Accuracy of data-based sensitivity indices

Thomas Bassine[*] Bryan Cooley[†] Kenneth Jutz[‡] and Lisa Mitchell[§]

Advisors: Pierre Gremaud[¶] and Joseph Hart[‖]

Abstract. When analyzing high-dimensional input/output systems, it is common to perform sensitivity analysis to identify important variables and reduce the complexity and computational cost of the problem. In order to perform sensitivity analysis on fixed data sets, i.e. without the possibility of further sampling, we fit a surrogate model to the data. This paper explores the effects of model error on sensitivity analysis, using Sobol' indices (SI), a measure of the variance contributed by particular variables (first order indices) and by interactions between multiple variables (total indices), as the primary measure of variable importance. We also examine partial derivative measures of sensitivity. All analysis is based on data generated by various test functions for which the true SI are known. We fit two non-parametric models, Multivariate Adaptive Regression Splines (MARS) and Random Forest, to the test data, and the SI are approximated using R routines. An analytic solution for SI based on the MARS basis functions is derived and compared to the actual and approximated SI. Further, we apply MARS and Random Forest to data sets of increasing size to explore convergence of error as available data increases. Due to efficiency constraints in the surrogate models, constant relative error is quickly reached and maintained despite increasing size of data. We find that variable importance and SI are well approximated, even in cases where there is significant error in the surrogate model.

1. Introduction.

In high dimensional problems, techniques that are used to analyze lower dimensional problems are too computationally costly. It is thus desirable to reduce the dimension of the problem, so identifying the relative importance of the variables is a high priority. If some variables can be deemed to have a relatively small effect on the function output, then these variables can be fixed and the dimension of the problem is reduced. Dimension reduction through identification of variable importance has many applications in engineering and the sciences. In this paper we seek, through computational experiments, to demonstrate how well sensitivity indices, a measure of variable importance, are approximated through the use of surrogate models when only a fixed amount of data is known.

[*]Department of Mathematics, University of Connecticut, Storrs, CT 06269, USA (thomas.bassine@uconn.edu)

[†]Department of Mathematics, East Tennessee State University, Johnson City, TN 37604, USA (cooleyj@goldmail.etsu.edu)

[‡]Department of Mathematics, North Carolina State University, Raleigh, NC 27695, USA (kfjutz@ncsu.edu)

[§]Department of Mathematics, Brigham Young University Idaho, Rexburg, ID 83460, USA (mit11014@byui.edu)

[¶]Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA (gremaud@ncsu.edu)

[‖]Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA (jlhart3@ncsu.edu)

Our general procedure in this paper is the following. We have a p dimensional test function $f : \mathbb{R}^p \to \mathbb{R}$. We sample data $\{\mathbf{x_i}, y_i\}_1^n$ from f, where $y_i = f(\mathbf{x_i})$ is the output of the ith data point, and then consider the black-box problem where we are restricted to only these inputs and outputs. It should be understood $\mathbf{x_i} = (x_{1i}, x_{2i}, ..., x_{pi})$ is the ith sampled data point. It is important to stress that we are no longer allowed access to f after we sample data from this function. We perform all proceeding analysis on a surrogate model, $\hat{f}$, which is fitted to the data. A surrogate model is an approximation of f from the data and can be evaluated anywhere in the domain. There are a variety of surrogate models and if some properties of the underlying function can be surmised, a particular choice may be advisable (i.e. linear regression when the function is essentially linear). See (5) for a review of surrogate models. In this paper, we utilize two non-parametric surrogate models: Multi-Adaptive Regression Splines (MARS) and Random Forest. We are motivated by the fact that these two surrogate models are popular in applications. Their properties are described in Section 2.

We perform sensitivity analysis on $\hat{f}$ to determine the sensitivity indices of the surrogate model. We then compare the results of this analysis to the known sensitivity indices of f. To ensure understanding of the precise definition of a sensitivity index, we shall offer a brief review of global sensitivity analysis. Broadly, global sensitivity analysis is the practice of quantifying variable importance of the inputs to a function as they are allowed to vary over their entire domains. Two prominent approaches in global sensitivity analysis are derivative based methods and variance based methods. We begin with an explanation of a derivative approach and then describe the variance based method that we utilize predominantly.

A logical way to define the sensitivity of x_i at a point $\mathbf{s} \in [0,1]^p$ is how much the function changes in response to a slight perturbation of x_i, while all other inputs are held constant. It is clear that this quantity is simply the partial derivative of f with respect to x_i evaluated at s. This quantity is called a local sensitivity. To extend this idea to a global measure over the entire domain, we consider

$$I_k = \int_{[0,1]^p} \left| \frac{\partial f}{\partial x_k} \right| d\mathbf{x} \tag{1.1}$$

where $d\mathbf{x} = dx_1, dx_2, ..., dx_p$ in this report. We use a means of estimating this quantity called Morris screening, which is described in (8). While this approach to global sensitivity is intuitive, it has drawbacks in practice. The surrogate model is an approximation of f, not its derivative. Hence even if we could find the true derivative of $\hat{f}$, this may be a poor approximation of the derivative of f.

We closely follow the variance based approach, pioneered by I.M Sobol and others (12). We now assume $f \in L^2$. The function admits the ANOVA decomposition:

$$f(x_1, ..., x_p) = f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{i<j} f_{i,j}(x_{i,j}) + \cdots + f_{i_1,...,i_p}(x_{i_1,...,i_p}) \tag{1.2}$$

While there are many different ways to write f in this form, the ANOVA decomposition requires the following constraint to make it unique:

$$\int_0^1 f_{i_1,...,i_n} dx_k = 0 \text{ for } k = i_1, ..., i_n.$$

Following Sobol's example, we view our problem through a probabilistic framework. Each input, x_j, is treated as a random variable with distribution Uniform[0,1] and hence f is a random variable as well. The ANOVA decomposition allows us to attribute variance in the function f to variance in the input variables. We use two measures of sensitivity, first order indices (S_i) and total indices (T_i). The first order indices measure the effect of each variable acting alone on the variance of the function while the total indices take into account variables acting together. They are defined below. Note that in the definition of T_i, K is the power set of $\{x_1, ..., x_p\}$, and integrating with respect to dx_k means to integrate with respect to each $x_j \in K$.

$$S_i = \frac{\int_0^1 f_i^2 dx_i}{\int_{[0,1]^p} f^2 dx_1, ..., dx_p - f_0^2}$$

$$T_i = \frac{\sum_{k \in K, x_i \in k} \int {f_k}^2 dx_k}{\int_{[0,1]^p} f^2 dx_1, ..., dx_p - f_0^2}$$

To understand the meaning of S_i and T_i, it is useful to note an equivalent definition:

$$S_i = \frac{\text{Var}\left(\mathbb{E}(f|x_i)\right)}{\text{Var}(f)} \tag{1.3}$$

$$T_i = \frac{\mathbb{E}\left(\text{Var}(f|x_i)\right)}{\text{Var}(f)} = 1 - \frac{\text{Var}\left(\mathbb{E}(f|x_i)\right)}{\text{Var}(f)} \tag{1.4}$$

A central theme of this paper is that we perform sensitivity analysis on $\hat{f}$ to estimate the first order indices and total indices of the model. We refer to these indices as $\hat{S}_i$ and $\hat{T}_i$ respectively. We are interested in how well these indices approximate S_i and T_i which are computed analytically. This idea is important because in applications it is desirable to use $\hat{S}_i$ and $\hat{T}_i$ to make statements about variable importance of f, in the pursuit of reducing dimensionality. Thus we compare $\left|S_i - \hat{S}_i\right|$ and $\left|T_i - \hat{T}_i\right|$ for $i = 1, ..., p$.

