13th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP13
Seoul, South Korea, May 26-30, 2019

Surrogate modeling and model selection in irreducible high dimensions with small sample size

Anindya Bhaduri
Graduate Student, Dept. of Civil Engineering, Johns Hopkins University, Baltimore, United States of America

Lori Graham-Brady
Professor, Dept. of Civil Engineering, Johns Hopkins University, Baltimore, United States of America

Michael D. Shields
Assistant Professor, Dept. of Civil Engineering, Johns Hopkins University, Baltimore, United States of America

ABSTRACT: There exist a number of high dimensional problems in which the dimensions cannot be effectively reduced, since all of them are more or less equally important. On top of that, when the computational models are expensive, it is not practical to perform more than a small number of model evaluations. In situations like this, a good space filling design is needed that provides maximum coverage over the input domain. In surrogate modeling methods, like kriging interpolation or radial basis function interpolation, a good sampling design can help improve the condition number of the kernel matrix by placing samples as far apart from each other as possible. In this study, the performance of three hierarchical space filled designs, namely Refined Latinized Stratified Sampling (RLSS), Hierarchical Latin Hypercube Sampling (HLHS) and Sobol quasi-random sequence, are compared using the Rosenbrock function in different dimensions. Ordinary kriging interpolation is chosen as the surrogate modeling method with different choices of correlation functions. The AIC criterion is used for model selection and the accuracy of selection is cross-verified using the root mean squared (RMS) error values.

Expensive computational models are often used to replicate many complex physical systems existing in nature. The models are more often than not associated with a number of parameters, ranging from as few as 2 or 3 to as high as few thousands. The input parameters can be random in nature and can follow a distribution structure. It might be of interest to account for the variations in the quantity of interest with respect to the input parameters, each of which varies over a certain range. It can become cumbersome to achieve the above goals when dealing with computationally expensive models. An efficient approach is to perform non-intrusive surrogate modeling. A non-intrusive surrogate modeling approach deals with establishing a cheap mathematical input/output relationship using a limited number of the expensive model evaluations. There are surrogate models which use structured sampling designs [Babuška et al. (2007); Agarwal and Aluru (2009); Bhaduri and Graham-Brady (2018); Bhaduri et al. (2018)], while others use unstructured designs [Santner et al. (2013); Zhang et al. (2013); Shields (2018)]. Choice of sampling points is an important step of any surrogate modeling procedure. When the broad choice is that of an unstructured sample design, a favourable design is

one with good non-collapsive space filling properties. Popular designs include Sobol quasi random sequences [Sobol' (1976)], Halton quasi random sequences [Halton (1960)], and Latin Hypercube Sampling (LHS) [McKay et al. (1979); Helton and Davis (2003)], among others.

In this work, two hierarchical non-collapsive space filling designs, hierarchical latin hypercube sampling (HLHS) [Sallaberry et al. (2008); Vořechovskỳ (2015)] and refined latinized stratified sampling (RLSS) [Shields (2016)] are considered along with the popular quasi-random Sobol sequence, and their performance is assessed when used as input designs for surrogate modeling. In this study, the surrogate modeling method considered is kriging. The paper is divided into the following sections. The HLHS and RLSS designs are described in brief in section 1. Section 2 talks about the kriging interpolation method in brief. In section 3, we demonstrate the performance of the RLSS, HLHS and Sobol designs in kriging metamodeling of the Rosenbrock function of different dimensions. Section 4 concludes the paper with discussions.

1. HLHS AND RLSS DESIGNS

LHS design [McKay et al. (1979)] has good projective properties but does not guarantee good space filling properties. Stratified sampling (SS) [McKay et al. (1979)] design, on the other hand, has good space filling properties but poor projective properties. Refined stratified sampling (RSS) design [Shields et al. (2015)] is the sequential version of the SS design where new strata are formed one at a time and a new sample is added by randomly sampling from each newly formed stratum. HLHS designs [Sallaberry et al. (2008); Vořechovskỳ (2015)] help in sequential addition of LHS design samples by gradual refinements of the dimension-wise stratifications. The minimum refinement that can be performed is to subdivide each dimension-wise stratum into two sub-strata. Thus, HLHS is not a purely sequential design and at least doubles the current sample size at each extension. Latinized stratified sampling (LSS) [Shields and Zhang (2016)] design attempts to achieve good non-collapsive and space filling properties by combining the advantages of the LHS and SS designs. In LSS, there is dimension-wise stratifications (as in LHS design) as well as full-dimensional stratifications (as in SS design) and sampling is performed taking both stratifications into account simultaneously. The sample size extension version of LSS design is referred to as Hierarchical Latinized Stratified Sampling (HLSS) design [Shields (2016)]. HLSS design involves the LHS-type refinements and subdivision of the SS-type strata in an exisiting LSS design and then sampling in the permissible regions. Finally, RLSS is essentially a sequential version of HLSS where one stratum is chosen from the strata subdivision and a sample is added there.

