overfitting is suggested by [7] and in [25] an optimisation method for selecting the optimal polynomial degrees is presented.

For correlation ratios we expect that all the conditional means $\bar{y}_r$ are near the global mean $\bar{y}$ if the functional influence of $\mathbf{X}$ on Y is non-significant. In this

situation, a test for the null hypothesis $H_0 : \bar{y}_r \equiv \bar{y}$ is known from ANOVA [8]. Its test statistics for comparing the conditional means against the global mean is given by

$$F = \frac{n-q}{q-1} \cdot \frac{\sum_{r=1}^{q} n_r (\bar{y}_r - \bar{y})^2}{\sum_{r=1}^{q} (n_r - 1) s_r^2}$$

which is tested against an F-distribution with $q-1$ numerator degrees of freedom and $n-q$ denominator degrees of freedom. This F-test is asymptotically robust so that it may also be used for non-normal distributions of Y. With (2) and (3) the test statistics may be expressed in terms of $\hat{\eta}^2$,

$$F = \frac{n-q}{q-1} \cdot \frac{\hat{\eta}^2}{1-\hat{\eta}^2}.$$

For a given test niveau α a critical value of η^2 can be computed using the upper α quantile of the F-distribution,

$$\eta^2_{\text{crit}}(\alpha; n, q) = \left(\frac{n-q}{q-1} h(\alpha) + 1 \right)^{-1}, \qquad h(\alpha) = (F^{q-1}_{n-q}(1-\alpha))^{-1}. \tag{13}$$

Estimates below this threshold are non-significant. From the example discussed above we take $n = 200$ and $q = 15$. Using $\alpha = 5\%$ gives a critical value of $\eta^2_{\text{crit}} = 0.1167$ which is not very convincing. However, reducing the number of partitions lowers this threshold value.

From (13) with $q = 2$ intervals and (10) we obtain an elliptical bound for unimportant factors given by

$$z(k)^2 \leq \frac{k(n-k)}{n^2} \eta^2_{\text{crit}}(\alpha; n, 2) = \frac{k(n-k)}{n^2((n-2)h(\alpha)+1)}, \quad k = 1, \ldots, n. \tag{14}$$

For $q = 2$ the upper α quantile satisfies $F^1_{n-2}(1-\alpha) = t^2_{n-2}(1-\alpha/2)$, yielding a Student t-distribution. If n is large then the quantile of the t-distribution is approximated by the quantile of the normal distribution Φ. Using this approximation, (14) then reads

$$|z(k)| \leq \frac{1}{n} \sqrt{\frac{k(n-k)}{(n-2)\Phi(1-\alpha/2)^{-2} + 1}}, \qquad k = 1, \ldots, n,$$

which can be even further simplified when $n \gg 1$,

$$|z(k)| \leq \Phi(1-\alpha/2) \frac{1}{n} \sqrt{\frac{k(n-k)}{n}}, \qquad k = 1, \ldots, n. \tag{15}$$

A niveau of $\alpha = \frac{1}{2} n^{-1/2}$ is suggested to compensate the error of the Monte-Carlo sampling. Note that α has to be adapted if quasi-Monte-Carlo sampling schemes like Sobol's LPτ sequences [17] are used.

In Figure 2 the ellipsoidal bound given by (15) shows that parameters x_3 and x_4 are insignificant and that parameter x_2 also hardly leaves this range. However, it oscillates between the lower and the upper bound. Such an effect cannot be dealt with when using a two-interval partition approach and therefore cannot be detected by the suggested significance ellipsis.

