Structure and Infrastructure Engineering, Vol. 2, Nos. 3–4, September–December 2006, 175–189

The promise and peril of uncertainty quantification using response surface approximations

A. A. GIUNTA*†, J. M. MCFARLAND‡, L. P. SWILER† and M. S. ELDRED†

†Sandia National Laboratories, PO Box 5800, Mail Stop 0828, Albuquerque, NM 87185-0828, USA
‡Vanderbilt University, Department of Mechanical Engineering, VU Station B 351592, 2301 Vanderbilt Place, Nashville, TN 37235, USA

(*Received 6 August 2005; accepted in revised form 2 December 2005*)

Conventional sampling-based uncertainty quantification (UQ) methods involve generating large numbers of random samples on input variables and calculating output statistics by evaluating the computational model for each set of samples. For real world applications, this method can be computationally prohibitive due to the cost of the model and the time required for each simulation run. Using response surface approximations may allow for the output statistics to be estimated more accurately when only a limited number of simulation runs are available. This paper describes an initial investigation into response surface based UQ using both kriging and multivariate adaptive regression spline surface approximation methods. In addition, the impact of two different data sampling methods, Latin hypercube sampling and orthogonal array sampling, is also examined. The data obtained from this study indicate that caution should be exercised when implementing response surface based methods for UQ using very low sample sizes. However, this study also shows that there are clear cases where response surface based UQ provides a gain in accuracy versus conventional sampling-based UQ methods.

Keywords: Uncertainty quantification; Response surface; Kriging; Multivariate splines; Latin hypercube; Orthogonal array

1. Introduction

Simulation-based uncertainty quantification (UQ) for real-world engineering applications often is difficult due to the computational expense of the simulations (a.k.a. 'code runs'). For example, current high-fidelity finite element models used in industry often involve millions of elements and require hours or days of compute time on a massively parallel computer for a single simulation run. The simulated physics is complex, as it often involves coupled systems of nonlinear, time-dependent partial differential equations that arise in disciplines such as fluid flow, structural mechanics, and radiation transport. Phenomena such as shock waves, multi-material contact (e.g. crushing), and random scattering are common in such simulations. Typically, each simulation involves tens to hundreds of uncertain input parameters involving various types of probability density functions.

Due to the complexity and nonlinearity of these simulations, analytic solutions for UQ statistics such as mean value, standard deviation, and probability of failure on the simulation outputs are almost never possible. In some fortunate scenarios, analytic reliability methods can be used for calculating failure probabilities, especially if the limit state function is linear or weakly nonlinear, and the uncertain parameters have normal distributions. However, in many scenarios the limit state function is nonlinear and/or the uncertain parameters are non-normal, and thus

*Corresponding author. Email: aagiunt@sandia.gov

Structure and Infrastructure Engineering
ISSN 1573-2479 print/ISSN 1744-8980 online © 2006 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/15732470600590507

analytic reliability methods are more prone to inaccuracy. In such cases, often the only viable UQ approach involves sampling-based methods. Due to budget considerations, one can usually afford only a limited number of simulation runs and, of course, one would like to extract as much information as possible from this limited amount of data. Accordingly, this study attempts to answer the following question, 'Given a set of N simulation output data samples, is it more accurate to estimate UQ statistics directly from the N data points, or is it more accurate to extract trend information from the N data points, and then estimate UQ statistics from the trend models?'.

The use of trend models (a.k.a. response surface approximations, metamodels, surrogate models) to understand complex phenomena has its origins in the statistical design of experiments literature (cf. Myers and Montgomery 1995). Over the past 30 years, response surface approximations have found increasing use in engineering design optimization studies (cf. Simpson *et al.* 2001) and are becoming increasingly common in engineering UQ studies as well (cf. Gomes and Awruch 2004, Romero *et al.* 2004, Kaymaz 2005). Indeed, it seems completely reasonable to employ some sort of trend model for UQ purposes when data are scarce. To date, only a few comprehensive surveys of response surface approximation methods have been published (cf. Jin *et al.* 2001). However, these studies typically focus on function approximation accuracy and not on statistical metric accuracy, thus motivating this study.

Due to the generality of the engineering UQ problems being addressed (nonlinear physics simulations involving non-normal uncertain parameters), there is limited insight to be gained by a strictly theoretical investigation. Rather, a more pragmatic approach is chosen in which a combination of different UQ sampling methods and response surface approximation methods is used to analyze a well-known test function. While the information gained from this initial study is largely qualitative and of limited general applicability, it is expected that the insights gained on these simple test problems will lead to more interesting research topics, and will provide some guidance for performing UQ in engineering design applications.

This study employed the publicly available DAKOTA (Design Analysis Kit for Optimization and Terascale Applications) software toolkit (Eldred *et al.* 2001) for both sampling-based UQ methods and response surface approximation methods. That is, DAKOTA can be used to estimate UQ statistics directly from N data samples, and it also can be used to create response surface approximations from the N data samples from which approximate UQ statistics can be obtained. Additional detail on DAKOTA is provided below.

2. Background

2.1 *Sampling methods for UQ*

This study investigates three sampling methods that generate data for UQ. The sampling methods are Monte Carlo (MC) sampling (Metropolis and Ulam 1949), Latin hypercube (LH) sampling (McKay *et al.* 1979), and orthogonal array (OA) sampling (Owen 1992, Hedayat *et al.* 1999).

The MC, LH, and OA sampling methods are based on random sampling according to user-prescribed joint probability density functions on the uncertain parameters. However, the LH and OA sampling methods are refinements to basic MC sampling in that they partition the parameter space into bins of equal probability, with the goal of attaining a more even distribution of sample points in the parameter space than typically occurs with MC sampling. For example, given N sample points in an n-dimensional parameter space, LH sampling partitions the parameter space into an N^n grid of bins, with N bins along each axis. Then, sample points are distributed such that all one-dimensional projections of the samples yield one sample per bin (figure 1). Note that the LH sampling method used in

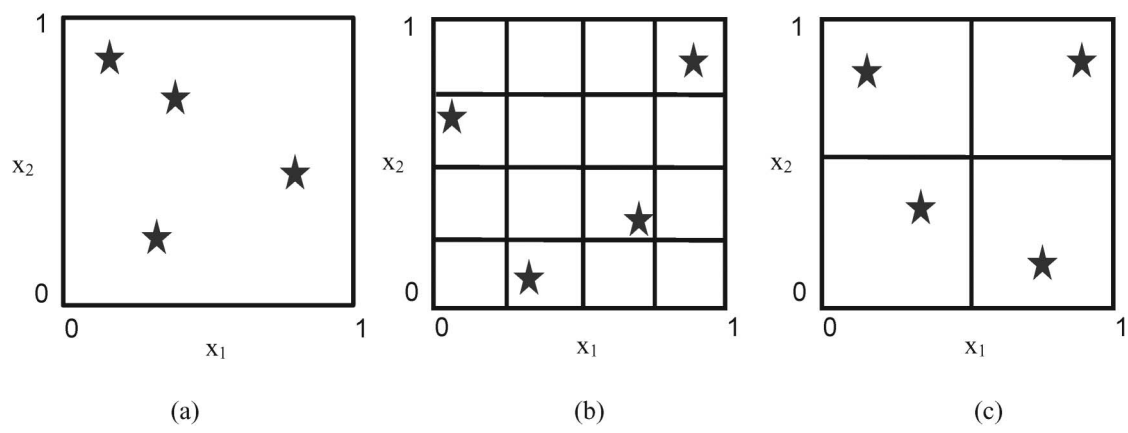

Figure 1. Examples of (a) MC sampling, (b) LH sampling, and (c) two-dimensional OA/Stratified sampling.

this study placed the sample randomly within each selected bin, although some LH sampling methods place the sample at the center of each selected bin. The bin-centered LH sampling method was not investigated in this study.