2. KRIGING INTERPOLATION

Kriging [Krige (1951); Matheron (1963)], also known as Gaussian process modelling, is an interpolation algorithm which tries to build a metamodel $\hat{y}(x)$ corresponding to an unknown true function $y(x)$ by assuming it to be a realization of a Gaussian process $Y(x)$. $Y(x)$ is given by:

$$Y(x) = \boldsymbol{\beta}^T \boldsymbol{f}(\mathbf{x}) + \sigma^2 Z(\mathbf{x}, \omega) \tag{1}$$

where term $\boldsymbol{\beta}^T \mathbf{f}(\mathbf{x})$ represents the mean value of the Gaussian process and $\boldsymbol{\beta}$ is the regression vector and $\mathbf{f}(\mathbf{x})$ is the basis function vector. The term $\sigma^2 Z(\mathbf{x}, \omega)$ represent the local variations of the function about the mean $\boldsymbol{\beta}^T \mathbf{f}(\mathbf{x})$ where σ^2 is the process variance and $Z(x, \omega)$ is a stationary Gaussian process with zero mean $[\mathbb{E}(Z(x)) = 0]$ and a correlation function given by:

$$Cov[Z(x), Z(x')] = K(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) \tag{2}$$

where $K = K(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta})$ is a measure of the similarity between two samples of the input space, e.g. x and x' and depends on the hyperparameters $\boldsymbol{\theta}$. We consider ordinary kriging in our study where $\boldsymbol{f}(\boldsymbol{x}) = 1$ and $\boldsymbol{\beta}^T \boldsymbol{f}(\boldsymbol{x}) = \beta_0$. Then the expression of the mean estimate of the kriging predictor is given by:

$$\mu_{\hat{Y}}(\mathbf{x}) = \hat{\beta}_0 + r(x)^T R^{-1} (\boldsymbol{y} - \mathbf{1}\hat{\beta}_0) \tag{3}$$

and the prediction variance estimate is given by:

$$s^2_{\hat{Y}}(x) = \hat{\sigma}^2 (1 - r^T(x) R^{-1} r(x)) \tag{4}$$

where,

$$\hat{\beta}_0 = (\mathbf{1}^T R^{-1}\mathbf{1})^{-1}\mathbf{1}^T R^{-1} y$$

$$\hat{\sigma}^2 = \frac{1}{n}(y-\mathbf{1}\hat{\beta})^T R^{-1}(y-\mathbf{1}\hat{\beta}) \quad (5)$$

In kriging prediction, the choice of covariance function K is important. In this study, four families of ellipsoidal type covariance functions are used: Gaussian, Exponential, Matérn 3/2 and Matérn 5/2, whose expressions are given in Table 1. The Kriging module in UQlab [Marelli and Sudret (2014); Lataniotis et al. (2015)] was used to perform Kriging surrogate modeling.

Table 1: Stationary covariance functions

Covariance functions	Expression
Gaussian	$exp\big(-\sum_{i=1}^{d}\theta_i p_i^2\big)$
Exponential	$exp\big(-\sum_{i=1}^{d}\theta_i p_i\big)$
Matern 3/2	$(1+\sqrt{3}\sum_{i=1}^{d}\theta_i p_i)exp\big(-\sum_{i=1}^{d}\theta_i p_i\big)$
Matern 5/2	$(1+\sqrt{5}\sum_{i=1}^{d}\theta_i p_i + ... \frac{5}{3}\sum_{i=1}^{d}\theta_i^2 p_i^2)exp\big(-\sum_{i=1}^{d}\theta_i p_i\big)$