Once we have $\hat{f}$, we compute $\hat{S}_i$ and $\hat{T}_i$ in one of two ways. In the case when $\hat{f}$ is a degree one MARS model, we analytically compute the exact value of $\hat{S}_i$ using an R routine discussed in Section 3. In all other cases, we follow the commonly used practice of approximating $\hat{S}_i$ and $\hat{T}_i$ by Monte Carlo integration. Broadly, Monte Carlo integration computes an integral by approximating an expected value. The expected value of a function is approximated by taking a random sample of size N and averaging the function at these data points. The desired integral is then computed by multiplying the expected value by the volume of the domain. For more information on Monte Carlo integration, see (3).The advantage to this approach is that there is a theoretical error bound that is $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$. Moreover, this rate of convergence is irrespective of the dimension. While there is an estimable error bound, this convergence rate is slow and obtaining high accuracy is thus computationally costly.

Some work has already been done on quantifying the error in sensitivity indices computed from surrogate models. Storlie et al. (14) introduced the idea of bootstrapping to build confidence intervals (CI) for S_i estimated from surrogate models. They used bootstrapped datasets from the surrogate model itself, however, which may limit the effectiveness of this

approach when the surrogate model poorly approximates the function. In our work, we analyze the index approximations along with model error. They also find that MARS is a relatively inexpensive algorithm that often performs well for estimating variable importance. Thus focusing a large part of our work on MARS is reasonable and we build on their results by testing two new functions.

Janon et al. (6) developed a procedure that allows one to find an analytic upper and lower bound on estimators of the first order indices by sampling from the surrogate model. Their method was successful on lower dimensional problems, but it also requires the error in the surrogate model to be small. In contrast, we examine higher dimensional problems where the model is a poor approximation of the true function.

The following is a brief outline of the topics discussed in this paper. In Section 2, we provide a basic overview of how the MARS and Random Forest routines work. We next explain, in Section 3, the methods that we use to perform our numerical experiments. The three test functions are introduced, an explanation of error measures is provided, and we touch upon the R code created to compute S_i of degree one MARS models.

In Section 4, we offer numerical results as to the effectiveness of surrogate models preserving S_i and T_i. We find that MARS and Random Forest accurately predict the order of variable importance for two of the test functions. Interestingly, we find $\left|S_i - \hat{S}_i\right|$ can be small even when $\hat{f}$ is not a good approximation of f. We also find the degree one MARS models approximate S_i well in two cases when we scale to account for the fact that we are using an additive model. We find Random Forest overestimates $\hat{S}_i$ for important variables and does well on the unimportant variables, for one test function. We also examine the effects of increasing the dimension of the problem. For the g-function, MARS does surprisingly well approximating S_i even in higher dimensions.

In Section 5, we examine how a derivative based sensitivity approach compares to the variance based method. We provide an example where local oscillations of the test function cause the two approaches to provide contradictory results as to variable importance. We also provide an example of a Morris screening approximation of partial derivatives that is very inaccurate in 15 dimensions. In Section 6, we summarize our findings.

2. Surrogate Models.

2.1. Multivariate Adaptive Regression Splines. Multivariate adaptive regression splines (MARS) is a form of non-parametric regression analysis that fits a basis of weighted piecewise linear splines to data. Each basis function has the form $(x-t)_+$ or $(t-x)_+$, where

$$(x-t)_+ = \begin{cases} x-t, & \text{if } x > t, \\ 0, & \text{otherwise,} \end{cases} \quad \text{and } (t-x)_+ = \begin{cases} t-x, & \text{if } x < t, \\ 0, & \text{otherwise.} \end{cases}$$

Let

$$C = \{(x_j - t)_+, (t - x_j)_+\}_{\substack{t \in \{x_{1j}, x_{2j}, \ldots, x_{Nj}\} \\ j=1,2,\ldots,p}}$$

be the set of all possible basis functions. The MARS algorithm makes two 'passes' over the data. At each step in the first pass the algorithm adds a pair of basis functions from C to the model that most decreases the residual error. The second pass is the 'pruning' pass,

where to avoid over fitting, the algorithm removes basis functions that contribute the least to minimization of residual error. The maximum order d of the basis functions can be decided as an input to 'earth' in R as degrees of interaction. One degree of interaction results in an additive model of univariate functions. It is important to note that basis functions of higher degree are always products of splines of different variables, rather than polynomials of one variable. The result is a model of the form

$$\hat{f}(\mathbf{x}) = \beta_0 + \sum_{m=1}^{M} \beta_m \mathrm{h}_m(\mathbf{x})$$

where

- $\beta_0 \in \mathbb{R}$ is the intercept.
- $\beta_{1,\ldots,M} \in \mathbb{R}$ are weights.
- $\mathrm{h}_m(\mathbf{x}), m = 1, \ldots, M$ are products of d functions in C where d is a user input. (4).

2.2. Random Forest. Random Forest is an ensemble learning approach proposed by Breiman (2). Essentially, it is a modification to the method of bagging (1) which constructs regression trees from random data samples, using bootstrap sampling. Random Forest takes the bagging method one step further, averaging a large collection of trees by taking a random sample of predictor variables in choosing where each node is best split. By the use of random sampling of predictor variables, in addition to bootstrap sampling, instability is subsequently reduced. Thus by additionally averaging over a large number of trees it is obvious that the variance is reduced. Consequently, however, as the well known bias-variance tradeoff describes, the crux of this result is that averaging trees that are built using only a subset of variables increases bias while reducing variance. We implement the R package **randomForest**(7) in our computations involving Random Forest.

3. Methods.

3.1. Test Data Functions. In order to explore the effects of surrogate model error on sensitivity analysis, it is necessary to use data for which the variable importance is known. We employ test data functions for which Sobol' indices and derivative sensitivity measures can be expressed analytically in terms of the function parameters. By manipulating the parameters associated with each variable, test data sets are generated with exact sensitivity measures that can be compared to sensitivity measures approximated from surrogate models. We make use of three such functions.

