Sample–based Estimation of Correlation Ratio with Polynomial Approximation

DANIEL LEWANDOWSKI
Delft University of Technology
ROGER M. COOKE
Delft University of Technology and Resources For The Future
and
RADBOUD J. DUINTJER TEBBENS
Harvard School of Public Health

Sensitivity analysis has become a natural step in the uncertainty analysis framework. As there is no general sensitivity measure that would capture all information on impact of input factors on model output, analysts tend to combine various measures to obtain a broader image of interactions between different modes. This paper concentrates on the correlation ratio, demonstrates methods for calculating this quantity efficiently and accurately and compares the results. A new method inspired by artificial intelligence techniques emerges as outperforming the familiar methods.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: Correlation and regression analysis; G.1.6 [**Optimization**]: Least squares methods

General Terms: Algorithms, Theory

Additional Key Words and Phrases: Correlation ratio, Sobol' indices, Sensitivity analysis

1. INTRODUCTION

Suppose a model is defined as a function $G = G(X_1, X_2, \ldots)$. Aim of the sensitivity analysis is to investigate how much the uncertainty in X_i's, $i = 1, 2, \ldots$, or combinations thereof, contributes to the uncertainty in G. In this paper we concentrate on the notion of the so–called *correlation ratio* - a variance based measure.

The correlation ratio (CR) of random variable G with respect to random variable X is defined as

$$\eta^2(G|X) = \frac{Var(E(G|X))}{Var(G)}.$$

Evidently, this is not a correlation coefficient of random variables; it is not symmetric and it is always non-negative. The variable G is the *explanandum* (the variable to be explained) and the variable X is the *explanans* (the variable doing

Author's address: D. Lewandowski, Department of Mathematics, Delft University of Technology, Mekelweg 4, Delft, The Netherlands

the explaining).

Thanks largely to the work of McKay [1997] the correlation ratio is becoming recognized as a key notion in global sensitivity analysis. Other authors have studied this subject as well (e.g. [Chan et al. 1997; Ishigami and Homma 1990; Cooke and Lewandowski 2001]). Saltelli et al. [2000] offer an extensive overview of sensitivity analysis methods, including variance–based approaches. Recently correlation ratio has been applied and compared with other sensitivity measures in [Duintjer Tebbens et al.]. Theoretically, the correlation ratio is an attractive tool for quantifying importance because it represents the fraction of the variance of G that can be attributed to variation of X. However, there is an evident problem with computing it in a simple and accurate manner - estimation of the conditional expectation $E(G|X)$ is the real challenge. A number of algorithms have been developed to overcome this difficulty, some more successful than others. Instructive among the lesser successful are the methods proposed by Kendall and Stuart [1961] and Sobol' [1993]. The first relies on a user-selected parameter (the number of bins for discretizing the model) which fully controls the value of the estimates. The second leads to very large deviations in the results and possible negative values although some may consider this as a strength of this method as it gives unbiased estimates. In general there is no need to approximate $E(G|X)$ in order to estimate the correlation ratio (methods like FAST and Sobol' explained in [Saltelli et al. 2000] do not deal with that at all, for instance). However the regression curve $E(G|X)$ arises naturally in sensitivity analysis and having that determined is a useful byproduct.

Recognizing the drawbacks of the standard estimation methods, we look for methods which:

(1) are based only on samples and does not require additional simulation and/or special simulation methods,
(2) give an approximation of $E(G|X)$ in analytical form,
(3) do not require any input from the user, as this could control the result,
(4) are generic, ie. not model specific,
(5) are easy to implement in computer code,
(6) have the accuracy at least on par with other known methods.
(7) have little computational cost

The first point really means that we are interested in methods of estimating the correlation ratio from pseudo random or fully random samples only. The Bayesian method of Oakley and O'Hagan [2004] performs best if the samples for input variables are carefully chosen and therefore needs a special sampling algorithm. However, it also works with pseudo random samples very well, and therefore we include this method in our comparison. Theorems introduced in this paper help develop a new method of estimating correlation ratio complying with this specification. The main object therefore is to present and compare 3 variants of this new method and decide which one performs best. The best adaptation of the method will be compared with two already known state-of-the-art methods of estimating the correlation ratio on an example of a multivariate model.

The paper is organized as follows. Section 2 places the correlation ratio into a broader context of global sensitivity measures. Section 3 presents a general defini-

tion of the correlation ratio. Next, in section 4 we describe methods proposed by Oakley and O'Hagan [2004] and Li et al. [2002]. Section 5 introduces 3 variations of the new method of estimating the correlation ratio. The performance of this method is investigated in section 6 with conclusions and discussion following in section 7.

2. GLOBAL SENSITIVITY MEASURES

The correlation ratio belongs to a family of global quantitative measures of importance of input factors for a given model; it is a variance–based non–parametric method closely related to Sobol' indices [Sobol' 1993; Chan et al. 2000]. Sobol's method relies on decomposing the model function $G(\mathbf{U})$ into orthogonal summands of increasing dimensionality with zero mean, where $\mathbf{U} = (U_1, U_2, \ldots, U_N)$ is a vector of length N of statistically independent uniform random variables on $[0, 1]$ with realizations $\mathbf{u}$,

$$G(\mathbf{u}) = G_0 + \sum_{i=1}^{n} G_i(u_i) + \sum_{1 \le i < j \le n} G_{ij}(u_i, u_j) + \ldots + G_{1,2,\ldots,n}(u_1, u_2, \ldots, u_n), \quad (1)$$

where G_0 denotes the expectation of $G(\mathbf{U})$ and

$$\begin{aligned} G_i(u_i) &= E(G|u_i) - G_0; \\ G_{ij}(u_i, u_j) &= E(G|u_i, u_j) - G_i(u_i) - G_j(u_j) - G_0; \text{ etc.} \end{aligned}$$

Similarly, higher–order terms can be obtained. This is the starting point for the high-dimensional model representations (HDMR), tools for estimating G_i's. HDMR expresses the model output G as a function expansion as given in eq.(1). It can be generalized to non–uniform and correlated inputs (see [Li et al. 2006]). Li et al. [2002] approximate the HDMR component functions by orthonormal polynomials, polynomial spline functions and ordinary polynomials. They do not, however, consider the problem of overfitting which is evidently possible if the order of the polynomial is too high.