Note: $p_i = |x_i - x_i^{'}|$; for isotropic case, $\theta_i = \theta$

3. NUMERICAL RESULTS

In this section, a commonly used benchmark function for high-dimensional applications, the Rosenbrock function, is used as an example problem given by:

$$f(x) = \sum_{i=1}^{d-1} 100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2 \quad (6)$$

where $\mathbf{x}$ is a d-dimensional vector. 4 Rosenbrock functions of dimensions 2, 5, 10 and 20 are considered. Sample points are generated according to three designs: Sobol, HLHS and RLSS. For a given dimension case and a given design, the function was evaluated at the generated sample points and kriging was applied to the data set to predict the function values at 10^6 Monte Carlo (MC) test samples in the problem domain. The performance of each case was measured using Root Mean Squared Error (RMSE) given by:

$$RMSE = \sqrt{\frac{1}{N_t}\sum_{i=1}^{N_t}\left(y_{true}^{(i)} - y_{predicted}^{(i)}\right)^2} \quad (7)$$

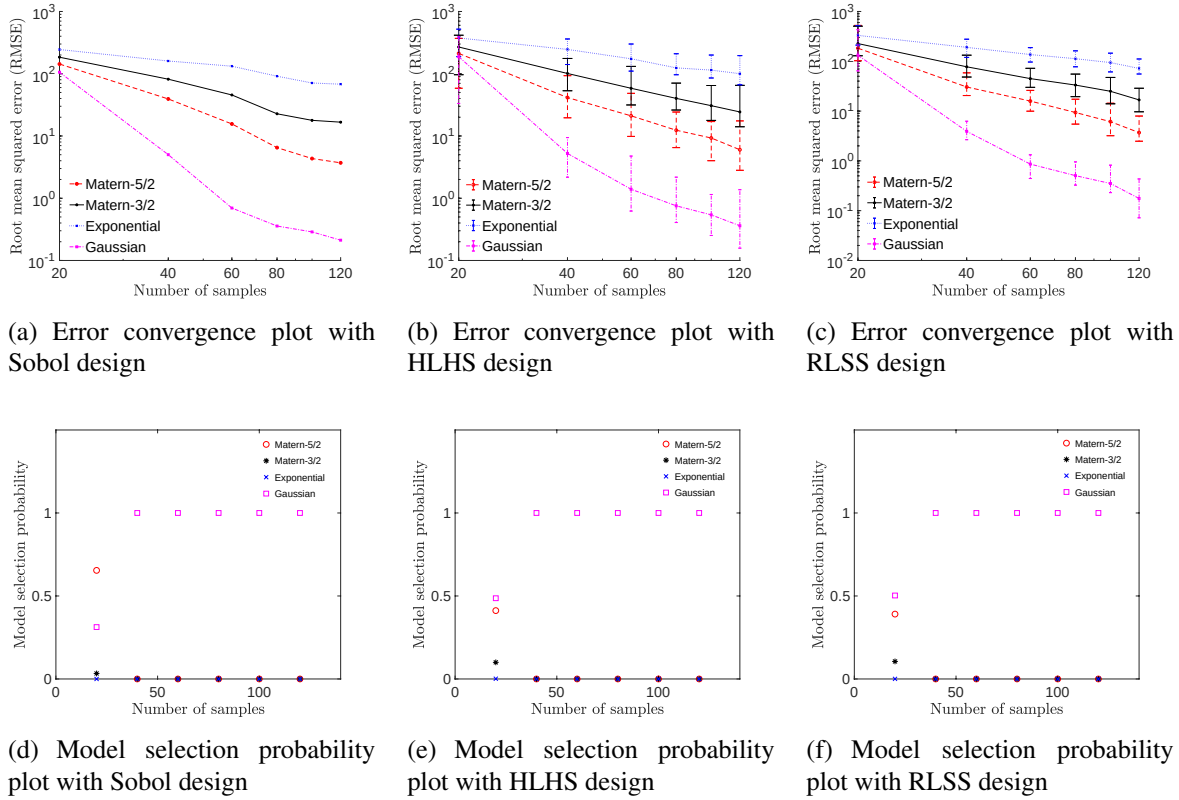
where N_t is the total number of test samples, $\mathbf{y}_{true}$ is the vector of true values of the function at the N_t points and $\mathbf{y}_{predicted}$ is the vector of kriging-predicted values at the same N_t points. Here, $N_t = 10^6$. The results from the different designs are combined together to compare their performance with different choice of kriging covariance functions. The covariance functions are considered to be isotropic. Maximum Likelihood (ML) was chosen as the estimation method [Santner et al. (2013)] and an interior point gradient-based optimization method with L-BFGS Hessian approximation [Nocedal (1980); Byrd et al. (1999)] was used to obtain the optimized parameter $\hat{\theta}$. The Akaike information criterion (AIC) [Akaike (1974); Burnham and Anderson (2003)] was used for model selection from candidate covariance functions and the best models from each design were compared with each other. AIC for small datasets [Cavanaugh et al. (1997)] is given by:

$$AIC_c = -2log(L(\hat{\theta}|D,M)) + 2n + \frac{2n(n+1)}{N-n-1} \quad (8)$$

where L is the likelihood function, $\hat{\theta}$ is the maximum likelihood estimate, D is the data, M is the model, n denotes the number of model parameters and N denotes the sample size. The model selection probability is then given by:

$$p_i = p(M_i|D) = \frac{exp(-\frac{\triangle_A^{(i)}}{2})}{\sum_{i=1}^{K} exp(-\frac{\triangle_A^{(i)}}{2})} \quad (9)$$

where, $\triangle_A^{(i)} = AIC_c^{(i)} - AIC_c{}^{min}$. In this study, $n = 1$ since isotropic covariance functions are considered as candidate models, and the model performance is compared at a fixed N. Thus, essentially, from the AIC criterion, the most suitable model is the one with the highest maximum likehood function value.



(a) Error convergence plot with Sobol design

(b) Error convergence plot with HLHS design

(c) Error convergence plot with RLSS design

(d) Model selection probability plot with Sobol design

(e) Model selection probability plot with HLHS design

(f) Model selection probability plot with RLSS design

Figure 1: 2-dimensional Rosenbrock function

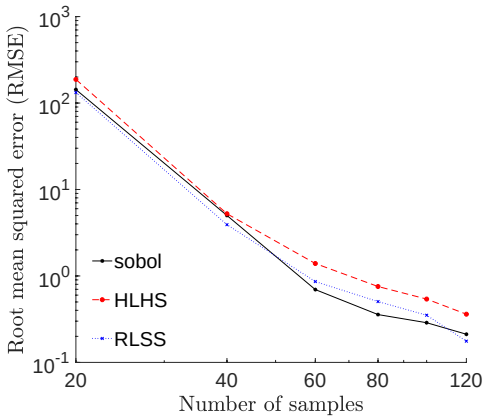
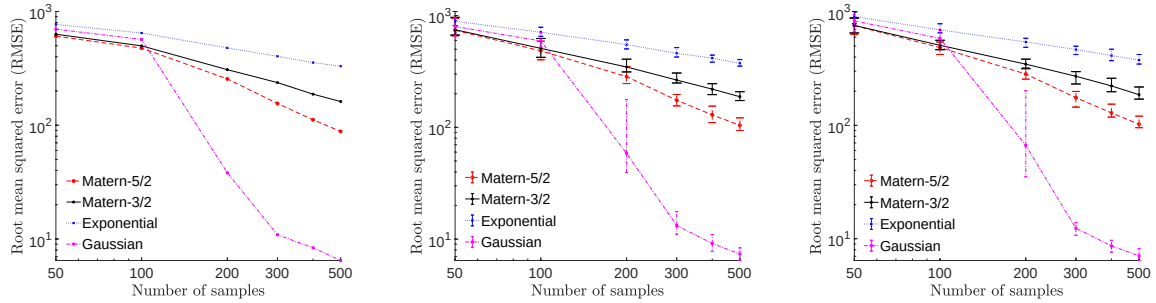


Figure 2: 2-dimensional Rosenbrock function

For HLHS and RLSS designs, 30 independent set of samples are generated. For each set, the meta-modeling procedure is conducted and the RMSE value is calculated. From these values, an error-bar plot is generated using the minimum, maximum and median RMSE values. On the other hand, Sobol sequence is deterministic and generation of only a single set of samples is sufficient for performance comparisons.