1. **Sobol *g*-function**

$$f(\mathbf{x}) = \prod_{i=1}^{15} \frac{|4x_i - 2| + a_i}{1 + a_i}, \text{ where } a_i = \frac{i-2}{2}, \tag{3.1}$$

$$x_i \sim \text{ i.i.d. on } \mathrm{U}(0,1) \text{ for all } i = 1, \ldots, 15;$$

2. **Gaussian Function**

$$f(\mathbf{x}) = \prod_{i=1}^{20} 1.2 \exp\left(\frac{-(x_i - b_i)^2}{c_i}\right), \tag{3.2}$$

$$x_i \sim \text{ i.i.d. on } \mathrm{U}(-1, 1) \text{ for all } i = 1, \ldots, 20,$$

$$b_i \in \{0.4, 0.3, 0.2, 0.1, 0.4, 0.3, 0.2, 0.1, 0.4, 0.3, 0.2, 0.1, 0.4, 0.3, 0.2, 0.1, 0.4, 0.3, 0.2, 0.1\},$$

$$c_i \in \{0.5, 0.5, 0.5, 0.5, 1, 1, 1, 1, 2.5, 2.5, 2.5, 2.5, 3, 3, 3, 3, 5, 5, 5, 5\},$$

and b_i and c_i are ordered;

3. **Oscillatory Function**

$$f(\mathbf{x}) = \prod_{i=1}^{10} 1000^{0.1} (\exp(d_i x_i) \sin^2(2\pi c_i x_i) + .3), \tag{3.3}$$

$$x_i \sim \text{ i.i.d. on } \mathrm{U}(0, 1) \text{ for all } i = 1, \ldots, 10,$$

$$c_i \in \{1, 1, 2, 2, 2, 3, 3, 10, 10, 10\},$$

$$d_i \in \{5, 5, 2, 2, 2, 1, 1, 1, -1, -1\},$$

and c_i and d_i are ordered.

3.2. Comparison of approximation methods for Sobol' indices. Computing the integrals necessary to determine the exact, analytical indices of our surrogate model can sometimes be impractical. For this reason, a variety of methods, relying on Monte Carlo integration, have been devised to approximate Sobol' indices. The R package **Sensitivity**(10) provides eight different methods for performing these approximations; each method utilizes a different approach to the Monte Carlo estimation. To determine which was most effective, we tested each method on our test functions and calculated the 1-norm relative error for each, defined as:

$$E = \sum_{j=1}^{p} \frac{\left| S_j - \hat{S}_{j,N} \right|}{S_j} \tag{3.4}$$

Our results show that **sobolowen**(9) is by far the best method, as can be seen in Figure 3.1. The advantage of **sobolowen** is that it was designed specifically to do a better job of approximating smaller indices. When applicable, **sobolowen** is the method we use to approximate first-order and total Sobol' indices.

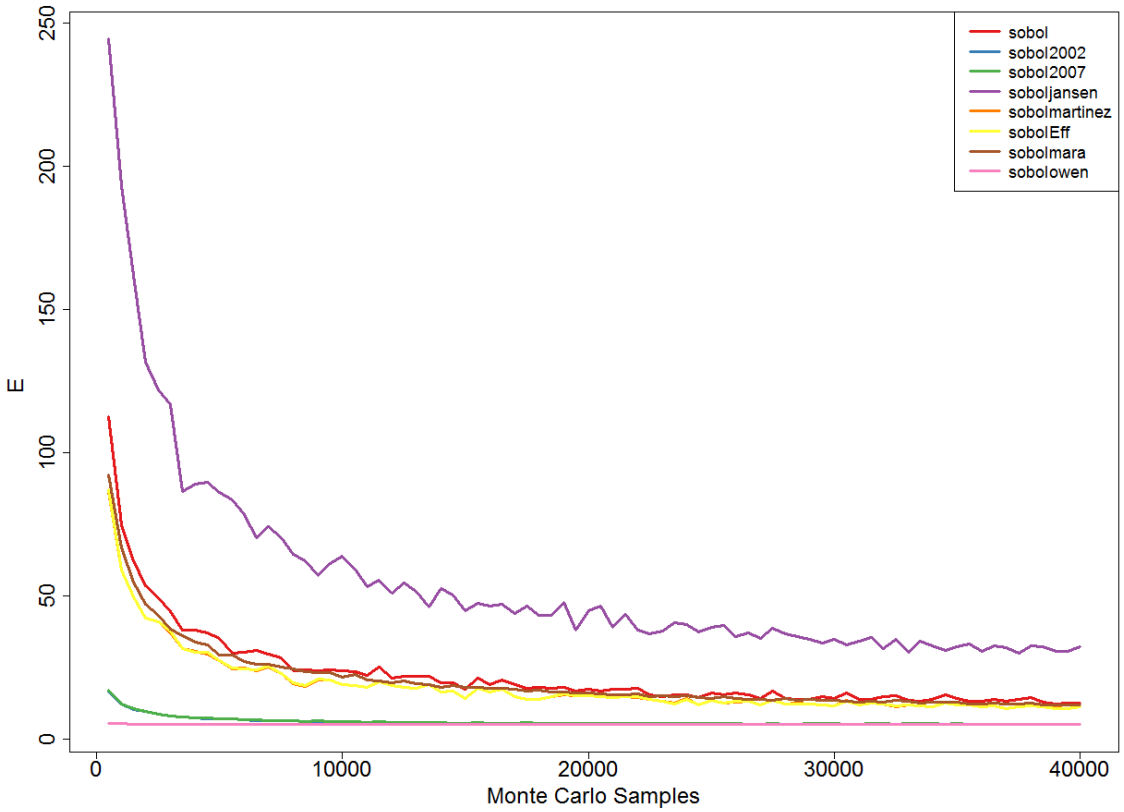


Figure 3.1: Convergence of R package **Sensitivity**(10) Sobol' methods to the Sobol' indices of the Sobol' g-function. Each point is the mean of 200 iterations, the range is 500 to 40000 in steps of 500. The error is measured as the 1-norm relative mean, see equation (3.4). It is evident that **sobolowen** has the least error. Note, the methods **2002** and **2007** are overlapping and **martinez** and **Eff** are overlapping.