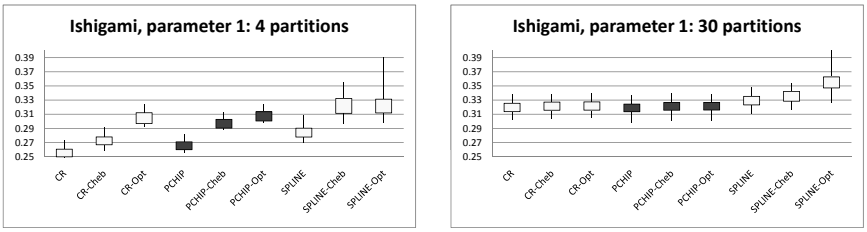

Figure 3: Parameter 1: Expected value is 0.3139

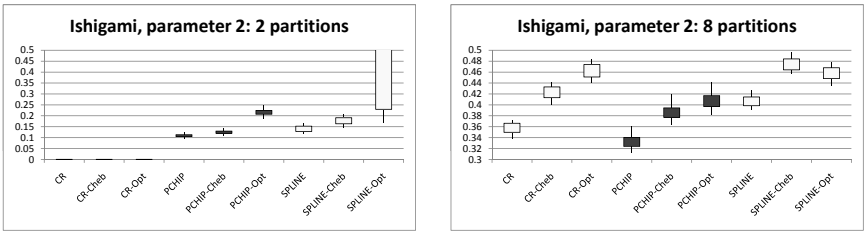

Figure 4: Parameter 2: Expected value is 0.4424

7. Estimators based upon smooth interpolation curves

The summation in (11) can easily be identified as a Riemannian sum, hence it may also be written formally as an integral, $\hat{\eta}^2 = \int_0^1 (\nabla z(nt))^2 \, dt$. We can use an interpolating function through the index points of a partition list like the one derived in Subsection 5.3. If this function is a piecewise polynomial then differentiation and integration can be performed analytically which may give better results in case of a convex z function.

Fitting an interpolation curve through the CUSUNORO curve seems to be a much simpler (and more robust) task than to create a regression curve through the scatter-plot.

We study if the quality of the estimates improves with the following setups. Estimators using

- a piecewise linear approximation (correlation ratios, CR),

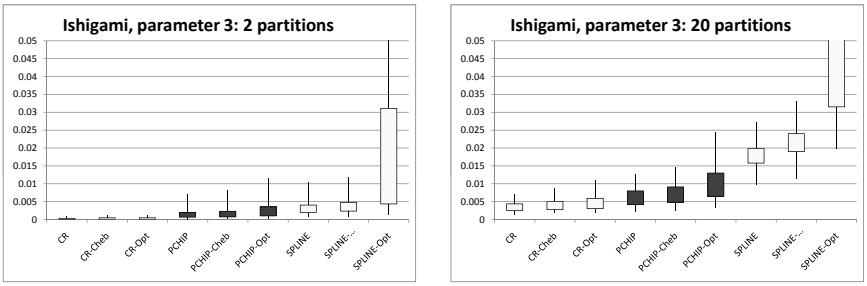

Figure 5: Parameter 3: Expected value is 0

- a cubic spline approximation (SPLINE), and
- a piecewise cubic Hermitian polynomial (PCHIP)

are applied to an index set composed of

- equally spaced indices,
- scaled Chebychev points (-Cheb), and
- an optimised partition layout (-Opt).

Here, Chebychev interpolation points are used to keep the polynomial interpolation error at the boundaries small. The polynomial regression models are computed by MatLab. As an example we again use the Ishigami function (5).

Thirty runs of sample size $n = 5000$ were simulated with simple random sampling. For this ample size the Monte-Carlo integration error is of order $\pm 2\%$ so that the differences are mostly due to the suggested CR estimators. The box-and-whiskers plots in Figures 3–5 show the extreme values and the first and third quartiles of the estimates. Some plots have been truncated to keep the same scale.

For small partition sizes the use of Chebychev spacing or the optimised spacing is advantageous to equidistant spacing, however spline interpolation is not suited for the optimised spacing and introduces overshoots. Parameter 2 with its four inner extrema (see Fig. 2) cannot be treated with very small partition sizes, see Fig. 4. For parameter 3 with its zero expectation, the interpolation curves only introduce unwanted effects. Moreover, a slight bias in the sizes of the partition is visible. For a large partition size, nearly all suggested methods produce comparable results, see the right plot of Fig. 3.

8. Comparison of methods

In this section we compare the proposed method with established ones. To distinguish this new method derived from the CUSUNORO curve we will call it CRA (**c**orrelation **r**atio using an **a**daptive partition layout). Again using the Ishigami function (5) with a fourth dummy parameter, let us consider for the same set of simple random samples

- CRA (using up to four pairs of inner points), with and without bias correction [7, 5.2.1: Adjusted R^2],
- CRA with Hermitian piecewise cubic splines,
- CR with rule of thumb "partition size is the square root of sample size",
- EASI (with maximum harmonic 8) [11], and
- RS-HDMR (with polynomial degree 6) [12, 25].