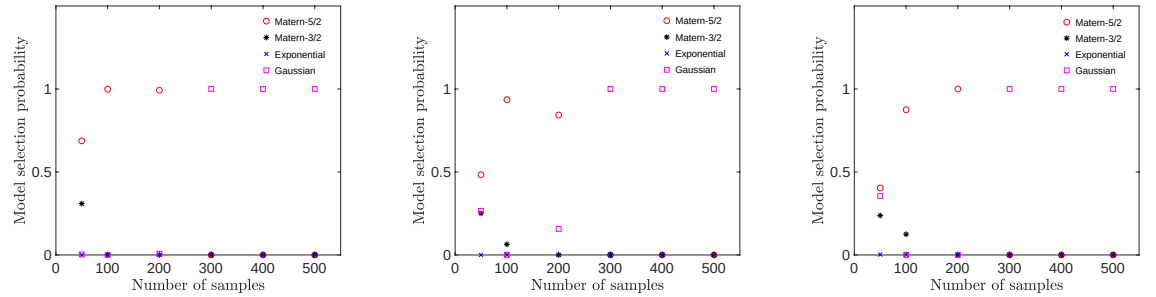
Figure 1(a) shows the RMS error convergence plot comparison between different types of covariance function models with Sobol design points for 2-dimensional Rosenbrock function, and the surrogate model with the Gaussian covariance function is found to be most accurate for all the sample cases. Figure 1(d) shows the probabilities of model selection with Sobol design. Except the 20-sample case, the AIC criterion chooses the Gaussian model as the most appropriate model which matches with the true RMSE error estimates in Figure 1(a). For the 20-sample case, the Matern-5/2 model is the most probable model but it has a higher RMS error than the Gaussian model, although the Gaussian model also has a finite probabiity of being selected. Figures 1(b) and 1(c) show the RMS error convergence plot comparisons between different types of covariance function models with HLHS design points and RLSS design points, respectively, where the Gaussian model again has the minimum error. As shown



(a) Error convergence plot with Sobol design

(b) Error convergence plot with HLHS design

(c) Error convergence plot with RLSS design



(d) Model selection probability plot with Sobol design

(e) Model selection probability plot with HLHS design

(f) Model selection probability plot with RLSS design

Figure 3: 5-dimensional Rosenbrock function

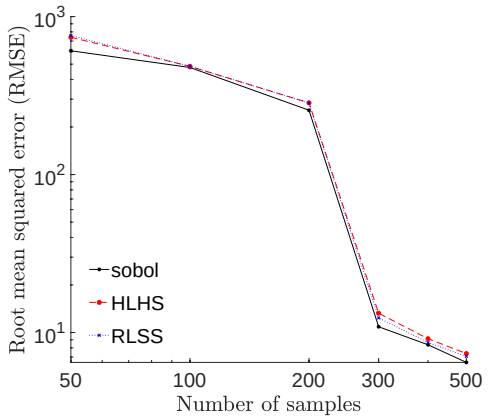
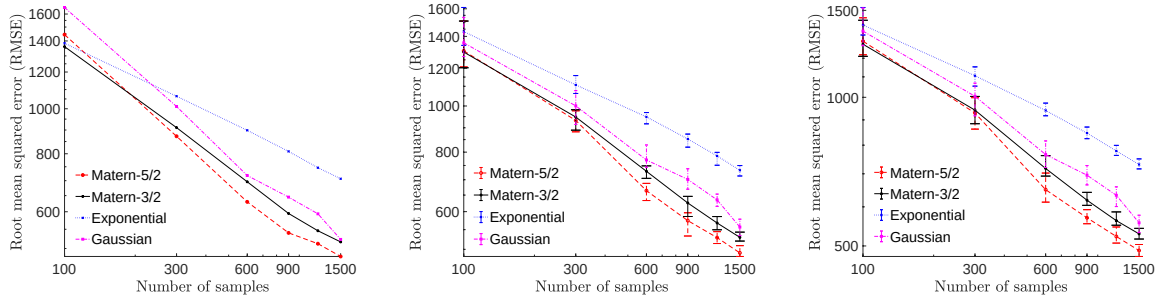


Figure 4: 5-dimensional Rosenbrock function

in figures 1(e) and 1(f), the AIC criterion selects the Gaussian model as the most suitable model which is in agreement with the RMS error values. Figure 2 shows the RMS error values of the most probable models selected from the AIC criterion using each of the 3 different designs for 2-dimensional Rosenbrock function.

Figures 3(a), 3(b) and 3(c) shows the RMS error convergence plot comparison between different types of covariance function models with Sobol, HLHS and RLSS design points respectively for 5-dimensional Rosenbrock function. The corresponding model selection probability plots in figures 3(d), 3(e) and 3(f) behave similarly in the sense that the selected (most probable) models are the ones with the minimum RMS error among the candidates. The only exception is the 200-sample case where the Matern-5/2 model is the selected model for all the 3 designs even though the RMS error is lowest for the Gaussian model for each design case. Figure 4 shows the RMS error convergence plot comparison between the best models from the AIC criterion using the 3 different designs for 5-dimensional Rosenbrock function.