3.3. Analytic indices of MARS. While Monte Carlo methods are robust in that their efficacy is not affected by the dimension of a problem, they are rather costly, especially when examining problems of large dimension. This challenge makes examining the properties of surrogate models in problems of high dimension very time consuming. To expedite our research, we developed an R routine to compute the exact Sobol' indices of MARS. To simplify the integrals that need to be evaluated, we've chosen to examine the analytic Sobol' indices for MARS models restricted to first-order interactions with each $x_i \sim$ i.i.d. on $\mathrm{U}(0,1)$ for all $i = 1, \ldots, 20$. Using the definition of Sobol' indices, the analytic indices of MARS are defined on the next page.

$$\hat{S}_j = \frac{\mathrm{Var}(\mathrm{E}(\hat{f}|x_j))}{\mathrm{Var}(\hat{f})} = \frac{\mathrm{E}(\mathrm{E}^2(\hat{f}|x_j)) - \mathrm{E}^2(\hat{f})}{\mathrm{E}(\hat{f}^2) - \mathrm{E}^2(\hat{f})}$$

$$\mathrm{E}(\hat{f}) = \int \left(\beta_0 + \sum_m \beta_m \mathrm{h}_m(\mathbf{x})\right) d\mathbf{x}$$

$$\mathrm{E}(\hat{f}^2) = \int \left(\beta_0 + \sum_m \beta_m \mathrm{h}_m(\mathbf{x})\right)^2 d\mathbf{x} = \int \left(\beta_0^2 + 2\beta_0 \sum_m \beta_m \mathrm{h}_m(\mathbf{x}) + \left(\sum_m \beta_m \mathrm{h}_m(\mathbf{x})\right)^2\right) d\mathbf{x}$$

$$= \beta_0^2 + 2\beta_0 \sum_m \beta_m \int \mathrm{h}_m(\mathbf{x}) d\mathbf{x} + \sum_m \beta_m^2 \int \mathrm{h}_m^2(\mathbf{x}) d\mathbf{x} + 2 \sum_m \sum_{k>m} \beta_m \beta_k \int \mathrm{h}_m(\mathbf{x}) \mathrm{h}_k(\mathbf{x}) d\mathbf{x}$$

$$\mathrm{E}(\mathrm{E}^2(\hat{f}|x_j)) = \iint (\beta_0 + \textstyle\sum_m \beta_m \mathrm{h}_m(\mathbf{x}))(\beta_0 + \sum_k \beta_k \mathrm{h}_k(\mathbf{x}')) d\mathbf{x} d\mathbf{x}'_{\sim j}$$

$$= \beta_0^2 + \beta_0 \sum_m \beta_m \iint \mathrm{h}_m(\mathbf{x}) d\mathbf{x} d\mathbf{x}'_{\sim j} + \beta_0 \sum_k \beta_k \iint \mathrm{h}_k(\mathbf{x}') d\mathbf{x} d\mathbf{x}'_{\sim j} + \sum_{m,k} \beta_m \beta_k \iint \mathrm{h}_m(\mathbf{x}) \mathrm{h}_k(\mathbf{x}') d\mathbf{x} d\mathbf{x}'_{\sim j}$$

$$= \beta_0^2 + 2\beta_0 \sum_m \beta_m \int \mathrm{h}_m(\mathbf{x}) d\mathbf{x} + \sum_{m,k} \beta_m \beta_k \iint \mathrm{h}_m(\mathbf{x}) \mathrm{h}_k(\mathbf{x}') d\mathbf{x} d\mathbf{x}'_{\sim j}$$

A property of MARS models allowing only one-degree interactions is that the first-order indices will sum to one and as a result lead to a definite error between the actual and surrogate model first-order indices. Therefore it is necessary to scale the first-order indices of the test function so that they sum to one, i.e.

$$\hat{S}_j = \frac{S_j}{\sum_{i=1}^{p} S_i}$$

These scaled indices provide a more meaningful comparison for how well MARS approximates the Sobol' indices as shown in Section 4.3.

4. Numerical Results.

4.1. Convergence of MARS to First Order ANOVA. For each test function, we explore the error between MARS with one degree of interaction, an additive model, and the first order ANOVA decomposition of the test function. Recalling the full ANOVA decomposition (Eq. 1.2) we may call the truncated ANOVA a function of the form

$$f_T(x) = f_0 + \sum_{i=1}^{p} f_i(x_i). \tag{4.1}$$

A theorem from Rabitz and Aliş (11) states that, for any additive function h in L^2,

$$\|f - f_T\|_2 \leq \|f - h\|_2 \tag{4.2}$$

In other words, the truncated ANOVA decomposition of a function may be considered the optimal additive representation. Since MARS of degree one is an additive model, we expect

that MARS will converge to the truncated ANOVA. To explore this, we use the L^2 error, represented by

$$\left\|\hat{f} - f_T\right\|_2 = \left(\int_{\mathbb{R}^p} (\hat{f} - f_T)^2 dx\right)^{\frac{1}{2}} \tag{4.3}$$

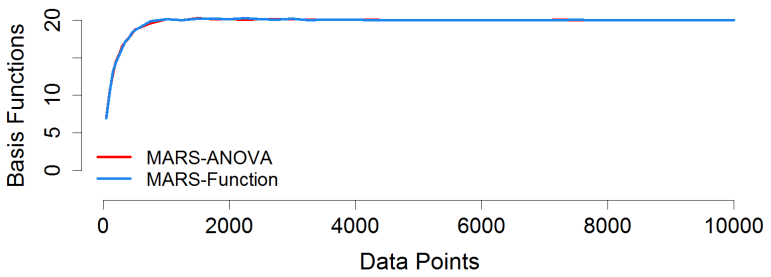
where f_T is the truncated ANOVA of the test function and $\hat{f}$ is the MARS approximation. The error was scaled by the L^2 norm of f_T to yield what will be referred to as the L^2 relative error.

Figure 4.1 demonstrates convergence of MARS to both the truncated first order ANOVA decomposition (red) and the g-function (blue) as the number of data points n is increased. Figure 4.1a demonstrates convergence of the number of basis functions MARS uses to model the data. Figure 4.1b demonstrates decreasing error between the MARS model and both the g-function and its first order ANOVA decomposition to a constant value. Figure 4.1c demonstrates the convergence rate of the relative error to a constant value. The minimum average relative error was taken as a horizontal asymptote for the relative error and subtracted from the relative error. The plot shows the log of this difference plotted against the log of the number of data points n.

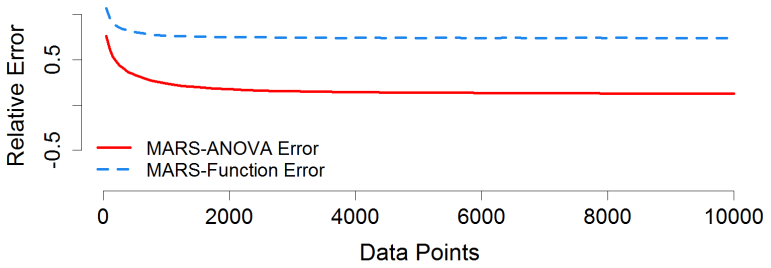
We see that the MARS model of the Sobol' g-function improves as the amount of available data increases, but that improvement quickly levels off and approaches a constant relative error for higher n. The limiting value for the approximation to ANOVA is about zero, which was consistent with our expectation since first degree MARS is additive. The limiting value of the error of the MARS approximation to the g-function is greater than 0.5, due to the complexity of the g-function. The reason for this leveling off can be seen in Figure 4.1a where the self-limiting constraints within MARS are reached and the number of basis functions is capped at about 20. In Figure 4.1c, a regression line is fitted to the log scaled axes to demonstrate near linearity of the convergence rate up to the capping of the basis functions. By inspection of 4.1a, we observe that the average number of data points at which MARS caps the basis functions is approximately 1000 to fit to both ANOVA and the g-function. The upper bound of the interval to which the line was fitted was thus chosen as $n = 1000$. The convergence rate on that interval is nearly linear until the maximum number of basis functions is reached, as shown by the Pearson Correlation Coefficients in Table 4.1. Beyond $n = 1000$, the relative error becomes much more noisy and the correlation coefficient is closer to zero. We observe similar trends with both the Gaussian and Oscillatory Functions.