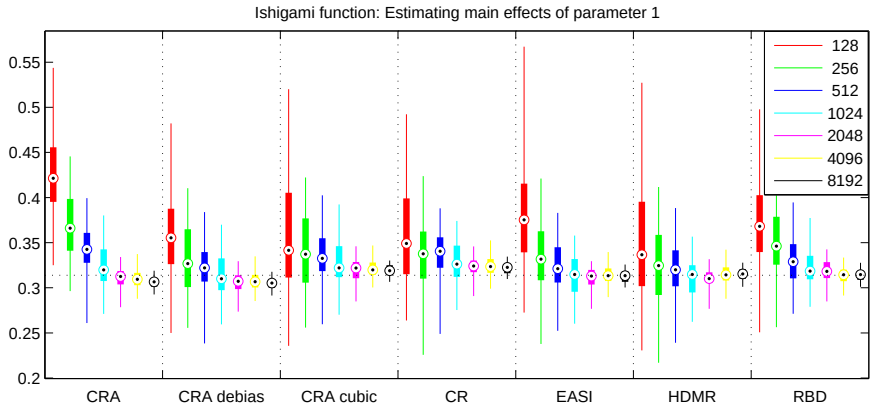

Figure 6: Convergence of correlation ratio estimates for parameter 1

Furthermore, we consider the following method using designed samples, but with the same sample sizes

- RBD (with maximum harmonic 8) [23].

Figures 6 to 9 show box-and-whiskers plots of the results from 50 repetitions of different sample sizes for parameters 1 to 4. The outlier detection uses 3 times the interquartile range.

The correlation ratio for an adaptive partition layout estimates are biased for small sample sizes, while the bias-adjusted version performs slightly better, however still exposes a visible bias. We also see that the polyline approximation using only a few inner interpolation points fails to capture details, while the cubic approximation using the same set of points overestimates the first order effect.

The bias can also be observed for the other methods of estimating first order effects. Nearly all methods show a bias which is most prominent for the zero value of parameters 3 and 4 which emphasises the need for detecting non-influential inputs as presented in Section 6. Both EASI and RDB use harmonic functions for a regression model so they are well suited for the Ishigami function. Also, the rule-of-thumb CR has comparable properties to other methods. For the polynomial regression with HDMR we used the analytical marginal cumulative distribution functions to transform the inputs to $[0, 1]^k$ such that a basis of shifted Legendre polynomials could be used.

9. Summary and Conclusions

We have developed a new graphical representation named CUSUNORO plot to compactly visualise sensitivity properties. This curve leads naturally to an adaptive partition layout for CR estimation by following the idea of "maximising the gradients." Other methods of estimating the variance-based sensitivity