With the assumption of independence of inputs and given eq.(1) the variance of G may be written:

$$\begin{aligned} Var(G(\mathbf{U})) &= \textstyle\sum_{1 \le i < j \le n} \{Var(G_i(U_i)) + Var(G_{ij}(U_i, U_j)) \\ &\quad + \ldots + Var(G_{1,2,\ldots,n}(U_1, U_2, \ldots, U_n))\}. \end{aligned} \quad (2)$$

The Sobol' k-th order *sensitivity index* is defined as

$$S_{i_1,\ldots,i_k} = \frac{Var(G_{i_1,\ldots,i_k}(U_{i_1}, \ldots, U_{i_k}))}{Var(G)}.$$

Sobol' indices sum up to unity.

The first order Sobol' indices were used already by Pearson [1903].

The role of the correlation ratio in quantifying importance is based on the well-known relation (which does not require $\{U_i\}$ to be independent):

$$Var(G) = E(Var(G|U_i)) + Var(E(G|U_i)),$$

If the expected reduction in variance of G with U_i fixed is small, then the variance $Var(E(G|U_i))$ is large. Normalizing by $Var(G)$, $\frac{Var(E(G|U_i))}{Var(G)}$ represents the frac-

tion of the variance of G which is "explained" by U_i. The use of Sobol' indices as a sensitivity measure is then motivated by the fact that they explain *all* the variance, according to eq.(2). For a more detailed overview of Sobol' indices see [Chan et al. 2000]. Bedford [1998] addresses the problem of sensitivity indices for dependent random variables. The following section suggests another motivation of the correlation ratio, not based on variance reduction, but on optimal prediction.

3. DEFINITION OF CORRELATION RATIO

Building on the concept of Sobol' indices, we more generally define for any random vector $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ and any subset $\mathbf{X}^{(k)}$ of k components of $\mathbf{X}$, $(1 \leq k \leq N)$:

Definition 3.1. The correlation ratio η^2 of $G = G(\mathbf{X})$ with respect a to random vector $\mathbf{X}^{(k)}$ is

$$\eta^2\left(G|\mathbf{X}^{(k)}\right) = \frac{Var\left(E\left(G|\mathbf{X}^{(k)}\right)\right)}{Var(G)}. \tag{3}$$

The correlation ratio can be motivated in terms of optimal prediction. One may ask for which function $f : \mathbb{R}^k \mapsto \mathbb{R}$ with $\sigma^2_{f(\mathbf{X}^{(k)})} < \infty$ is the correlation $\rho^2(G, f(\mathbf{X}^{(k)}))$ maximal? The answer is given by the generalized result of Cooke and Lewandowski [2001] (similar to a result of Whittle [1992]).

THEOREM 3.2. *Let $\mathbf{X}^{(k)}$, G and and $f(\mathbf{X}^{(k)})$ have finite variance. Then*

$$\max_f \rho^2(G, f(\mathbf{X}^{(k)})) = \rho^2(G, E(G|\mathbf{X}^{(k)})) = \frac{Var(E(G|\mathbf{X}^{(k)}))}{Var(G)} = \eta^2(G|\mathbf{X}^{(k)}).$$

PROOF. Let $\delta(\mathbf{X}^{(k)})$ be any function with finite variance and write $f(\mathbf{X}^{(k)}) = E(G|\mathbf{X}^{(k)}) + \delta(\mathbf{X}^{(k)})$. Put $A = \sigma^2_{E(G|\mathbf{X}^{(k)})}$, $B = Cov(E(G|\mathbf{X}^{(k)}), \delta(\mathbf{X}^{(k)})) = Cov(G, \delta(\mathbf{X}^{(k)}))$, $C = \sigma^2_G$, and $D = \sigma^2_\delta$. Then

$$\begin{aligned}
\rho^2(G, E(G|\mathbf{X}^{(k)}) + \delta(\mathbf{X}^{(k)})) &= \frac{(A+B)^2}{C(A+D+2B)}, \\
\frac{\sigma^2_{E(G|\mathbf{X}^{(k)})}}{\sigma^2_G} &= \frac{A}{C}, \\
\frac{(A+B)^2}{C(A+D+2B)} \leq \frac{A}{C} &\iff B^2 \leq AD.
\end{aligned}$$

The latter inequality follows from the Cauchy-Schwarz inequality. □

If $k = 1$ then the conditioning set of variables $\mathbf{X}^{(k)}$ contains only one element which we denote by X. If the optimal regression of G on $\mathbf{X}^{(k)} = (X)$ is linear, that is, $E(G|X) = aX + b$, then

$$\begin{aligned}
&Var(E(G|X)) = Var(aX + b) = \\
&= \frac{Cov^2(aX + b, X)}{Var(X)} = \frac{Cov^2(E(G|X), X)}{Var(X)} = \frac{Cov^2(G, X)}{Var(X)},
\end{aligned}$$

and eq.(3) becomes the product moment correlation squared $\rho^2(G, X)$.

Sobol' indices coincide with the correlation ratio when the explanatory variables are independent uniforms. However, when the variables are not independent, the motivation of Sobol' indices in terms of variance decomposition, as in eq.(2) is lost. It suffices to consider $G = X + Y$ with $X = Y$. Then $\eta^2(G|X) = \eta^2(G|Y) = \eta^2(G|(X,Y)) = 1$, and they obviously do not sum to one. The correlation ratio admits a more general motivation in terms of prediction, according to Theorem 3.2.