Orthogonal arrays have been developed to study main effects by creating experimental design matrices where the columns (the levels or settings of the uncertain variables) are orthogonal, meaning statistically independent. For each level of one variable, all levels of the other variables occur an equal number of times. This allows one to test if the influence of variable X, for example, at level A has a significant influence on the output vs the setting of variable X at level B. The OA sampling method is similar to the LH approach in that it employs a partitioning scheme to subdivide the n-dimensional parameter space into a grid of equal-probability bins. Specifically, in OA sampling the n-dimensional parameter space is partitioned into a $\left(\sqrt[t]{N}\right)^n$ grid of bins, with $\sqrt[t]{N}$ bins along each axis. In an OA, the samples are distributed such that all t-dimensional projections (for $t < n$) of the samples yield at least one sample per bin. This produces stratified random sampling (one sample per bin) in all t-dimensional subsets of the n-dimensional parameter space. The term t is called the 'strength' of the OA. Due to the special structure of the OA sampling, it is not always possible to create a set of OA samples for an arbitrary value of N. Tables of commonly used OA samples sets for various combinations of N, n, and t values are available (cf. Hedayat *et al.* 1999). Note that the LH sampling approach is equivalent to a strength-1 OA. For the special case where the dimension of the test function and the strength of the OA are the same value (e.g. $n = t = 2$ in this study), the OA produces stratified random sampling on a $\sqrt{N} \times \sqrt{N}$ grid of equal-probability bins. It should be stressed, however, this is not the case for a general n-dimensional test function and strength-t orthogonal arrays. The introduction of OA sampling in this study where $n = 2$ is carried out in preparation for future studies involving test problems with $n > 2$.

Figure 1 illustrates the difference between MC, LH and OA/Stratified sampling, where, for simplicity, we have two ($n = 2$) uniform uncorrelated uncertain parameters x_1 and x_2 and four samples ($N = 4$) on the parameter space $[0,1]^2$. For MC sampling, the stars represent randomly selected sampling locations in the parameter space. For LH sampling and OA/Stratified sampling, the grid lines inside the parameter space denote equal-probability bins. Due to the assumption of uniform uncorrelated marginal distributions, all of the bins are of equal area as well. The LH sampling approach generates a grid of 4×4 bins, and four bins are chosen such that each row or column contains exactly one sample. The OA sampling approach with $n = t = 2$ creates stratified MC sampling on a grid of 2×2 bins, with the one sample placed randomly in each bin.

2.2 *Response surface approximation methods*

This study employs two general response surface approximation methods: kriging interpolation and multivariate adaptive regression splines (MARS). Each is discussed below.

2.2.1 Kriging interpolation. Kriging interpolation techniques were originally developed in the geostatistics and spatial statistics communities to produce maps of underground geologic deposits based on samples obtained at widely and irregularly spaced borehole sites (Cressie 1991). The basic notion that underpins kriging is that the sample response values exhibit spatial correlation, with response values modeled via a Gaussian process around each sample location (i.e. samples taken close together are likely to have highly correlated response values, whereas samples taken far apart are unlikely to have highly correlated response values). Kriging methods have found wide utility due to their ability to accommodate irregularly spaced data, their ability to model general surfaces that have many peaks and valleys, and their exact interpolation of the given sample response values.

The specific form of the kriging model used in this study is described in Giunta and Watson (1998) and Romero *et al.* (2004), and it is based on the work of Koehler and Owen (1996). The form of the kriging model is

$$\hat{f}(x) = \hat{\beta} + r(x)^T R^{-1}(f - \hat{\beta}\{1\}) \tag{1}$$

where $\hat{\beta}$ is the generalized least squares estimate of the mean response; $r(x)$ is an $N \times 1$ vector of correlations between the current point x, and all N sample sites in parameter space; R is the $N \times N$ correlation matrix of all N sample sites; f is the vector of N sample site response values; and $\{1\}$ is an $N \times 1$ vector with all values set to unity. The terms in the correlation vector and matrix are computed using a Gaussian correlation function. The ith term in $r(x)$ is given by

$$r_i = \exp\left[-\sum_{k=1}^{n} \theta_k |x_k - x_k^{(i)}|^2\right] \tag{2}$$

and, similarly, the i, j^{th} term in R is given by

$$R_{i,j} = \exp\left[-\sum_{k=1}^{n} \theta_k |x_k^{(i)} - x_k^{(j)}|^2\right] \tag{3}$$

where n is the dimension of the parameter space; k is the index on the dimension of the parameter space; $i = 1, \ldots, N$; $j = 1, \ldots, N$; and θ is the $n \times 1$ vector of correlation parameters. In this study, all values of θ are set to unity in order to simplify the use of the kriging method. In the most general approach to kriging, the values of θ are computed using maximum likelihood estimation.

However, finding optimal θ values can be problematic, as it requires an iterative search scheme that sometimes fails to converge.

Note that the kriging interpolation method is prone to ill-conditioning in the correlation matrix R as the number of sample points increases. This occurs because of the distance measure that is computed in equation (3). As the distance between any two sample points i and j decreases, then the ith and j^{th} rows in matrix R become linearly dependent, and in the limit where the points are the same, the matrix R becomes singular. Thus, this basic kriging method works well for a sparse set of sample points in an n-dimensional parameter space, but as the number of samples increases (and the inter-point distances decrease), the kriging method becomes unstable.

2.2.2 Multivariate adaptive regression splines. The multivariate adaptive regression splines (MARS) function approximation method (Friedman 1991) is based on a complex, recursive partitioning algorithm involving truncated power spline basis functions. The form of the MARS model is:

$$\hat{f}(x) = a_o + \sum_{m=1}^{M_1} a_m B_m(x_i) + \sum_{m=1}^{M_2} a_m B_m(x_i, x_j) + \ldots \qquad (4)$$

where the B_m terms are the basis functions, the a_m terms are the coefficients of the basis functions, M_1 is the number of one-parameter basis functions, and M_2 is the number of two-parameter basis functions. The MARS software allows the user to select either linear or cubic spline basis functions. Cubic spline basis functions are used for this study. The regression aspect of the MARS algorithm involves a forward/backward stepping process to adaptively add/remove spline basis functions from the model. It is this regression process that generates the a_o and a_m terms in equation (4). The resulting MARS model is a C^2-continuous function of piecewise cubic splines, but it will not exactly interpolate the data points that were used in calculating the coefficients. Thus, like polynomial regression, MARS has the ability to create smooth approximations to noisy data. Unlike kriging, MARS appears to have no upper limit on the number of samples that can be used in the function approximation process.

Admittedly, the mathematical and statistical algorithms within MARS are somewhat of a black box, given the lack of a detailed user's manual. However, we elected to use MARS in this study since it is one of the few publicly available function approximation software packages that, like kriging, can handle arbitrarily spaced data. In addition, MARS has received some use in the engineering community (cf. Jin *et al.* 2001) and we were interested in exploring both its utility and its drawbacks.

2.3 *Software tools*

As noted above, the DAKOTA toolkit (Eldred *et al.* 2001) was used in this study. DAKOTA provides an integrated suite of data sampling methods and response surface approximation methods. It is a publicly available software package that can readily be employed for similar scientific studies.

Within DAKOTA, the MC and LH sampling capabilities are provided via the LHS software library (Iman and Shortencarier 1984, Swiler and Wyss 2004). The OA sampling capability is provided via the DDACE library (Martinez-Canales and Williams 2005), which leverages the original 'oa.c' software package created by Prof. Art Owen of Stanford University (Owen 1992).