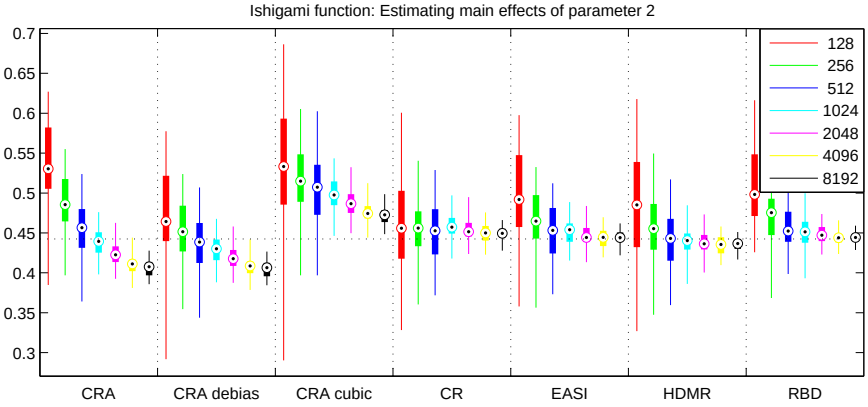

Figure 7: Convergence of correlation ratio estimates for parameter 2

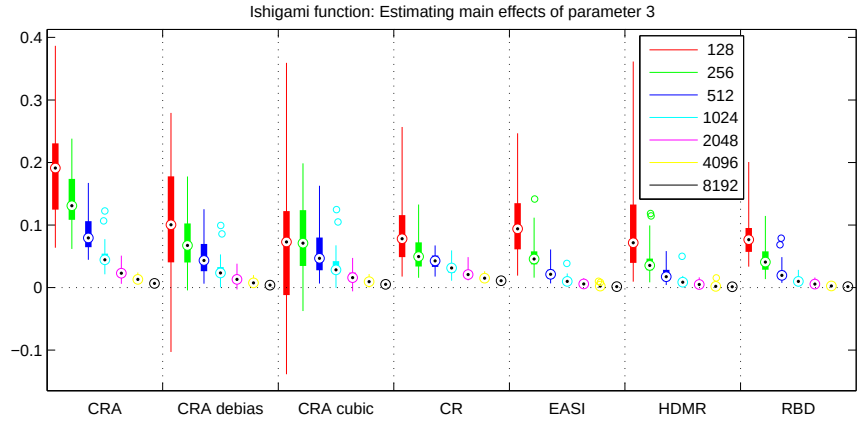

Figure 8: Convergence of correlation ratio estimates for parameter 3

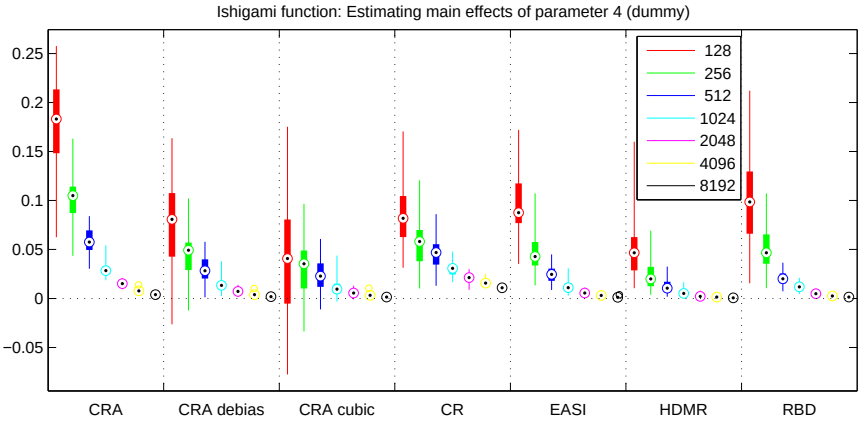

Figure 9: Convergence of correlation ratio estimates for parameter 4

index (1) which are based on interpolation of this plot rather than a non-linear regression of a scatter-plot are also immediate consequences of the graphical representation. CR methods are competitive with advanced methods of computing variance-based first order sensitivity indices. Small partitions with cleverly chosen subsamples are performing as good as equi-distant partition layouts with many subsamples. There is still room for improvements of CR methods and adaptive partition layout strategies. For example, alternatively an empirical estimator of $1 - \mathbb{V}[Y]^{-1}\,\mathbb{E}[\mathbb{V}[Y|\mathbf{X}]]$ might be used,

$$\begin{aligned}\hat{\eta}^2_{\mathrm{ECV}} &= 1 - \frac{n-1}{s_{yy}} \sum_{r=1}^{q} \frac{n_r}{n} s_r^2 \\ &= 1 - \sum_{r=1}^{q} \frac{n-1}{n_r - 1} \left(\frac{n_r}{n} \left(v(j_r) - v(j_{r-1}) \right) - \left(z(j_r) - z(j_{r-1}) \right)^2 \right).\end{aligned}$$

A different layout strategy could combine the significance test with bisecting partition intervals. Further work is also needed to jugde the quality of the different available debiasing techniques which might also be of interest for non-CR methods.

Instead of testing for the influence of a single input parameter, one may also be interested in the combined effect of two or more parameters. For these higher order effects, the idea of combining the CUSUNORO plot with space-filling curves looks promising [13]. Here, more practical experience has to be gathered.

Correlation ratios are a direct method of estimating first order effects. They can be combined with a sampling plan, or a meta-model approach which may increase the quality of the estimates.

References

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

[2] E. Borgonovo. A new uncertainty importance measure. *Reliability Engineering&System Safety*, 92(6):771–784, 2007.