4. STANDARD METHODS OF ESTIMATING CORRELATION RATIO

State-of-the-art methods for computing the correlation ratio include the Bayesian approach of Oakley and O'Hagan [2004] and State Dependent Parameter (SDP) model by Ratto et al. [2006]. We describe them both briefly here. The HDMR method of Li et al. [2002] stops where we start. It approximates the component functions but does not deal with the prevention of overfitting. It must be noted that a variety of other approaches exist for carrying out this task, like *FAST* [Saltelli et al. 1999]. We do not consider these in this paper in view of the requirements formulated in section 1.

It is assumed from now on that the sample size is m. Symbol $\mathbf{x}_j$ denotes the j-th vector of realizations of $\mathbf{X}$ and $\mathbf{x}_{i,j}$ is the j-th realization of X_i.

4.1 Bayesian approach

The first method employs the Bayesian paradigm by emulating G as a Gaussian process whose parameters are assigned hyper prior distributions and updating using model evaluations $G(\mathbf{x}_j)$, $j = 1, \ldots, m$. For further reading on this method please refer to the article of Oakley and O'Hagan [2004].

The biggest advantage of this approach is that it does not require a large number of simulations and therefore it is best suited for applications when computing the model evaluations is rather complicated and time consuming. On the other hand it requires the user to specify many parameters and to implement routines for numerical integration. The sensitivity of this method to various specifications of the input parameters is yet to be determined as there is no study on this subject (the choice of samples for instance).

4.2 State Dependent Parameter models

The State Dependent Parameter modelling developed in [Ratto et al. 2006], in turn, can be applied to any Monte Carlo sample and can be seen as one of the *postprocessing* methods, ie. the analysis is done after the creation of the sample. The idea is to extract the signal ($E(G|X_i)$) from noisy data ($G(X_i)$). In order to prepare simulation data, which does not need to exhibit any *temporal* order, for smoothing with this method one has to sort the values of X_i in an increasing order (with $Y = G(X_i)$ sorted accordingly) and *pretend* that this ordered statistic specifies a time series. The change in Y as X_i changes its value from $x_{i,j}$ to $x_{i,j+1}$ is modelled as a random walk process. The forward filtering algorithm has been coupled with backward recursive smoothing (in this case Fixed Interval Smoothing) algorithm since the data is available for the whole range and does not come sequentially.

Unfortunately, there is no computer implementation of this method available at the time of writing this article. Therefore a full comparison of the new method

with the SDP approach is not possible although it clearly has potential.

5. POLYNOMIAL APPROXIMATION METHODS

The problems of estimating the correlation ratio are, for the most part, caused by the recurring issue of estimating the regression curve $E(G|X)$ based on data. There is a great deal of literature on the latter [Draper and Smith 1998; Kleinbaum et al. 1998]; but we propose a simpler strategy that can be easily implemented.

For simplicity, we restrict attention to the case, where the explanandum X is a one–dimensional random variable rather then a vector.

The method we propose assumes that the regression function is *analytic*, that is it can be approximated as a Taylor expansion, ie. a polynomial function. Having said that one can immediately observe that Theorem 3.2 gives a good instrument for estimating $E(G|X)$. Intuitively, since the regression curve is a function that maximizes $\rho^2(G, f(X))$ over all possible $f(X)$, then under the above assumption of smoothness we are searching for a polynomial $g_d(X)$ of degree d that maximizes $\rho^2(G, g_d(X))$. Optimization methods can be implemented with the coefficients of the polynomial as independent variables.

5.1 Polynomial fit

For fixed d the optimization problem can be formulated as:

$$\text{maximize } \rho^2(G, p_0 + p_1 X + \ldots + p_d X^d) \tag{4}$$

Optimization routines are time consuming, however. The following theorem states that equivalent results can be obtained by simply applying the least–squares error method to fit the polynomial.

THEOREM 5.1. *Let $G = G(\mathbf{X})$ with $\sigma_G^2 < \infty$. Then*

$$\arg\min_f E(G - f(X))^2 = E(G|X).$$

PROOF. Decompose the variance of $G - f(X_i)$ in order to obtain

$$E(G - f(X))^2 = Var(G - f(X)) + E^2(G - f(X)).$$

Minimizing the right hand side of the above equation implies setting $E(G) = E(f(X))$ (hence $E^2(G - f(X)) = 0$). Express f as

$$f(X) = E(G|X) + \delta(X),$$

where $E(\delta(X)) = 0$, and note that

$$\begin{aligned} E(G \cdot E(G|X)) &= E(E(G|X)^2) \\ E(\delta(X) \cdot E(G|X)) &= E(E(G\delta(X)|X)) = E(G \cdot \delta(X)). \end{aligned}$$

Then

$$E(G - (E(G|X) + \delta(X)))^2 = E(G^2 + \delta(X)^2) - E(E(G|X)^2)$$

attains its minimum when $\delta(X) = 0$ and hence $f(X) = E(G|X)$. □

Henceforth we use the least squares error method to fit a polynomial to data.