	Pearson CC	Regression Line
MARS-ANOVA	-0.9807418	$1.955895 - 0.575647X$
MARS-Function	-0.9873448	$2.128972 - 0.785233X$

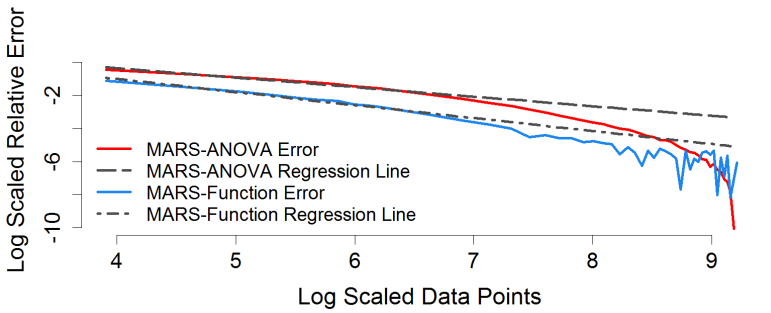
Table 4.1



(a)



(b)



(c)

Figure 4.1: (a): Capping of the number of basis functions used to model the g-function data. Both MARS-ANOVA and MARS-Function reach the cap at approximately $n = 1000$ data points. (b): Convergence of MARS to a constant relative error with respect to the g-function (blue) and the first order ANOVA decomposition of the g-function (red) as n is increased. (c): Log-log plot of (b) with regression lines fitted to the interval $[50, 1000]$. The rate of convergence of the model to both the function and its first order ANOVA up to the capping of the basis functions is nearly linear (see Table 4.1). Note: $\log 1000 \approx 6.9$, which is where the log-scaled relative error becomes noisier.

4.2. Convergence of First-Order Sobol' Indices for MARS. A question of particular interest is: how well does MARS capture variable importance with different amounts of available data? Using the analytic indices from Section 3.3, we discover that it depends. In Figure 4.2, the Sobol' indices are well approximated for the g-function. The order of variable importance is preserved and converges rather quickly. However, it is clearly evident from the second test case that particular problems, in this case the oscillating test function, compromise MARS' ability to preserve variable importance. We postulate that the introduction of oscillatory elements causes MARS difficulty. To test this, we performed an experiment of smaller dimension, $p = 5$, varying the c_i for each variable. In the case of $c = \{1, 1, 1, 1, 1\}$, the results were similar to that of the g-function: the indices were well approximated. However, further variations of the $c_i > 1$ showed that MARS performed worse as the c_i increased, suggesting that oscillating variables may be the cause.

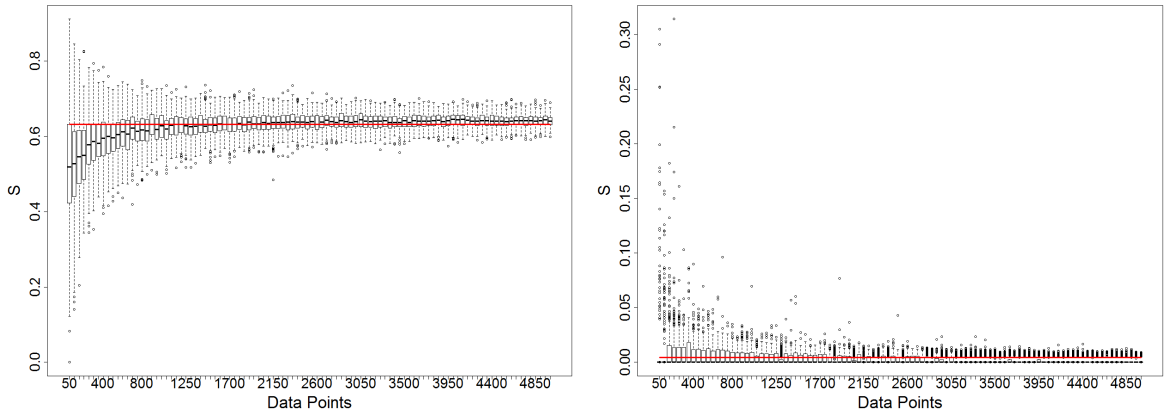
In the case of increasing dimension, we observe in Figure 4.3 that MARS is indifferent to the dimension of the problem. This is an encouraging result, as it demonstrates that Sobol' indices can be well approximated and variable importance conserved in especially high-dimensional problems. For this experiment, the g-function is utilized with a modified form of the a_i so that the first 10 variables are important and the remaining variables are unimportant.

$$a_i = \frac{i - \frac{30}{p}}{\frac{30}{p}}$$

4.3. Total Indices Approximations for MARS. We also explore how well Total Sobol' Indices (Section 1) were approximated based on MARS models. Recall from Section 2.1 that the degrees of interaction can be decided as an input argument to MARS. By allowing the MARS model to account for interactions between pairs and groups of variables in a given data set, we expect that the approximation of Total indices will improve. We first noted that for the g-function and Gauss function, the relative error between the model and the function decreased as the degrees of interaction were increased from one to two. This was reflected by a decrease in the error of S_j and T_j for these functions. However, the relative error and indices approximation only improve negligibly, at best, when degrees of interaction are further increased. We partially contribute this effect to the large amount of variance contributed by first and second order effects in both these functions. To demonstrate this, we define variance contributed by the ith order effect as follows:

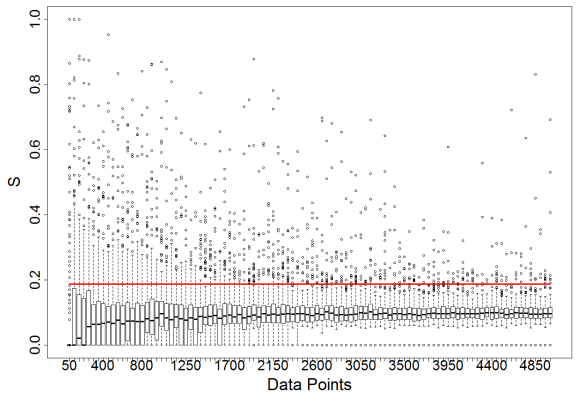
$$R_i = \frac{\sum_{j_1<j_2<\ldots<j_i} \int f^2_{j_1,j_2,\ldots,j_i}}{\mathrm{Var}(f)}$$

In the above definition, the sum is taken over all distinct subsets of $\{x_1, \ldots, x_p\}$ of size i. Also, $f_{j_1,j_2,\ldots,j_i}$ is a term from ANOVA, equation (1.2). For the Gauss function, $R_1 = .4262$, $R_2 = .3480$, and hence .7742 of the variance is explained by these two effects combined. The g-function has $R_1 = .5611$ and $R_2 = .3350$, explaining a combined .8961 of variance. We conjecture that the large value of $R_1 + R_2$ is a factor in explaining the negligible effects of increasing MARS degrees of interaction beyond 2 in these cases.