The kriging interpolation method in DAKOTA is an extension of the work of Giunta and Watson (1998). The MARS approximation method is available through the DDACE library, which leverages the original 'mars36.for' software package (MARS version 3.6) created by Prof. Jerome Friedman of Stanford University (Friedman 1991). The 'mars36.for' program was obtained from the StatLib online software repository.

This study employed DAKOTA version 3.3 on a 32-bit Intel microprocessor-based computer workstation running the Red Hat Linux Fedora Core 3 operating system. This version of DAKOTA is available to the public, under the restrictions of the GNU General Public License, from http://endo.sandia.gov/DAKOTA. The DAKOTA input files used in this study are listed in Appendix A and Appendix B.

3. Technical approach

3.1 *Test function*

Figure 2 shows a plot of the Rosenbrock function (Gill *et al.* 1981) that is used in this study since it provides a simple, computationally inexpensive response function that exhibits some of the nonlinear trends often found in data from computationally expensive engineering/physics simulation codes (cf. computational shock physics studies in Giunta *et al.* 2002). The formula for the Rosenbrock function is:

$$f(x_1, x_2) = 100(x_2 - {x_1}^2)^2 + (1 - x_1)^2 \qquad (5)$$

where both x_1 and x_2 are bounded by $[-2, 2]$. In this study, three statistics of f are examined: the mean, the standard deviation, and the failure probability, p_f, which is defined as the probability that $f < 4$. Figure 3 illustrates the portion of the parameter space that falls below this threshold. Given independent uniform

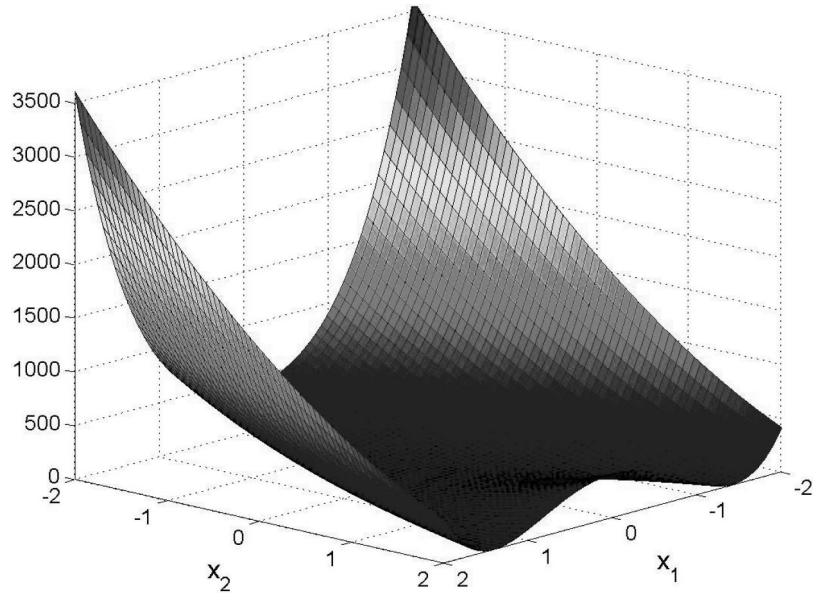

Figure 2. Rosenbrock's function.

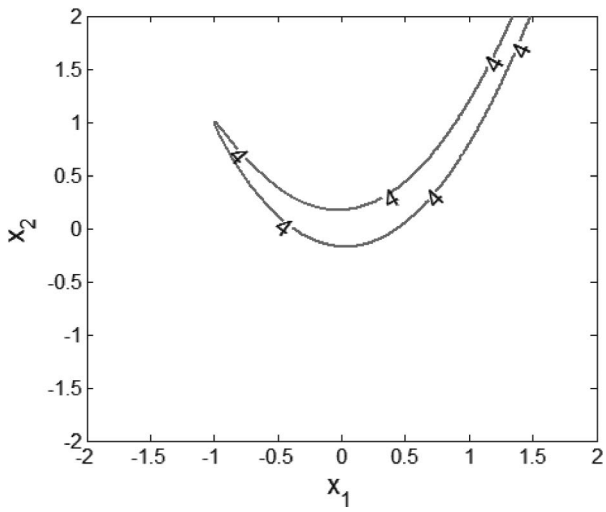

Figure 3. Contour plot of Rosenbrock's function at the failure threshold value.

distributions for the input parameters, the mean and standard deviation of $f(x_1, x_2)$ over the region $[-2, 2]^2$ are approximately 455.67 and 606.56, respectively, which were calculated analytically. The failure probability is approximately 0.04940 (approximately 5%), which was estimated using one million MC samples of the Rosenbrock function.

Although the true functional form of f is known, this study does not take advantage of this knowledge, since in general engineering applications the form of f is unknown. Also, it is important to note that the probability density function (PDF) of f is generally not known. Clearly, when the form of f is unknown, simply knowing the PDFs of the input parameters does not in itself provide that the PDF of the output f is known. Thus, the mean, standard deviation, and failure probability must be estimated solely from samples of f. Note that analytic reliability methods are not addressed in this study. Due to the nonlinearity of the limit state function (see figure 3), analytic reliability methods would be inaccurate.

3.2 *Statistical metrics*

The mean value of function f is estimated using sampled values of f via

$$\tilde{\mu} = \frac{1}{N}\sum_{i=1}^{N} f(x_1^{(i)}, x_2^{(i)}) \tag{6}$$

where N is the number of samples, and the values for x_1 and x_2 are drawn randomly from their respective PDFs.

When the mean value is estimated using samples taken from a response surface approximation, the mean value is denoted as

$$\hat{\mu} = \frac{1}{M}\sum_{i=1}^{M} \hat{f}(x_1^{(i)}, x_2^{(i)}), \tag{7}$$

where $\hat{f}$ is the response surface approximation function computed from N data points, and M is the number of samples taken of the response surface approximation (note: $M \gg N$). That is, $\hat{f}$ is either a kriging model or a MARS model that is computed from the N original data samples.

Once the $\hat{f}$ approximation is constructed, $M = 10\,000$ samples are used to estimate the mean value. It is important to note that in certain cases where $\hat{f}$ has a simple functional form (e.g. polynomial response surfaces), it is easy to analytically integrate the $\hat{f}$ function to obtain the mean value. However, the functional forms of the kriging and MARS methods are not readily amenable to analytic integration as we are treating these approximation functions as 'black boxes' inside the DAKOTA software.

Following the mean value estimation approaches given above, the standard deviation of the function f is estimated directly from the samples as

$$\tilde{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}[f(x_1^{(i)}, x_2^{(i)}) - \tilde{\mu}]^2} \tag{8}$$

and using the response surface approximations as

$$\hat{\sigma} = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}[\hat{f}(x_1^{(i)}, x_2^{(i)}) - \hat{\mu}]^2} \tag{9}$$

The failure probability, $P(f < 4)$, is given by the integral

$$p_f = \int_{\substack{-2 \\ f(x_1,x_2)<4}}^{2}\int_{-2}^{2} h_X(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \tag{10}$$

where $h_X(x_1, x_2)$ is the joint probability density of x_1 and x_2 and $h_X(x_1, x_2) = 1/16$ because x_1 and x_2 are independent uniform random variables. Equation (10) is difficult to evaluate analytically because of the complex integration region (see figure 3). Thus, a very large number of random samples were generated to approximate the true value of p_f. This approximation is given by

$$\tilde{p}_f \approx \frac{N_{fail}}{N}, \tag{11}$$

where N_{fail} is the number of samples in the failure region. Similarly, the failure probability is estimated using the response surface approximations as

$$\hat{p}_f \approx \frac{M_{\mathrm{fail}}}{M}, \tag{12}$$

where M_{fail} is the number of samples in the failure region of the response surface approximation function.