5.2 Prevention of overfitting

Fitting a polynomial to data introduces a problem of overfitting. The challenge is not to fit (in the least–squares error sense) a function that predicts perfectly values of the fitted sample, but a function that will be representative for the whole population from which the samples were drawn. Therefore there is a need for introducing a mechanism to prevent overfitting. Since it has been assumed that the model is a polynomial, the only parameter that can be used for controlling the overfitting is the degree of the polynomial. Once the degree is fixed the coefficients are uniquely determined by applying the least–squares error method. We consider three methods for determining the optimal degree of the polynomial estimation of the regression curve:

Adjusted R^2. This statistic is a rather standard tool used in regression analysis for evaluating impact of additional variables on a model's performance. The multiple correlation R^2 can be computed as the squared correlation $\rho^2(G, g_d(X))$. The adjusted R^2, accounting for the number of parameters in the model, is

$$adj\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-d-1}$$

The adjusted R^2 can decrease if increasing the polynomial degree d is not associated with a sufficient increase in R^2. Choose the degree d maximizing the adjusted R^2.

Early stopping. This idea is based on an approach applied in machine learning models such as neural networks. The sample is split into two subsets: a test sample and a validation sample. A polynomial of degree d is fitted to the test sample and used to estimate the correlation ratio for the validation sample. Choose the lowest d such that the correlation ratio with polynomial of degree d on the validation set is greater than with $d + 1$.

Wilcoxon rank sum test. The one–sided Wilcoxon rank sum test also compares two data sets of estimates of correlation ratio based on the test sample (data set 1) and the validation sample (data set 2) and tries to detect the shift in their distributions. The null hypothesis is that both distributions are equal. The alternative is that data set 1 is statistically larger than data set 2. The test statistic is the sum of ranks of the test observations among all combined and sorted test and validation observations. Its distribution can be easily tabulated or approximated by the normal distribution [Hodges and Lehmann 1970].

Our specific application of this test relies on the following reasoning. First split a given sample into two equally sized subsets (T - test sample and V - validation sample), then fit a polynomial of degree d to the test sample. Now divide both T and V into 10 smaller data sets of equal size and calculate the approximate correlation ratios for each of these based on the polynomial fitted on the test sample. In the end 10 values of correlation ratio for the test sample and 10 corresponding values of correlation ratio for the validation sample are obtained. They form two sets that will be compared with the help of the Wilcoxon rank sum test. The sum of the ranks W_T of the test group is expected to be larger than this sum W_V for the validation group. We use the following p–value as an indication of overfitting

$$P(W_T \geq w_T) = p_W,$$

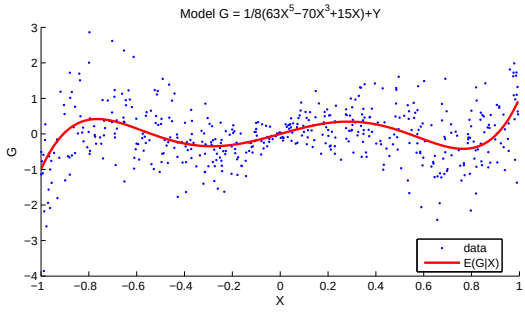


Fig. 1. Scatter plot of 500 samples generated given model $G = \frac{1}{8}(63X^5 - 70X^3 + 15X) + Y$.

where w_T is the realization of the rank sum of the test sample correlation ratios. Small p–value indicates overfitting. For the calculations presented next, we choose degree d for which the p–value is closest to 0.05 from above.

6. SIMULATIONS AND RESULTS

The performance of all of the variations of the polynomial method introduced in section 5.2 is compared in terms of their ability to estimate the true correlation ratio. The search algorithm is restricted to polynomials of degree from 1 to 20, as the fitting algorithms in generally available programs (eg. MATLAB) experience numerical instabilities for degrees greater than 20. The sample sizes that expose sensitivity for overfitting are therefore also relatively small. Of course, if higher degree polynomials can be reliably fitted, the overfitting issues will apply to larger sample sizes. The synthetic benchmark model used for simulations is chosen such that the true $\eta^2(G|X)$ can be easily calculated analytically (X is the explaining variable and Y is added noise). Let $G = f(X) + Y = \frac{1}{8}(63X^5 - 70X^3 + 15X) + Y$ where $X \sim U[-1, 1]$ and $Y \sim \mathcal{N}(0, \sqrt{|X|})$, $E(Y|X) = 0$. Thus the true regression function $E(G|X) = f(X)$ is known. This highly non–linear model presented in Fig.1 exhibits heteroscedasity in error variance, $\eta^2(G|X) \approx 0.1538$.

6.1 Influence of sample size

The sample size is a crucial factor in estimating any statistical quantity, therefore we study its influence on the accuracy of the estimations. It can be observed in Figure 2 that small sample sizes cause problems in estimating the correlation ratio accurately, as expected. The estimations of the correlation ratio are compared against the sample correlation ratio computed on the whole data set rather than the true η^2 in order to avoid penalizing the estimator for features of the data. Since the regression function is given the sample correlation ratio can be computed as the ratio of the sample variances $Var(f(X))$ and $Var(G)$. The polynomial approximation of degree 4 badly underestimates the sample correlation ratio of 0.155. It simply does not exhibit enough variability. On the other hand polynomial approximations of degree 5 (the degree of the model polynomial) and 15 yield good estimates for sample sizes greater than 400, indicating little sensitivity to the polynomial degree once it is at least equal to the degree of the true regression polynomial.

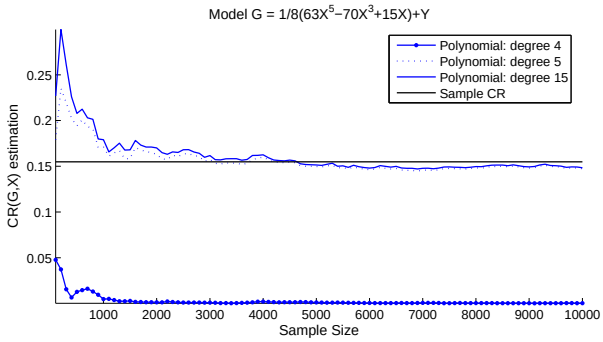


Fig. 2. Convergence of the estimation of correlation ratio $\widehat{\eta}^2(G|X)$ as a function of sample size n.