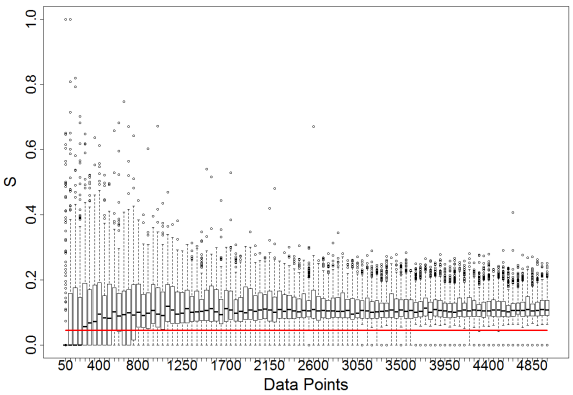


(a) Important variable, g-function.

(b) Unimportant variable, g-function.



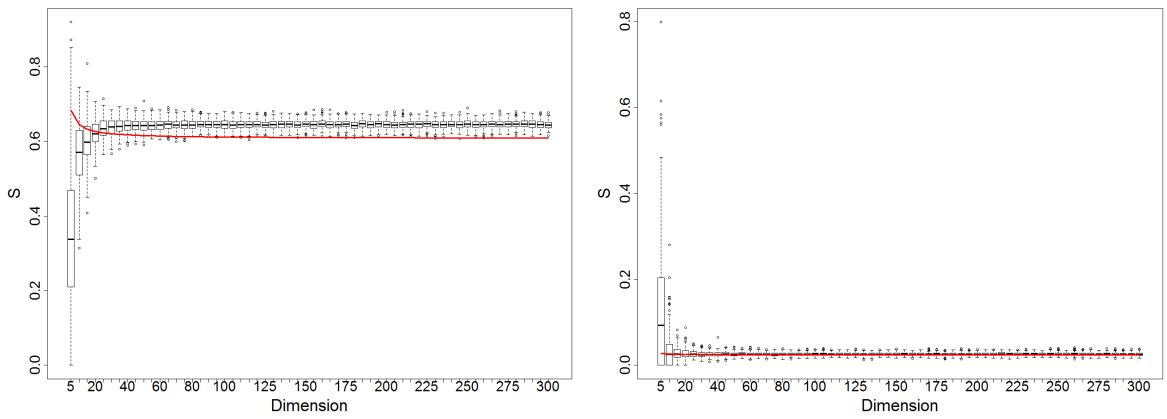
(c) Important variable, oscillating function.



(d) Unimportant variable, oscillating function.

Figure 4.2: Convergence of analytic MARS Sobol' indices to the scaled Sobol' indices as amount of data increases. The boxplots represent 200 samples of the analytic indices of MARS at each set of available data, which increases from data sample size of 50 to 5000 in steps of 50. The red line represents the scaled Sobol index for the corresponding variable. The Sobol' indices are well approximated for the g-function for important and unimportant variables. In the case of the oscillating test function, variables with high frequency cause MARS difficulty to preserve variable importance. Notice that the important variable is approximately the same value as the unimportant variable.

4.4. Convergence of Random Forest. We look at the function convergence of Random Forest for comparison with MARS which we computed in Section 4.1. Figure 4.5 shows the calculated L^2 relative error using the Sobol' g-function with respect to increasing values of n data points along the x-axis. Error is defined as in equation (4.3). We find that Random Forest



(a) Important variable, increasing dimension.

(b) Unimportant variable, increasing dimension.

Figure 4.3: Convergence of analytic MARS Sobol' indices to the scaled Sobol' indices as dimension increases. The boxplots represent 200 samples of the analytic indices of MARS at a particular dimension, which increases from dimension of 5 to 300 in steps of 5. The red line represents the scaled Sobol indice for the corresponding variable. MARS conserves and well-approximates variable importance regardless of dimension.

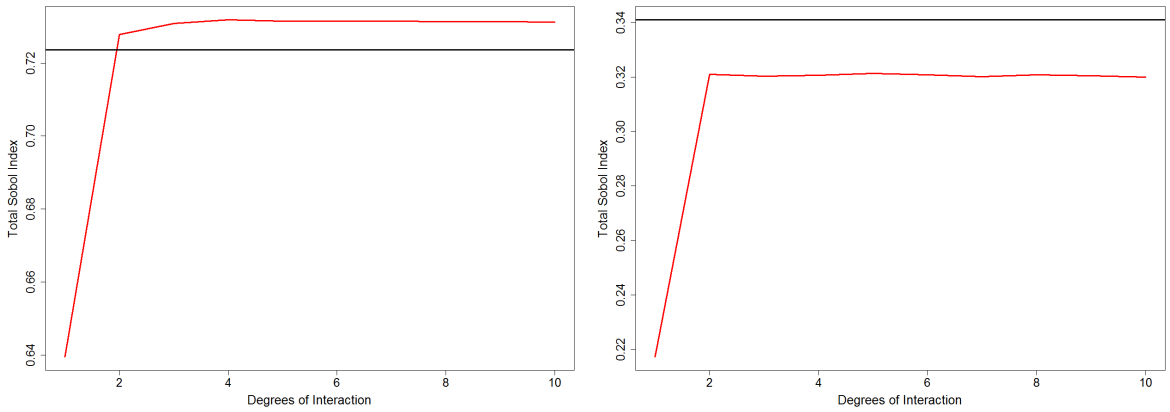


Figure 4.4: Convergence of total index approximation (red) to actual total index of most important variable (black) of g-function (left) and Gaussian Test Function (right) as degrees of interaction are increased. Increasing degrees of interaction to two results in a marked improvement in model based approximation of total indices for these two functions, but further increases do not seem to affect the approximation error.

does not converge by 20000 data points, whereas our previous results show MARS converges significantly faster at approximately 2000 data points. Random Forest does, however, reach significantly smaller values of L^2 relative error than MARS as available data increases. As a result this gives some indications that MARS is more appropriate for smaller data sets and Random Forest for larger data sets.