Haldar and Mahadevan (2000) show that the error in the failure probability approximation can be estimated as

$$\varepsilon\% = \sqrt{\frac{(1 - p_f^T)}{N \times p_f^T}} \times 200\% \tag{13}$$

where p_f^T is the true failure probability. Thus, the failure probability estimate of 0.04940 for the Rosenbrock function (where failure occurs for $f < 4$) made using one million random samples probably has an error less than 0.87%. For the purpose of this paper, the true value of p_f is assumed to be within 0.04940 ± 0.00043.

3.3 *Sampling methods*

The MC and LH sampling methods allow the user to choose an arbitrary number of sample points. For this study, we used MC and LH sample sizes of 10, 25, 50, 75, 100, and 121 points. Note that all samples are randomly generated, and no set of samples is a subset of any other set. Due to the unique structure of strength-2 OA sampling, the sample sizes are limited to specific values. In this study we used OA sample sizes of nine, 16, 25, 49, and 121. Note that in practice, one need not be limited to the specific sample sizes imposed by the OA structure. For example, with a budget of N samples, one could find the largest OA of K samples where $K < N$, and then generate the remaining $N-K$ samples with MC sampling.

3.4 *Surface approximation methods*

This study first examined statistical data generated using MC, LH, and OA/Stratified sampling alone without employing response surface approximation. This was carried out in order to illustrate the variance reduction offered by both LH and OA sampling over traditional MC sampling. This was followed by an examination of the statistical data found using response surface approximations. Since one would obviously employ LH or OA sampling rather than MC sampling, only response surfaces based on LH sample points and OA/Stratified sample points were investigated.

Note that $M = 10\,000$ LH samples are used in evaluating equations (7), (9), and (12) for this study. This sample size was selected somewhat arbitrarily, based on a few trials performed with M in the range of 5000 – 50 000 samples. Over this range, the estimates for mean value, standard deviation, and failure probability showed negligible variation. Thus, while there is some error in the statistical metrics introduced via a finite sample size, this error should be much smaller than the error introduced by either the kriging or MARS response surface approximation methods. In this study, $M > N$ by several orders of magnitude.

4. Results and discussion

4.1 *Comparison of sampling methods*

Figure 4 shows the trends in the estimated mean value of the Rosenbrock test function for MC, LH and OA/Stratified sampling. The MC and LH data were obtained for $N = 10$, 25, 50, 75, 100, and 121 samples; and the OA/Stratified samples were obtained for $N = 9$, 25, 49, and 121

samples. For each value of N, 10 runs (i.e. unique trials of N samples) were performed in order to examine the spread of the mean value estimate. Recall that equation (6) is used in these mean value estimates. As expected, both LH and OA/Stratified sampling show, generally, less spread (variance) in the estimated means for $N > 25$, although for $N = 121$ the variance reduction of LH versus MC sampling may not be statistically significant. Note that for $N = 10$ (MC and LH) and for $N = 9$ (OA) samples, there is no qualitative difference in the spread of the mean values predicted by the three sampling methods. For $N = 121$, there appears to be a qualitative difference between LH and OA/Stratified sampling, with the OA/Stratified approach having less spread in the mean value estimates.

Figure 5 shows the standard deviation estimates produced for the three sampling methods for increasing N, where standard deviation was computed via equation (8). From these results it is difficult to differentiate the performance of the three sampling methods, but for $N = 121$, OA sampling appears to be the most accurate.

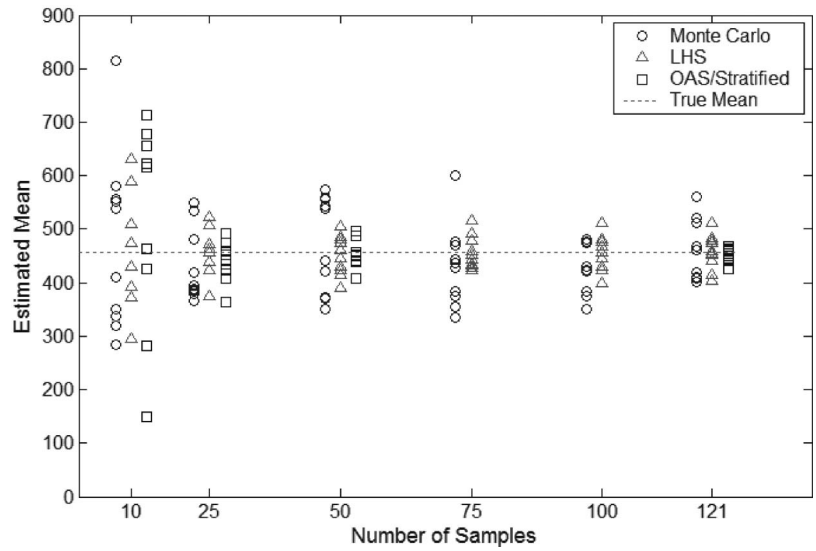

Figure 4. Comparison of mean value estimation using MC, LH, and OA (Stratified) sampling. (LHS, LH sampling; OAS, OA sampling).

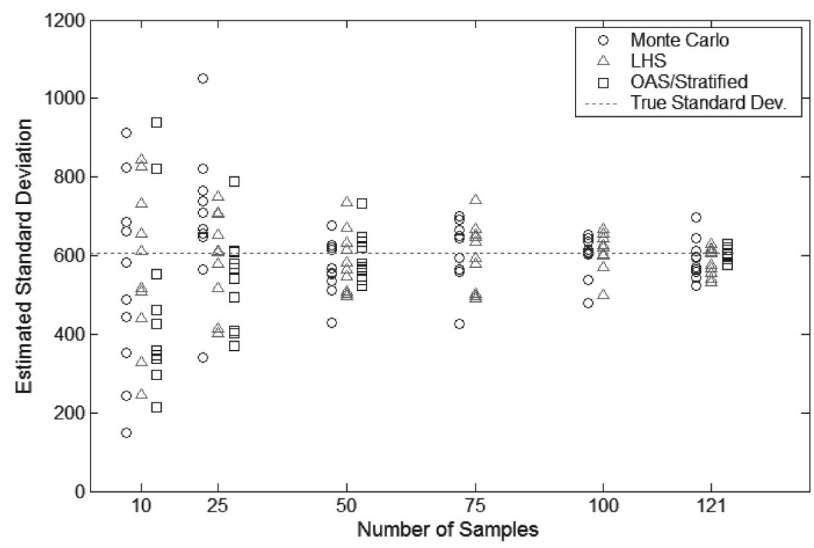

Figure 5. Comparison of standard deviation estimation using MC, LH, and OA (Stratified) sampling.

Figure 6 shows the failure probability estimates for increasing N, where the failure probability was computed using equation (11). Although 10 trials were made for each value of N, in some cases duplicate failure probability estimates were obtained due to the discrete nature of equation (11). For all values of N, there is no clear difference in the spread of the failure probability estimates produced by the three sampling methods.

The data in figure 6 illustrate the potential error in using equation (11) to estimate failure probabilities, especially for low values of N. That is, in several of the replicates for $N \leq 50$, the estimated failure probability was zero since none of the sample points landed in a region where $f < 4$ for the Rosenbrock function (recall, the true failure probability is 0.0494).

4.2 *Function approximations based on LH samples*

Figures 7–9 show, respectively, the trends in mean value, standard deviation, and failure probability for increasing N as estimated using kriging and MARS surface approximations based on LH sampling. Recall that the kriging and MARS models are constructed using N samples, but then the statistics for each function approximation are estimated using 10 000 LH samples (i.e. the kriging and MARS models are not readily amenable to exact integration as are simpler low-order polynomial approximations).

Figure 7 shows the trends in estimated mean value for an increasing number of available LH samples. Interestingly, for $N \leq 25$, there is no clear benefit to using either the kriging or the MARS function approximation. However, for $N \geq 50$, both the kriging and MARS approximations provide a more accurate mean value estimate than is obtained from the original set of N samples (i.e. comparable bias, but lower variance).