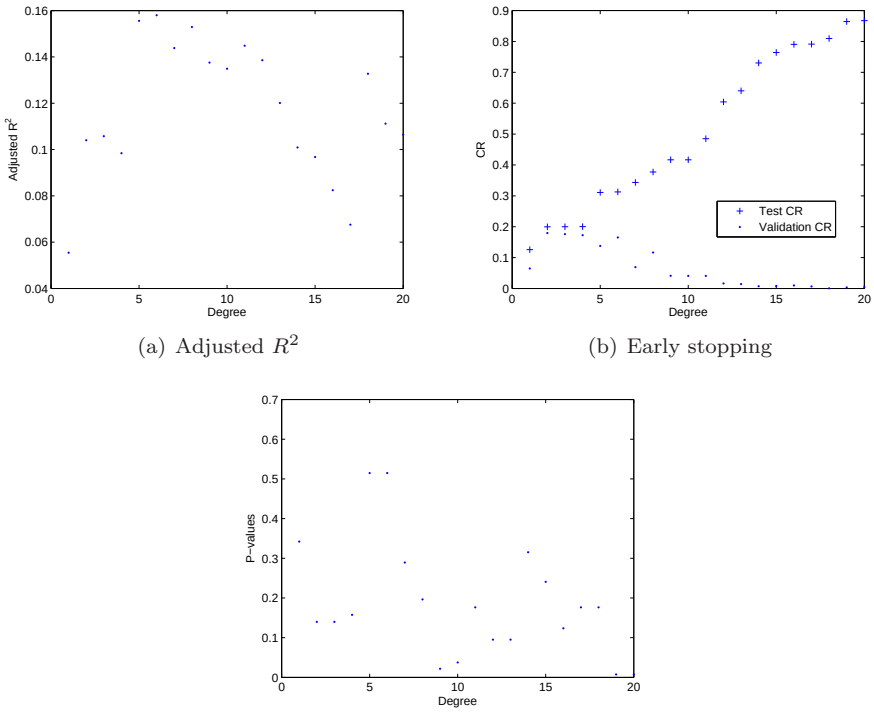
6.2 Overfitting

As it has been already mentioned an important issue for the polynomial methods of estimating the correlation ratio is the prevention of overfitting. Figure 3 shows a typical picture of what one may expect from the values of the adjusted R^2, the test and the validation CR's and p–values versus the degree of the fitted polynomial approximation for a small sample size, in this case 60. In this situation the adjusted R^2 statistic is rather unstable.

We proposed two other techniques for preventing overfitting designed with this specific issue in mind. Early stopping trains the polynomial approximation first on test data and then checks its performance on validation data. Figure 3b shows values of the estimates of the correlation ratio both on test data and validation data against the degree of the polynomial approximation. The validation–set correlation ratio gradually increases as the degree increases and eventually starts decreasing when the degree of the approximation becomes too high. We stop when the correlation ratio on the validation set starts to decrease. This method is more eager to penalize data overfitting by reducing the optimal degree of the polynomial approximation.

The Wilcoxon rank sum test for preventing overfitting is much more *forgiving* in a sense that it rejects the hypothesis of overfitting only after there is a clear evidence to do so. This evidence is the p-value being as close to 0.05 as possible, but not lower. The threshold value (0.05 in our case) should reflect analyst's particular risk attitude. The example we present in Figure 3c shows the p-values to be very noisy for this small sample but a general tendency for decreasing value as the degree increases can be observed.

Things become clearer with a larger data set of 200 samples (see Fig.4). There is a clear jump of the adjusted R^2 statistics when the degree of the approximation polynomial changes from 4 to 5. This jump can be explained by the fact that the true model is also a fifth order polynomial in X. The early stopping method also correctly detects the underlying model as the fifth order polynomial. The maximum of the validation–set correlation ratio is attained for degree equal to 5 and gradually decreases when the polynomial degree increases giving some evidence for overfitting. The behavior of the p-value of the Wilcoxon rank sum test is also



(a) Adjusted R^2

(b) Early stopping

(c) Wilcoxon rank sum test

Fig. 3. Various statistics vs degree of polynomial approximation for 60 samples.

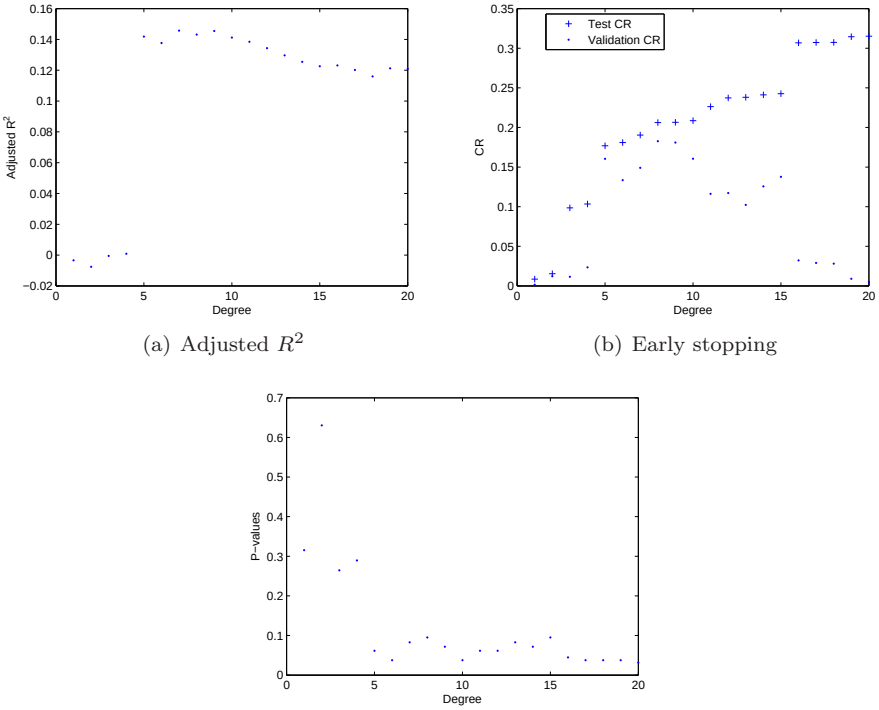
much more stable than with only 200 samples. The degree with the p-value closest to 0.05 from the top is 5 as well.