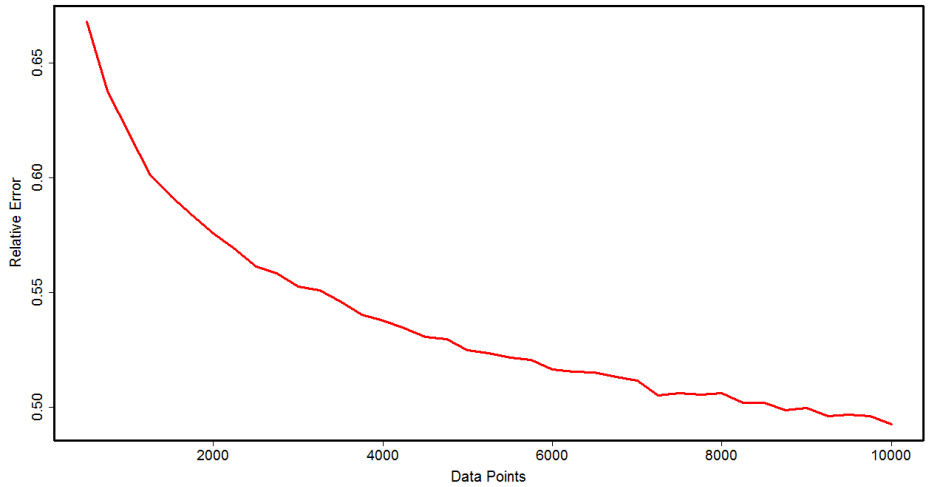
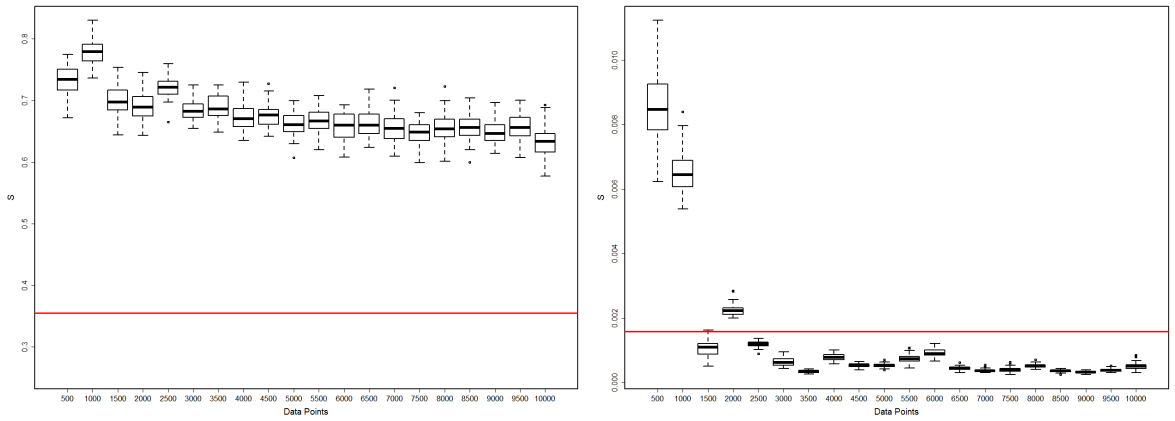


Figure 4.5: L^2 relative error convergence of Random Forest, approximated over 100 iterations for increasing values of n data points from 500 to 10000 by steps of 500.

4.5. Convergence of First-Order Sobol' Indices for Random Forest. Mirroring the trials performed in Section 4.2 we compute the first-order Sobol' indices for the test functions described in Section 3.1 using Random Forest, and examine the effects of increasing the amount of available data. Our results using the g-function as our test function indicate that as the amount of available data increases, approximation of Sobol' indices of important variables are overestimated (Figure 4.6a), and approximations of unimportant variables drop to values close to zero (Figure 4.6b). In comparison to the results preformed using MARS in Section 4.2 the approximation of important variables using Random Forest are substantially worse in their estimation than the MARS results. We obtain similar results using the Gaussian test function, but find that, similar to MARS, Random Forest does not capture variable importance for the Oscillatory test function. Our results indicate that for the test functions used in our examination, the use of MARS as a surrogate model outperforms the use of Random Forest in computation of Sobol' indices.



(a) Important variable, g-function.

(b) Unimportant variable, g-function.

Figure 4.6: Approximated Sobol' Indices of the g-function using Random Forest, as the number of avaliable data points increases.

5. Partial Derivative Measures of Sensitivity vs Sobol' Indices.

5.1. Introduction to Partial Derivatives as a Measure of Sensitivity. In this section we consider the sine squared test function (5.1).

$$f(\mathbf{x}) = \sin^2(2\pi\beta x_1) \prod_{i=2}^{15} \sin^2(2\pi x_i) \tag{5.1}$$

Each input x_i is i.i.d. and has the Uniform[0,1] distribution.
Moreover we also define a derivative based sensitivity index, D_j by the following:

$$D_j = \frac{I_j}{\sum_{k=1}^{p} I_k} \cdot \sum_{k=1}^{p} T_k \tag{5.2}$$

where

$$I_j = \int_{[0,1]^p} \left(\frac{\partial f}{\partial x_j} \right)^2 d\mathbf{x}$$

First, it is worth mentioning the reason behind the scaling in 5.2. We multiply by the sum of the total indices in 5.2 to ensure that D_j and T_j sum to the same value when the sum is taken over all indices. This allows us to make a fair comparison between the two sensitivity measures.

Now, it is natural to ask how these two measures of variable importance compare. If the the two measures always indicate the same order of variable importance for all functions then one should simply use whichever method is more convenient. Moreover, Sobol and Kucherenko show in (13) that the following relationship holds for functions f which satisfy $\frac{\partial f}{\partial x_j} \in L^2$:

$$T_j \leq \frac{\int_{[0,1]^p} \left(\frac{\partial f}{\partial x_j}\right)^2 d\mathbf{x}}{\mathrm{Var(f)}\pi^2}$$

Moreover, this bound is optimal in the sense that we can find a function for which the inequality becomes equality. The fact that the upper bound includes an integral of a partial derivative squared may indicate that the two measures are at least tangentially related.

We are interested in whether the two methods preserve order of variable importance. That is, are variables that are considered to be important(or unimportant) by one method always judged to be this way by the other? The answer to this is no, as evidenced by Figure 5.1.

In this section, we also demonstrate the perils of trying to approximate derivative sensitivity measures in high dimensions. A commonly used routine that estimates the partial derivatives of a function is Morris Screening, described in (8). We used Morris Screening to estimate I_1 in the g-function example. The routine is given access to a MARS approximation of the g-function. We then compare the Morris screening approximation of a derivative index with the true value. The results show the Morris screening approximation is very unreliable in this instance.

5.2. Summary of Derivative Sensitivity Results. Computations on the sine squared test function (5.1) demonstrate that the two measures of variable importance do not always agree. This function is a product of 15 univariate components which are each expressions involving sine squared. The first component has a frequency, β, which is a parameter we can control. As β is allowed to vary, T_1 remains very close to constant, particularly when $|\beta|$ is large. Every total index is approximately equal to .34 for all values of β, and hence the variables are deemed to be of equal sensitivity.