Figure 8 shows the trends in estimated standard deviation for increasing N LH samples. The 10 trials of LH samples without function approximation exhibit large scatter for low values of N, but the estimates are unbiased (i.e. are centered around the exact standard deviation value) and they converge to the correct value as N increases. In contrast, the kriging approximation method shows a clear bias in the standard deviation estimates for low values of N. Upon closer inspection of the kriging function approximation surfaces, it was noticed that for low values of N the kriging model was predicting $f < 0$ over some portions of the $[-2, 2]^2$ parameter space, whereas the true Rosenbrock function has $f \geq 0$ over the entire parameter space. (Note that this special knowledge about the Rosenbrock function was not exploited in this study since, in general, such knowledge would not be available about an unknown function.) The MARS approximation method also performs poorly for low values of N. In particular, for $N = 10$ all of the standard deviation estimates are zero. Upon further inspection, it was discovered that the MARS software generated a constant-valued function for each of the 10 trials. (Note that it is not clear from the output of the MARS code whether this is an error or simply a default condition when there are too few data points for MARS to construct a surface approximation. Clearly this is a hidden drawback of the MARS software.) The MARS-based standard deviation estimates converge to the correct value as N increases.

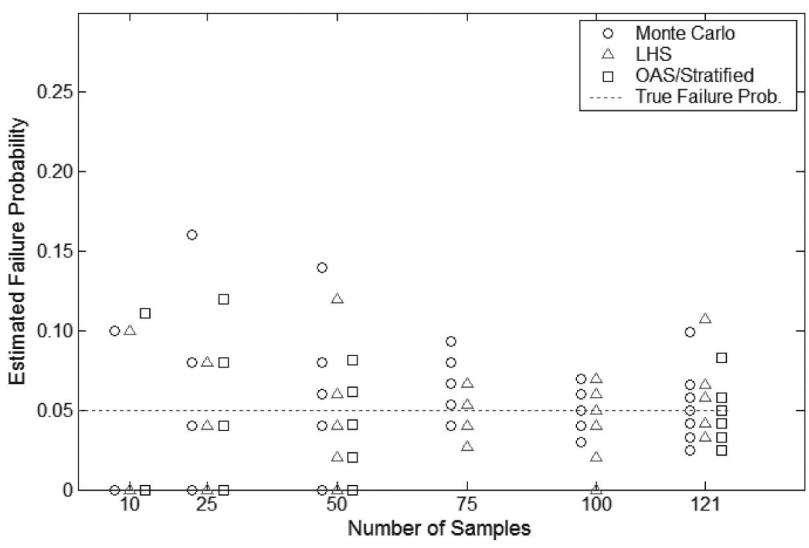

Figure 6. Comparison of failure probability estimation using MC, LH, and OA (Stratified) sampling.

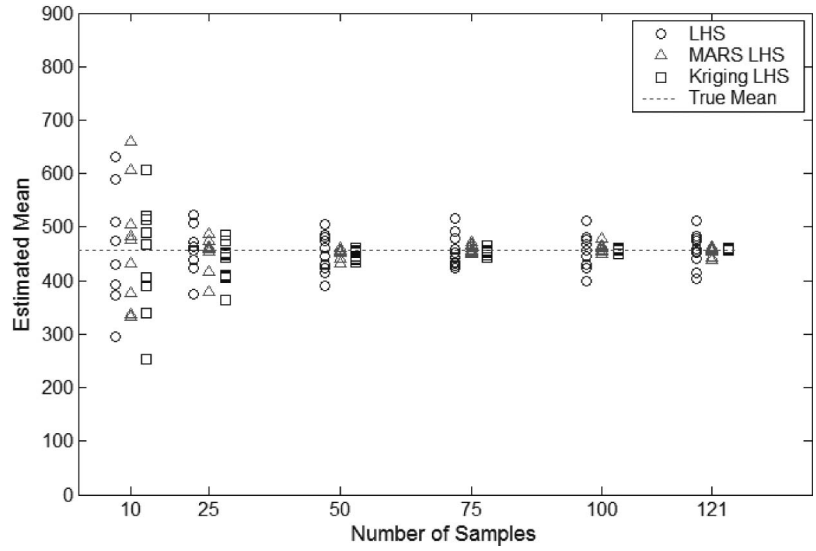

Figure 7. Comparison of mean value estimation using N LH samples vs estimation using response surface approximations generated with N LH samples.

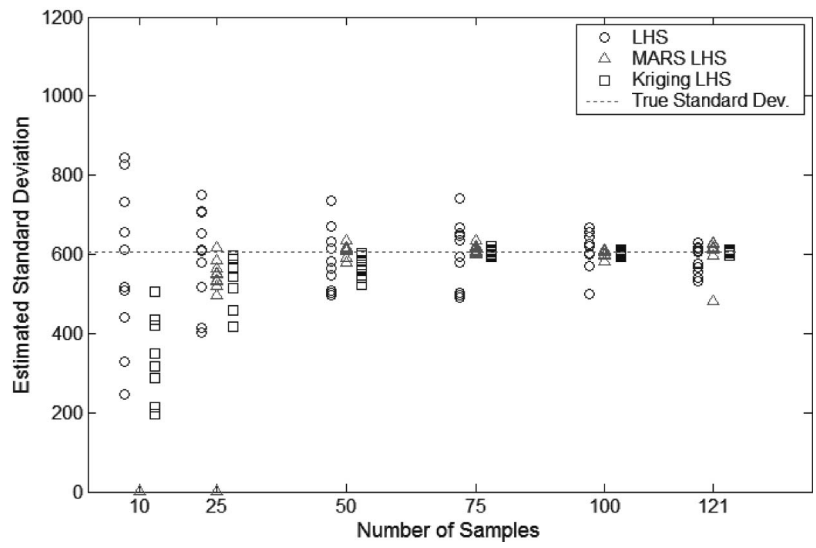

Figure 8. Comparison of standard deviation estimation using N LH samples vs estimation using response surface approximations generated with N LH samples.

Figure 9 shows the trends in estimated failure probability for increasing N LH samples. As before, the LH samples without function approximation show a large, but unbiased, scatter in failure probability estimates. This scatter decreases as N increases. The kriging-based failure probability estimates exhibit bias for low N (low bias for $N = 10$, and high bias for $N = 25$), but this bias disappears with increasing N. The MARS-based failure probability estimates are in gross error for low N, where either the failure probability is zero for all trials (e.g. for $N = 10$) or the failure probability is highly biased. As N increases, the MARS-based failure probability predictions improve, but remain highly scattered and biased even for $N = 121$. Note that the MARS-based failure probability estimates typically are worse for all values of N, when compared to the results produced by the LH samples used alone without any use of function approximation.

Figures 10 and 11 depict, qualitatively, the difference in accuracy between the MARS and kriging response surface approximations for $N = 121$ LH samples. Figure 10 shows the true Rosenbrock function (shaded contour plot) versus a sample-based view of the MARS surface approximation. That is, the MARS surface is first constructed from $N = 121$ samples, and then visualized in figure 10 by generating 10 000 samples in $[-2, 2]^2$ and plotting the predicted function value at each point. While the MARS function captures the general trends of the Rosenbrock function, there are portions of the parameter space where the MARS surface deviates from the true surface, particularly near the extreme high and low values of the surface. Hence, the mean value and standard deviation estimates are accurate for $N = 121$ samples, but the failure probability estimate is biased. While adding more samples (i.e. increasing N) would improve the accuracy of the MARS surface approximation, the trend shown in figure 9 indicates that the convergence rate is slow for failure probability estimates.