6.3 Robustness

The robustness of the estimation methods will be studied given three sample sizes - 60, 200 and 1000 samples. We estimate the statistical fluctuation of the estimation by iterating the estimation process 500 times. One iteration consists of the following steps:

(1) generate n samples of X, Y and compute $G = G(X, Y)$,
(2) fit polynomials of degree 1 to 20 to the whole sample and calculate adjusted R^2 for each polynomial (Adjusted R^2 method),
(3) fit polynomials of degree 1 to 20 to the first half of the sample and calculate the estimated correlation ratio on the other half of the sample for each polynomial (Early stopping method),
(4) fit polynomials of degree 1 to 20 to the first half of the sample, then split each half into 10 subsamples and calculate the p-value of the Wilcoxon rank sum test statistics for each polynomial (Wilcoxon rank sum test method).

Figure 5(a) shows the box plots of the estimates of correlation ratio calculated based on 60 samples using various polynomial methods presented in this paper. The



(a) Adjusted R^2

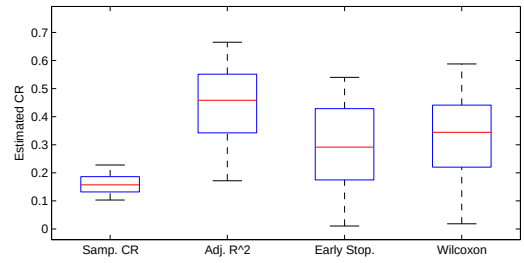
(b) Early stopping

(c) Wilcoxon rank sum test

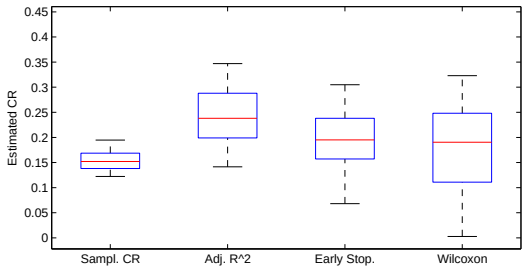
Fig. 4. Various statistics vs degree of polynomial approximation for 200 samples.

lower and upper lines of the boxes are the 25^{th} and 75^{th} percentiles of the sample and the whiskers are the 5^{th} and 95^{th} percentiles. The lines in the middle of the box plots show the medians. The first box plot (denoted as *Samp. CR* in Fig.5) represents the distribution of the estimates calculated using the true regression function, ie. the error of the estimates occurs only due to statistical fluctuation in samples. The remaining distributions contain variability also due to the model approximation. Selecting an optimal polynomial based on the adjusted R^2 tends to overestimate CR (data overfitting) compared to the early stopping and Wilcoxon methods. The best performing method both in terms of the accuracy and low variability is the early stopping algorithm.

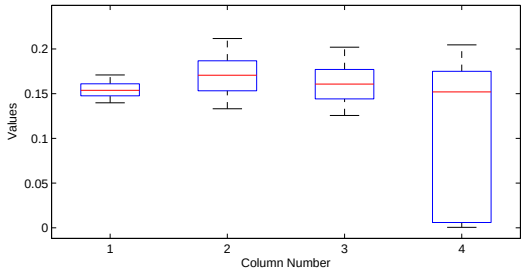
Figure 5(b) shows the box plots of the distribution of estimates given 200 samples. Quick comparison with Fig.5(a) shows that now the variability in the estimates is considerably smaller (maximal standard error of the order of 0.1 compared to 0.16). All methods perform better with larger number of samples providing more accurate estimates. Again, the best version of the polynomial method was early stopping. However, the Wilcoxon rank sum test starts to exhibit an undesirable feature - the erratic nature of the p-value causes in some cases to choose 4 or less as the optimal polynomial degree. The value of the correlation ratio is more heavily underestimated then (for example in Fig.2), making the box plots look very



(a) 60 samples



(b) 200 samples



(c) 1000 samples

Fig. 5. Box plots of the estimates of the correlation ratio.

stretched. Figures 5(b) and 5(c) confirm these observations.

The same model has been used for initial comparison of the Bayesian method with early stopping. The Bayesian method has been implemented in GEM-SA - Gaussian Emulation Machine for Sensitivity Analysis software and we use it in the analysis. 50 sets of samples were generated with 100 samples of X, Y and G per set. The estimates were converted to percentages and compared in this form in Table I. The Bayesian method underestimated the value of $\eta^2(G|X)$ (15.38) with estimates tightly concentrated around value 3.78. Increasing the number of samples to 400 (maximum supported by GEM-SA) did not cure this problem. This suggests that the Bayesian method may have problems with non-normal models and should be further explored. The early stopping algorithm on the other hand produced a more

Table I. Estimates of the correlation ratio - Bayesian and Early stopping methods

Method	Mean	RMSE
Bayesian	3.78	1.71
Early stop	18.97	10.11

sensible average estimate.