[3] R. Cukier, C. Fortuin, K. Shuler, A. Petschek, and J. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I. Theory. *J. Chem. Phys.*, 59:3873–3878, 1973.

[4] K. Doksum and A. Samarov. Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *The Annals of Statistics*, 23(5):1443–1473, 1995.

[5] A. N. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer Verlag, Berlin, 1933.

[6] B. Krzykacz. *SAMOS: A Computer Program for the Derivation of Empirical Sensitivity Measures of Results from Large Computer Models*. Garching, Germany, 1990. Report GRS-A-1700, Contract No. 73 708, 31 050.

[7] D. Lewandowski, R. M. Cooke, and R. J. Duintjer Tebbens. Sample-based estimation of correlation ratio with polynomial approximation. *ACM Transactions on Modeling and Computer Simulation*, 18(1):3:1–3:16, 2007.

[8] R. G. Miller, Jr. *Grundlagen der angewandten Statistik (Beyond ANOVA)*. Oldenbourg Verlag, München, 1996.

[9] J. E. Oakley and A. O'Hagan. Probabilistic sensitivity analysis of complex models: A Bayesian approach. *J. R. Statist. Soc. B*, 66(3):751–769, 2004.

[10] K. Pearson. *On the General Theory of Skew Correlation and Non-linear Regression*, volume XIV of *Mathematical Contributions to the Theory of Evolution, Drapers' Company Research Memoirs*. Dulau & Co., London, 1905.

[11] E. Plischke. An effective algorithm for computing global sensitivity indices (EASI). *Reliability Engineering&System Safety*, 95:354–360, 2010.

[12] H. Rabitz and Ö. F. Alış. General foundations of high-dimensional model representations. *J. Math. Chem.*, 25(2–3):197–233, 1999.

[13] M. Ratto, A. Pagano, and P. Young. State Dependent Parameter metamodelling and sensitivity analysis. *Comput. Phys. Commun.*, 177:863–876, 2007.

[14] A. Saltelli and R. Bolado. An alternative way to compute Fourier amplitude sensitivity test (FAST). *Computational Statistics&Data Analysis*, 26:445–460, 1998.

[15] A. Saltelli, K. Chan, and E. Scott. *Sensitivity Analysis*. John Wiley&Sons, Chichester, 2000.

[16] I. M. Sobol'. Estimation of the sensitivity of nonlinear mathematical models. *Mat. Model.*, 2(1):112–118, 1990.

[17] I. Sobol´. On the distribution of points in a cube and the approximate evaluation of integrals. *U.S.S.R. Comput. Math. Math. Phys.*, 7(4):86–112, 1967. Translation from Zh. Vychisl. Mat. Mat. Fiz. 7, 784-802 (1967).

[18] I. Sobol´, S. Tarantola, D. Gatelli, S. Kucherenko, and W. Mauntz. Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Engineering&System Safety*, 92:957–960, 2007.

[19] C. B. Storlie and J. C. Helton. Multiple predictor smoothing methods for sensitivity analysis: Description of techniques. *Reliability Engineering&System Safety*, 93:28–54, 2008.

[20] C. B. Storlie and J. C. Helton. Multiple predictor smoothing methods for sensitivity analysis: Example results. *Reliability Engineering&System Safety*, 93:55–77, 2008.

[21] B. Sudret. Global sensitivity analysis using polynomial chaos expansion. *Reliability Engineering&System Safety*, 93:964–979, 2008.

[22] K. Takezawa. *Introduction to Nonparametric Regression.* John Wiley&Sons, Hoboken, NJ, 2006.

[23] S. Tarantola, D. Gatelli, and T. Mara. Random balance designs for the estimation of first order global sensitivity indices. *Reliability Engineering&System Safety*, 91:717–727, 2006.

[24] S. Tarantola, V. Kopustinskas, R. Bolado-Lavin, A. Kaliatka, E. Ušpuras, and M. Vaišnoras. Sensitivity analysis using contribution to sample variance plot: application to a water hammer model. *Reliability Engineering&System Safety*, 2011. Submitted.

[25] T. Ziehn and A. S. Tomlin. GUI-HDMR - a software tool for global sensitivity analysis of complex models. *Environmental Modelling&Software*, 24:775–785, 2009.