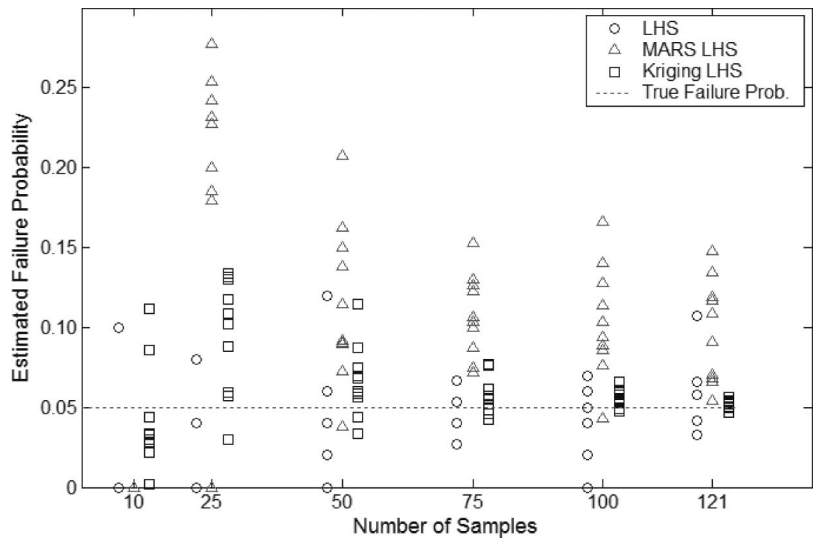

Figure 9. Comparison of failure probability estimation using N LH samples vs estimation using response surface approximations generated with N LH samples.

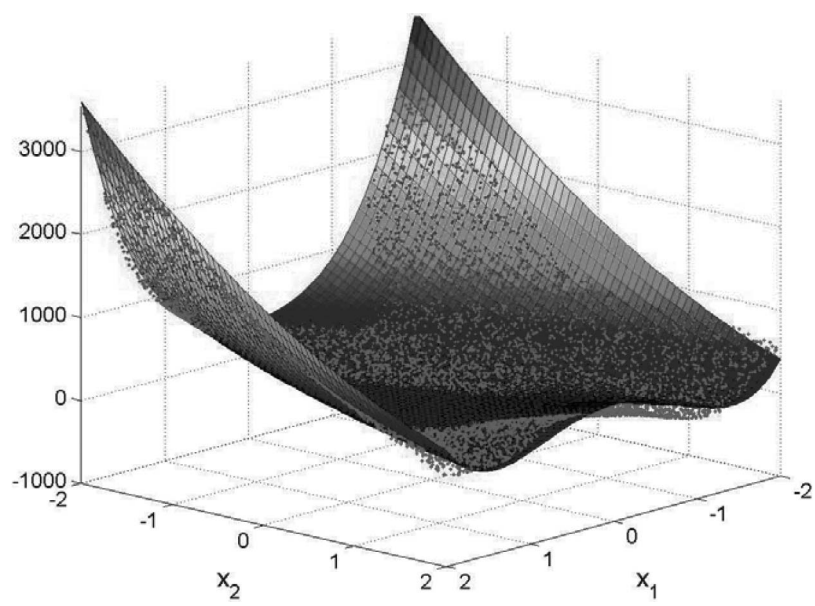

Figure 10. Rosenbrock's function and MARS approximation (dots) built with 121 samples.

Figure 11 shows a similar diagram of the kriging surface approximation for $N = 121$ samples, where again the kriging surface is visualized using 10 000 samples in $[-2, 2]^2$. Qualitatively, the kriging-based sample points appear to be much more faithful to the trends of the true Rosenbrock function, as compared to the MARS surface in figure 10. For the range of N values used here, the kriging model does not suffer from any ill-conditioning effects. However, when N is increased to 200 samples for this Rosenbrock function example, the correlation matrix in the kriging model becomes ill-conditioned, and the kriging function approximation results become erratic (large variations in predicted values at best, and code failures at worst).

4.3 *Function approximations based on OA/stratified samples*

Figures 12 – 14 illustrate the estimated mean value, standard deviation, and failure probability trends for OA/Stratified sampling, for $N = 9$, 25, 49, and 121. In general, these results for OA/Stratified sampling are similar to those for LH sampling (in figures 8 – 10). These observations are summarized as:

(a) Mean value estimates (figure 12): for low N ($N = 9$, $N = 25$), the mean value estimates are virtually the same between OA sampling alone, and OA sampling combined with kriging and MARS surface approximations. For large N ($N = 49$, $N = 121$), there is some improvement in accuracy afforded by using OA sampling with the surface approximations, versus OA sampling alone.

(b) Standard deviation estimates (figure 13): for low N, the OA-based kriging approximations of standard deviation are biased below the true value of standard deviation. This bias disappears as N increases. For $N = 9$, the OA-based MARS approximations of standard deviation are in gross error. This is due to the same constant-value MARS approximation problem observed with LH samples. For $N \geq 25$, the OA-based MARS estimates quickly converge to the true standard deviation.

(c) Failure probability estimates (figure 14): the OA-based kriging failure probability estimates converge non-monotonically to the true failure probability. For $N = 121$, the OA-based kriging estimates appear unbiased, with little scatter around the true failure probability. For $N = 9$, the OA-based MARS failure probability estimates are in gross error, and with larger N the OA-based MARS estimates slowly converge non-monotonically to the true failure probability. However, even at $N = 121$ there is clear bias and large scatter in the OA-based MARS failure probability estimates.

4.4 *Observations*

While it is difficult to draw general conclusions from this study given its limited scope, a few key observations can be made:

(a) For sampling-only estimates, the OA method tended to perform best out of the three sampling methods for a large number of samples, but worst at low numbers of samples. The LH method generally outperformed random MC sampling for all samples sizes. Also, the

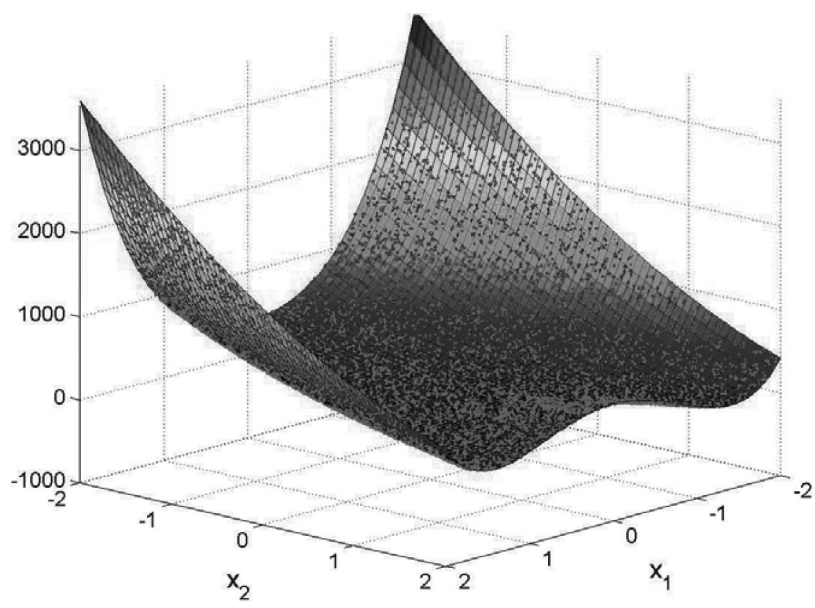

Figure 11. Rosenbrock's function and kriging approximation (dots) built with 121 samples.

LH results were close to the OA results at the higher number of samples.