6.4 The analytic function of Oakley and O'Hagan

This model for benchmarks has been proposed by Oakley and O'Hagan [2004]. It is a multivariate model with 15 inputs

$$G(\mathbf{X}) = \mathbf{a}_1^T\mathbf{X} + \mathbf{a}_2^T \sin(\mathbf{X}) + \mathbf{a}_3^T \cos(\mathbf{X}) + \mathbf{X}^T\mathbf{M}\mathbf{X}, \tag{5}$$

where $\mathbf{X}$ is a vector of independent standard normal random variables. Scalar vectors $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ and matrix $\mathbf{M}$ are chosen such that the importance of the inputs can be classified into 3 categories based on the appropriate values of the correlation ratio. The same model has also been studied in [Ratto et al. 2006].

The full analysis of methods described in section 4 is not viable at this moment as the authors of the SDP method could not supply the code with the implementation. Therefore we base our findings on the comments of the authors in [Ratto et al. 2006]. On the other hand, the method of Oakley and O'Hagan [2004] has been implemented in GEM-SA - Gaussian Emulation Machine for Sensitivity Analysis software and we use it in our analysis.

Note that the estimates of the correlation ratio are presented on the percentage scale rather than fractions and all the results are calculated based on percentages.

Oakley and O'Hagan [2004] report that given 250 evaluations of eq.(5) at carefully chosen design points for $\mathbf{X}$ the standard deviations for the correlation ratio estimates of $X_1, \ldots, X_5$ is about 0.2, for $X_6, \ldots, X_{10}$ is 0.5 and for $X_{11}, \ldots, X_{15}$ is about 1. Since our method does not require any specific methods of generating the sample we compare it with O'Hagan's method using pseudo random samples produced in Matlab 2006b. This is of course the situation less favorable for the Bayesian method, but it complies with the desiderata declared in section 1. The decisive factor when we chose to limit the number of runs to 24 was long execution time of GEM-SA software. Also, out of these 24 runs only 10 distinct vectors of 15 estimates (for each input variable) were returned by GEM-SA. This suggests that the maximum likelihood optimization routine for the hyperparameters of the Bayesian method gets stuck at some fixed points quite often. This may give a misleading picture of the mean and RMSE of the estimates.

Figure 6 shows the mean estimates of the correlation ratios for this model based on 24 iterations, 250 samples per variable each. The estimates produced by the Bayesian method are much closer to the true values despite the fact that the input sample was not chosen optimally. The early stopping algorithm tends to overestimate the correlation ratios, especially if the true value is close to 0. The power of the prevention of overfitting is limited for a small sample size like this. The RMSE's of the results are also smaller for the Bayesian method (0.5, 1.5 and 2.5 for each of the three groups of input variables respectively) although not on a par with the results reported by [Oakley and O'Hagan 2004] if the sample is carefully

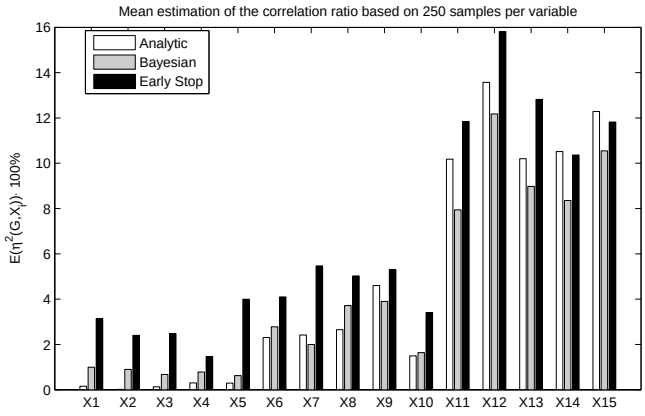


Fig. 6. Mean estimates of the correlation ratio based on 24 iterations and 250 samples per variable.

Table II. Mean and standard deviations of the estimates of the correlation ratio obtained with the early stopping method (750 and 1000 samples per variable, 100 iterations).

	Analytical	Mean		Standard deviation	
		750 samples	1000 samples	750 samples	1000 samples
X_1	0.1560	1.0499	0.7591	0.9011	0.6797
X_2	0.0186	1.0346	0.5218	0.8172	0.4934
X_3	0.1307	1.0311	0.6731	0.7570	0.5236
X_4	0.3045	1.1800	0.9251	1.1149	0.6570
X_5	0.2905	0.9849	0.7772	0.8972	0.6373
X_6	2.3035	2.9734	2.8181	1.2959	0.9932
X_7	2.4151	3.1584	2.9750	1.4744	1.2283
X_8	2.6517	2.8456	3.0997	1.3179	1.3230
X_9	4.6036	5.3172	5.5461	1.7640	1.6890
X_{10}	1.4945	2.0598	2.0152	1.0798	1.1111
X_{11}	10.1823	10.4025	10.7995	2.0375	2.1275
X_{12}	13.5708	13.9139	13.8106	2.3893	2.0873
X_{13}	10.1989	10.0289	10.3519	2.2431	1.9953
X_{14}	10.5169	11.0706	10.4579	2.4762	2.1103
X_{15}	12.2818	12.4564	12.4932	2.3133	2.0299

selected (0.2, 0.5 and 1). In order to achieve comparable RMSE with the early stopping method the number of samples would have to be increased to about 750 as Table II shows. This, however, is not enough to have similar mean estimates - for this 1000 samples have to be generated. Overall the early stopping method needs substantially more samples than the Bayesian approach. It will definitely not beat the SDP method either, which seems to perform very well in terms of determining values of the correlation ratios given 1000 samples per variable.

7. CONCLUSIONS AND DISCUSSION

There are many ways to quantify sensitivity. We have argued that the correlation ratio $\eta^2(G|X)$ is particularly attractive in this regard, although it cannot always be computed on-the-fly, and may be difficult to compute analytically.