However, when variable importance is judged by a partial derivative method, we get a contradictory result. The derivative index of x_1, D_1, is greatest for $|\beta|$ large, as demonstrated by Figure 5.1. As $|\beta|$ increases, the derivative approach assigns x_1 a sensitivity that approaches $\sum_{j=1}^{15} T_j$. This means the weighted relative importance of x_1 is approaching 100%. Hence the derivative sensitivity measure weights x_1 as substantially more important than the other variables for these β values.

This intuitively makes sense because I_1 increases as the frequency of oscillations is increased. A derivative approach to sensitivity in a sense 'values' these oscillations while in the perspective of variance based sensitivity the oscillations average out and do not change the underlying conditional expectation, $\mathbb{E}(f|x_1)$.

It should be mentioned that there are instances where a derivative approach and the total Sobol' index agree in measuring sensitivity. One example of this is the g-function. In Figure 5.2, we perform computations on the g-function with modified parameters [1]. We find that the two methods give the same value for sensitivity of variable x_1 as parameter β is changed.

[1] $a_i = \{\beta, 0, 0, 0, 0, 0, 0, 0, 0, 3, 3, 3, 3, 3\}$

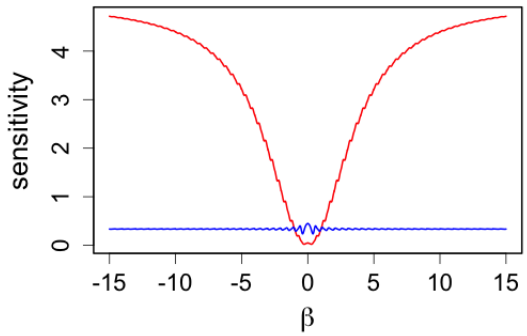


Figure 5.1: Sine squared test function. The Derivative Index, D_1 (red), assigns x_1 greater sensitivity for large $|\beta|$ while T_1 remains close to constant (blue).

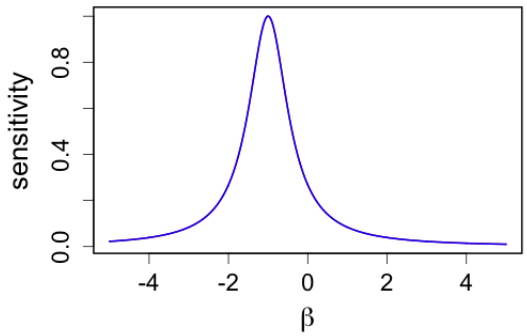


Figure 5.2: g function. T_1 identical to D_1 for all values of the parameter β

The differences in the methods notwithstanding, we also examine how well partial derivatives may be estimated in general at higher dimensions. We find a standard Morris screening routine does not approximate I_1 well when given access to a MARS surrogate model of the g-function. In Figure 5.3 we vary parameter β and build a MARS model from 1000 sampled data points. We then estimate the appropriate partial derivatives using a Morris screening routine to compare the approximated quantity $\frac{I_1}{\sum_{k=1}^{p} I_k}$ with the true value. Figure 5.3 shows the partial derivatives are not well approximated, especially when $a_1 > 0$. There are unpredictable overestimates of the partial derivative which indicate the Morris screening is volatile. This volatility may be compounded by the high dimension of the problem. This example demonstrates potential drawbacks of the derivative approach.

The derivative and variance based methods of global sensitivity measure two different

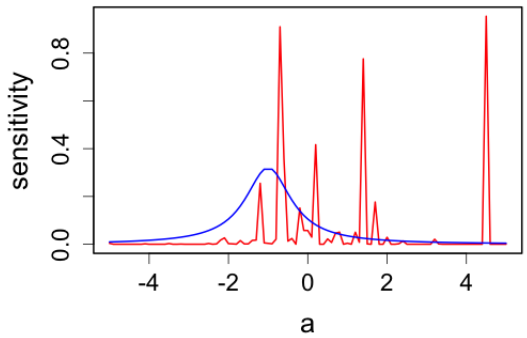


Figure 5.3: g-function. Morris screening approximation (red) of $\frac{I_1}{\sum_{k=1}^{p} I_k}$ when given access to only MARS model poorly estimates true value (blue).

quantities. They will not, in general, agree in determining the order of variable importance. Thus caution is advised when selecting a method of sensitivity analysis. One should carefully consider which type of variable sensitivity they wish to measure. This decision will probably be problem dependent. Moreover, there are limitations to the derivative method caused by the inaccuracy of approximating partial derivatives in high dimensions.

6. Conclusion. We found that variable importance and S_i can be well approximated even when there is significant error in a MARS approximation. S_i were also observed to be well approximated for problems ranging from 5-300 dimensions and T_i are well approximated using 2nd degree MARS models. Comparison of derivative measures to variance-based measures of sensitivity revealed that there are functions for which the measurements do not agree. Our results also showed that Morris screening, a method using finite difference approximations, can be unreliable when performed on a surrogate model.

Future work could determine what properties of MARS cause it to resolve the SI well despite having a large function error, and what MARS options will result in the optimal approximation of the SI. Exploration of how function properties cause differences between variance and derivative sensitivity measurements could lead to further understanding of when to choose a particular sensitivity measurement and yield results that are more meaningful to the problem at hand.

REFERENCES.

[1] Leo Breiman, *Bagging Predictors*, Machine Learning, 24 (1996), pp. 123–140.
[2] ——, *Random forests*, Machine Learning, 45 (2001), pp. 5–32.
[3] Russel E. Caflisch, *Monte carlo and quasi-monte carlo methods*, Acta Numerica, (1998), pp. 1–49.
[4] Jerome H. Friedman, *Multivariate adaptive regression splines*, The Annals of Statistics, 19 (1991), pp. 1–67.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning, Second Edition*, Springer, 2008.
[6] Alexandre Janon, Maelle Nodet, and Clementine Prieur, *Uncertainties assessment in global sensitivity indices estimation from metamodels.*, (2011).
[7] Andy Liaw and Matthew Wiener, *Classification and regression by randomforest*, R News, 2 (2002), pp. 18–22.
[8] Max D. Morris, *Factorial sampling plans for preliminary computational experiments*, Technometrics, 33 (1991), pp. 161–174.
[9] Art B. Owen, *Better estimation of small sobol' sensitivity indices*, ACM transactions on mathematical software, (2012).
[10] Gilles Pujol, Bertrand Iooss, and Alexandre Janon, *sensitivity: Sensitivity Analysis*, 2015. R package version 1.11.1.
[11] Herschel Rabitz and Ömer F. Aliş, *General foundations of highdimensional model representations*, Journal of Mathematical Chemistry, 25 (1999), pp. 197–233.
[12] I.M. Sobol', *Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates*, Mathematics and Computers in Simulation, 55 (2001), p. 271–280.
[13] I.M. Sobol' and Sergei Kucherenko, *Derivative based global sensitivity measures and their link with global sensitivity indices*, Mathematics and Computers in Simulation, (2009).
[14] Curtis B. Storlie, Laura P. Swiler, Jon C. Helton, and Cedric J. Sallaberry, *Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models*, (2008).