(b) When the number of samples, N, is small (e.g. ~ 10 samples in a two-dimensional space), the kriging-based and MARS-based surface approximation models do not necessarily provide a benefit over using the original set of N data samples directly to estimate statistical properties (mean, standard deviation, failure probability). That is, there is just as much bias and scatter in the statistical estimates made using the response surface approximations as there is in the statistical estimates made using the original data. This is a somewhat unexpected observation, and merits further investigation. That is, we expected the kriging- and MARS-based approaches to provide more benefit for small N than was observed in this study. Perhaps it is the nonlinearity of the Rosenbrock function that is the cause of this unexpected result, and more benign test functions may be more amenable to response surface approximation.

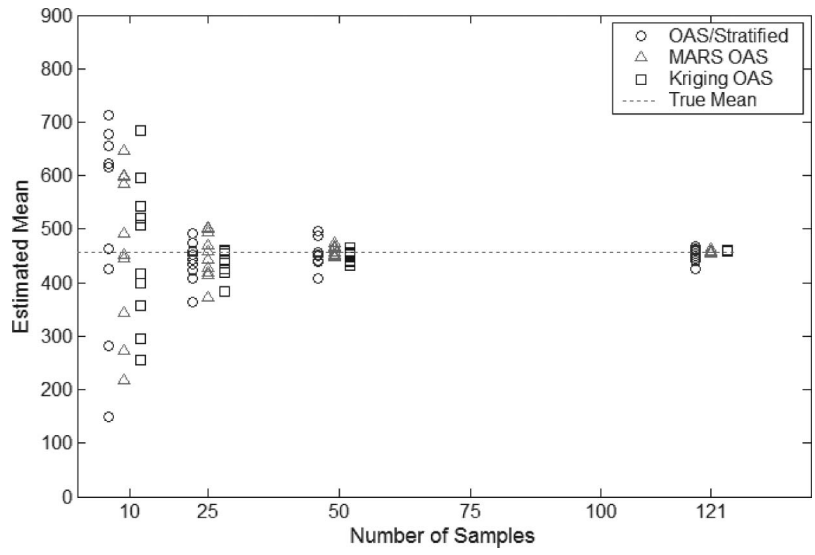

Figure 12. Comparison of mean value estimation using N OA samples vs estimation using response surface approximations generated with N OA samples.

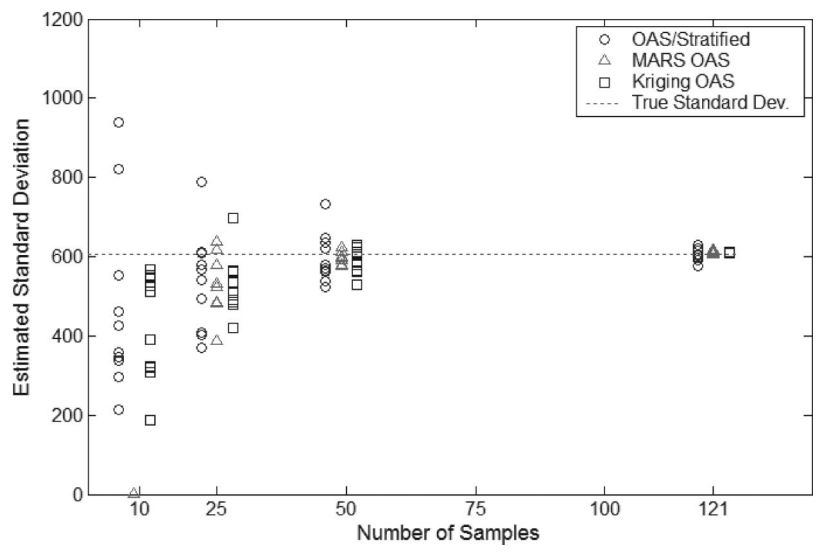

Figure 13. Comparison of standard deviation estimation using N OA samples vs estimation using response surface approximations generated with N OA samples.

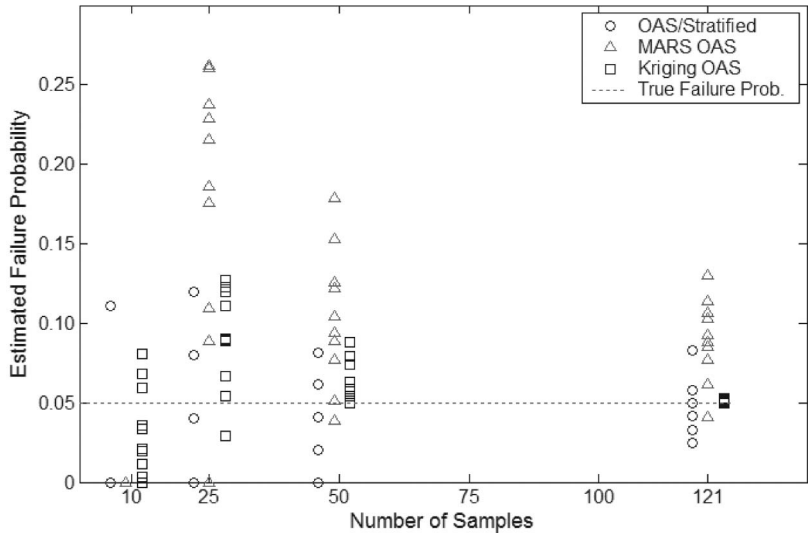

Figure 14. Comparison of failure probability estimation using N OA samples vs estimation using response surface approximations generated with N OA samples.

(c) When the number of samples is small, there is a danger of bias in the kriging-based and MARS-based statistical estimates, versus sampling alone. This bias effect was most pronounced in the standard deviation and failure probability estimates, but was not pronounced for mean value estimates.

(d) When the number of samples is large (e.g. ~ 100 samples in a two-dimensional parameter space), the kriging surface approximation method more accurately estimated the statistical properties as compared to both the MARS-based estimates and the sampling-alone estimates. This is not especially surprising, given our past experience with kriging methods (cf. Romero *et al.* 2004). However, we are fully aware of the limits of the kriging method, particularly that it is prone to mathematical ill-conditioning. From these studies, we are now aware of the potential bias in under-sampled kriging-based predictions.

(e) There is no strong evidence of interaction between the sampling schemes and the response surface approximation schemes. That is, kriging paired with OA samples did not outperform kriging paired with LH samples. The same is true for MARS/OA and MARS/LH.

(f) 'Black-box' response surface approximation methods are very useful, but they may contain hidden pitfalls. Thus, they should be used with caution and, if possible, in combination with other surface approximation methods in a sort of cross-validation capacity.

The results of this study indicate that caution should be exercised when applying response surface (RS) approximations for UQ. In particular, we suggest using a sampling-based UQ approach (i.e. without RS approximations) along with a RS approximation-based UQ approach, and we advocate using multiple types of RS approximation functions in order to cross-check the RS-based UQ predictions.

In such a combined UQ approach, if there is wide disagreement between the various UQ predictions, then it is likely that the sample size, N, is too small. Also, any large disagreements could be indicative of an inherent problem with a particular RS approximation method (see, for example, the MARS results in figures 9 and 14).

5. Summary

Real-world computer simulations often require hours or days for a single run, and the output may be highly nonlinear over the range of input parameter values. When the input parameters are uncertain, conducting an UQ study on the output using conventional sampling methods could require hundreds or thousands of simulation runs. A possible alternative to conventional UQ methods is the use of response surface approximations. When a limited number of simulation runs are available, the output can be used to build a response surface approximation model. Output statistics can then be obtained by sampling the approximation model, which is essentially free to evaluate when compared to the original computer simulation.

This study examined the accuracy of UQ statistics estimated using two sampling methods and two response surface approximation methods, kriging and MARS, based on various sample point sizes. The sample points were generated using two commonly used methods – LH

sampling and OA sampling. The two-dimensional Rosenbrock function was used to represent an expensive computer model. The performance of the sampling methods and of the response surface approximation methods was assessed based on the accuracy in estimating three output statistics on the Rosenbrock function: the mean, the standard deviation, and the probability of exceeding a 5% threshold.