The correlation ratio can be accurately estimated if the regression $E(G|X)$ of G on X is determined with sufficient accuracy. This paper develops a benchmark for testing candidates for good estimates of $E(G|X)$. The polynomial method assumes that the underlying model is sufficiently smooth and can be accurately approximated with a polynomial. In order to prevent overfitting we employ three

well motivated techniques based on: the adjusted R^2, early stopping algorithm, and Wilcoxon rank sum test. The early stopping method is most resistant to overfitting, has no "tweakable" parameters, is easy to implement and gave the best results. Therefore we used this specific algorithm for further comparison with the Bayesian method. The Bayesian method performed very well on the benchmark model proposed by Oakley and O'Hagan [2004], but experienced difficulties with the model in section 6. Without questioning the advantages of Bayesian methods for calculating the correlation ratio, there is a need for a simple generic method that works for a wide variety of models and sample sizes. Polynomial approximations perform decently in this regard, with early stopping as front runner and are very cheap to run when implemented in computer code. Obviously there is a trade-off here between the cost of needing a lot of samples (depends on how expensive the model is to run), and the cost of the algorithm itself. It should be noted that one run of GEM-SA takes 5 minutes to complete one calculation of estimates of $\eta^2(G, X_i)$ for the model described in section 6.4 on the current top-of-the-line dual core Intel processor (Intel Core 2 Extreme X6800) with only the option to calculate main effects selected and all the remaining program options set to default.

Polynomial approximation methods can also be extended for estimation of joint effects of 2 or more random variables on the output. One dimensional polynomial function would simply be replaced by its multidimensional counterpart. A possible future research can look more into the robustness of various methods of estimating the correlation ratio for different models as the choice of benchmark models mattered quite a lot in this study.

REFERENCES

Bedford, T. 1998. Sensitivity indices for (tree)-dependent variables. In *SAMO'98, Proceedings of Second International Symposium on Sensitivity Analysis of Model Output*, K. Chan, S. Tarantola, and F. Campolongo, Eds. 17–20.

Chan, K., Saltelli, A., and Tarantola, S. 1997. Sensitivity analysis of model output: variance-based methods make the difference. In *Proceedings of the 29th conference on Winter simulation*. 261 – 268.

Chan, K., Saltelli, A., and Tarantola, S. 2000. Winding stairs: A sampling tool to compute sensitivity indices. *Statistics and Computing 10*, 187–196.

Chan, K., Tarantola, S., Saltelli, A., and Sobol, I. 2000. *Sensitivity Analysis*. Wiley, Chapter Variance-Based Methods, 167–197.

Cooke, R. and Lewandowski, D. 2001. Bayesian sensitivity analysis. In *System and Bayesian Reliability - Essays in Honor of Professor Richard E. Barlow on His 70^{th} Birthday*, Y. Hayakawa, T. Irony, and M. Xin, Eds. Vol. 5. World Scientific, 315–331.

Draper, N. and Smith, H. 1998. *Applied regression analysis*. Wiley series in probability and statistics. John Wiley and Sons, Inc.

Duintjer Tebbens, R. J., Thompson, K. M., Hunink, M. G. M., Mazzuchi, T. M., Lewandowski, D., Kurowicka, D., and Cooke, R. M. Uncertainty and sensitivity analyses of a dynamic economic evaluation model for vaccination programs. *Medical Decision Making*. In press.

Hodges, J. and Lehmann, E. 1970. *Basic Concepts of Probability and Statistics*. Holden-Day.

Ishigami, T. and Homma, T. 1990. An importance quantification technique in uncertainty analysis for computer models. In *Proceedings of the ISUMA '90 First International Symposium on Uncertianty Modelling and Analysis*. 398–403.

Kendall, M. and Stuart, A. 1961. *The Advanced Theory of Statistics - Volume 2 Inference and Relationship*. Charles Griffin and Company Limited, London.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., and Nizam, A. 1998. *Applied Regression Analysis and Multivariable Methods.* An Alexander Kugushev book. Brooks/Cole Publishing Company.

Li, G., Hi, J., Wang, S.-W., Georgopoulos, P. G., Schoendorf, J., and Rabitz, H. 2006. Random sampling-high dimesional model representations (rs-hdmr) and orthogonality of its different order component functions. *Journal of Physical Chemistry 110,* 7, 2474–2485.

Li, G., Wang, S.-W., and Rabitz, H. 2002. Practical appraoches to construct rs-hdmr component functions. *Journal of Physical Chemistry 106,* 37, 8721–8733.

McKay, M. 1997. Nonparametric variance–based methods of assessing uncertainty importance. *Reliability Engineering and System Safety 57*, 267–279.

Oakley, J. E. and O'Hagan, A. 2004. Probabilistic sensitivity analysis of complex models: a bayesian approach. *Journal of the Royal Statistical Society 66*, 751–769.

Pearson, K. 1903. Mathematical contributions to the theory of evolution - on homotyposis in homologous but differentiated organs. *Proceedings of the Royal Society of London 71*, 288–313.

Ratto, M., Saltelli, A., Tarantola, S., and Yound, P. 2006. Improved and accelerated sensitivity analysis using state dependent parameter models. Eur 22251 en, Joint Research Centre, European Commission.

Saltelli, A., Chan, K., and Scott, E. M., Eds. 2000. *Sensitivity Analysis.* Wiley.

Saltelli, A., Tarantola, S., and Campolongo, F. 2000. Sensitivity analysis as an ingredient of modeling. *Statistical Science 15,* 4, 377–395.

Saltelli, A., Tarantola, S., and Chan, K. 1999. A quantitative, model independent method for global sensitivity analysis of model output. *Technometrics 41*, 39–56.

Sobol', I. 1993. Sensitivity analysis for nonlinear matematical models. *Mathematical Modelling and Computational Experiment 1*, 407–414.

Whittle, P. 1992. *Probability Via Expectation.* Springer Texts in Statistics. Springer.