Although this study is of limited scope, our preliminary observations indicate that, as expected, more accurate UQ statistics often can be estimated from the response surface approximation functions as compared to estimates made using the original data. However, this result did not hold true for all sample sizes. In particular, for very low sample sizes there were clear cases where the UQ statistics estimated from response surface approximations were less accurate, and biased, compared to estimates made using the original data. This set of conflicting results clearly motivates the need for future studies involving additional test functions. Of particular interest are high-dimensional ($n > 2$) test functions which will provide greater relevance to engineering UQ studies.

Acknowledgements

The authors wish to thank the anonymous reviewers along with Prof. Layne T. Watson of Virginia Tech and Dr James F. McEnerney of Lawrence Livermore National Laboratory for providing valuable feedback on draft versions of this manuscript. Also, the authors thank the following Sandia colleagues for their reviews of this document: Dr Joseph P. Castro, Dr Kevin J. Dowding, and Dr Vicente J. Romero.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

References

Cressie, N., *Statistics of Spatial Data*, 1991 (John Wiley and Sons: New York).

Eldred, M.S., Giunta, A.A., van Bloemen Waanders, B.G., Wojtkiewicz, S.F., Jr., Hart, W.E., and Alleva, M.P., DAKOTA users manual: version 3.1. Sandia Technical Report SAND2001-3796, 2001 (Sandia National Laboratories, Albuquerque, NM, USA) (online at: http://endo.sandia.gov/DAKOTA/software.html).

Friedman, J.H., Multivariate adaptive regression splines. *Ann. Stat.*, 1991, **19**, 1–141.

Gill, P.E., Murray, W., and Wright, M.H., *Practical Optimization*, 1981 (Academic Press: San Diego).

Giunta, A.A., Eldred, M.S., Trucano, T.G., and Wojtkiewicz, S.F., Jr., Optimization under uncertainty methods for computational shock physics applications, in *Proceedings of the 43rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, 2002, AIAA Paper 2002-1642.

Giunta, A.A. and Watson, L.T., A comparison of approximation modeling techniques: polynomial versus interpolating models, in *Proceedings of the 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, 1998, pp. 392–404, AIAA Paper 98-4758.

Gomes, H.M. and Awruch, A.M., Comparison of response surface and neural network with other methods for structural reliability analysis. *Struct. Saf.*, 2004, **26**, 49–67.

Halder, A. and Mahadevan, S., *Probability, Reliability and Statistical Methods in Engineering Design*, 2000 (John Wiley and Sons: New York).

Hedayat, A.S., Sloane, N.J.A., and Stufken, J., *Orthogonal Arrays: Theory and Applications*, 1999 (Springer: New York).

Iman, R.L. and Shortencarier, M.J., A Fortran 77 program and user's guide for the generation of Latin hypercube samples for use with computer models. Sandia Technical Report SAND83-2365, 1983 (Sandia National Laboratories, Albuquerque, NM, USA). Also listed as report number NUREG/CF-3624.

Jin, R, Chen, W., and Simpson T., Comparative studies of metamodeling techniques under multiple modeling criteria. *J. Struct. Optimiz.*, 2001, **23**, 1–13.

Kaymaz, I., Application of kriging method to structural reliability problems. *Struct. Saf.*, 2005, **27**, 133–151.

Koehler, J.R. and Owen, A.B., Computer experiments. In *Handbook of Statistics, 13*, edited by S. Ghosh and C.R. Rao, pp. 261–308, 1996 (Elsevier Science: Amsterdam).

Martinez-Canales, M. and Williams, P.G., DDACE user's guide. Available online at: http://csmr.ca.sandia.gov/projects/ddace (accessed 30 November 2005).

McKay, M.D., Beckman, R.J., and Conover, W.J., A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 1979, **21**, 239–245.

Metropolis, N. and Ulam, S., The Monte Carlo method. *J. Am. Stat. Assoc.*, 1949, **44**, 335–341.

Myers, R.H. and Montgomery, D.C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 1995 (John Wiley and Sons: New York).

Owen, A.B., Orthogonal arrays for computer experiments, integration, and visualization. *Statistica Sinca*, 1992, **2**, 439–452.

Romero, V.J., Swiler, L.P., and Giunta, A.A., Construction of response surfaces based on progressive-lattice-sampling experimental designs. *Struct. Saf.*, 2004, **26**, 201–219.

Simpson, T.W., Mauery, T.M., Korte, J.J., and Mistree, F., Kriging metamodels for global approximation in simulation-based multidisciplinary design optimization. *AIAA J.*, 2001, **39**, 2233–2241.

Swiler, L.P. and Wyss, G.D., A user's guide to Sandia's Latin hypercube sampling software: LHS UNIX library/standalone version. Sandia Technical Report SAND2004-2439, 2004 (Sandia National Laboratories, Albuquerque, NM, USA).

Appendix A

Listed below is the DAKOTA input file that performs a sampling-based uncertainty quantification study using the Rosenbrock function.

```
# DAKOTA INPUT FILE - dak_uq.in

strategy,                                          \
      single_method                                \
      tabular_graphics_data                        \
#           graphics

method,                                            \
      nond_sampling   # use for MC and LH samples  \
      samples = 121                                \
#           seed = 5                               \
#           sample_type random                     \
      sample_type lhs                              \
      response_levels = 4.0                        \
#     dace oas   # use for orthogonal arrays       \
#           samples = 121                          \
#           all_variables

variables,                                         \
      uniform_uncertain =  2                       \
      uuv_dist_lower_bounds =       -2.0  -2.0     \
      uuv_dist_upper_bounds =        2.0   2.0     \
      uuv_descriptors =             'x1'  'x2'     \
      uncertain_correlation_matrix = 1.0   0.0     \
                                     0.0   1.0

interface,                                         \
      application direct                           \
      analysis_driver = 'rosenbrock'

responses,                                         \
      num_response_functions = 1                   \
      no_gradients                                 \
   no_hessians
```

Appendix B

Listed below is the DAKOTA input file that performs a combined sampling-based and response surface approximation-based uncertainty quantification study using the Rosenbrock function.

```
# DAKOTA INPUT FILE - dakota_sbuq.in

strategy,                                          \
        single_method                              \
          tabular_graphics_data                    \
          method_pointer = 'UQ'

method,                                            \
        id_method = 'UQ'                           \
        model_type layered                         \
          interface_pointer = 'RSM'                \
        nond_sampling                              \
          samples = 10000                          \
          seed = 5                                 \
          sample_type lhs                          \
          response_levels = 4.0

variables,                                         \
        uniform_uncertain    =  2                  \
          uuv_dist_lower_bounds =  -2.0  -2.0      \
          uuv_dist_upper_bounds =   2.0  2.0       \
          uuv_descriptors       =   'x1' 'x2'      \
        uncertain_correlation_matrix =  1.0 0.0    \
                                        0.0 1.0

interface,                                         \
        id_interface = 'RSM'                       \
        approximation global,                      \
          dace_method_pointer = 'DACE'             \
          reuse_samples all                        \
#         kriging                                  \
#           correlations = 1.0 1.0                 \
          mars

responses,                                         \
        num_response_functions = 1                 \
        no_gradients                               \
        no_hessians

#######################################
# interface truth model and dace method #
#######################################

method,                                            \
        id_method = 'DACE'                         \
        model_type single                          \
          interface_pointer = 'I1'                 \
        nond_sampling                              \
          samples = 121                            \
#         seed = 50                                \
          sample_type lhs                          \
#       dace oas                                   \
#         samples = 121                            \
#         all_variables

interface,                                         \
        application direct,                        \
          id_interface = 'I1'                      \
          analysis_driver = 'rosenbrock'
```