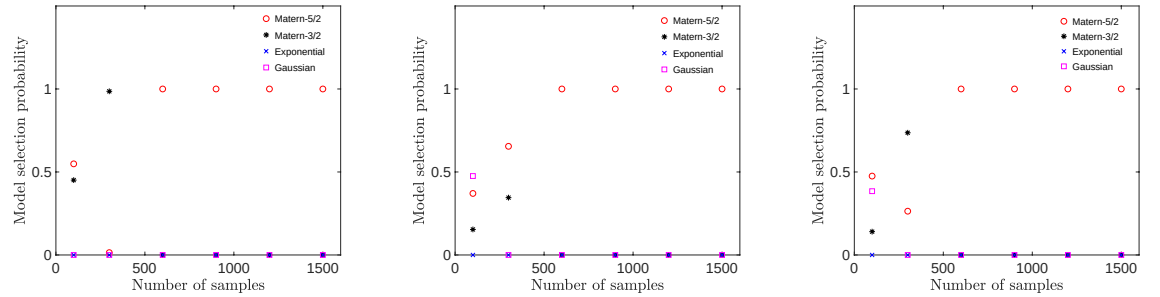
Figures 5(a), 5(b) and 5(c) shows the RMS error convergence plot comparison between different types of covariance function models with the 3 designs for 10-dimensional Rosenbrock function. In figures 5(d), 5(e) and 5(f), it is seen that there is



(a) Error convergence plot with Sobol design

(b) Error convergence plot with HLHS design

(c) Error convergence plot with RLSS design



(d) Model selection probability plot with Sobol design

(e) Model selection probability plot with HLHS design

(f) Model selection probability plot with RLSS design

Figure 5: 10-dimensional Rosenbrock function

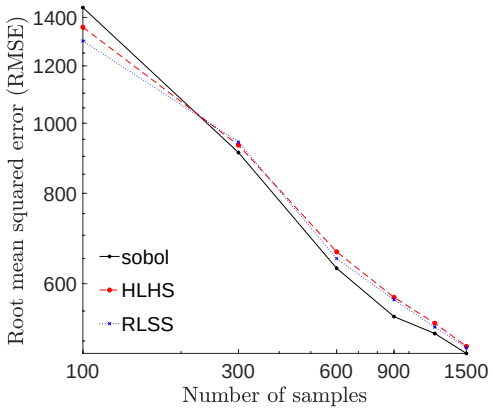


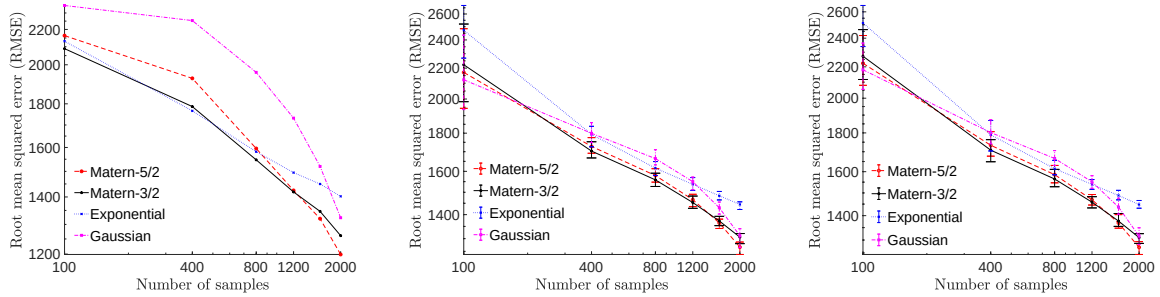
Figure 6: 10-dimensional Rosenbrock function

some discrepancy in the model selection and the corresponding RMS error for 100 and 300 sample cases, but the difference in error between the selected model and the most accurate (minimum RMS error) model is very small. However, there is good agreement between the maximum likelihood function value and the RMS error for the higher sample cases. Figure 6 shows the RMS error convergence plot comparison between the best models from the AIC criterion using the 3 different designs for 10-dimensional Rosenbrock function.

Figure 7 shows the RMS error convergence plot comparison between different types of covariance function models with the 3 designs and the corresponding model selection probabilities for 20-dimensional Rosenbrock function. Apart from the 100-sample case with the Sobol design, there is a general agreement between the maximum likelihood function value and the RMS error and in cases of disagreement, the difference in the RMS error is fairly small. Figure 8 shows the RMS error convergence plot comparison between the best models from the AIC criterion using the 3 different designs for 20-dimensional Rosenbrock function.

4. CONCLUSIONS

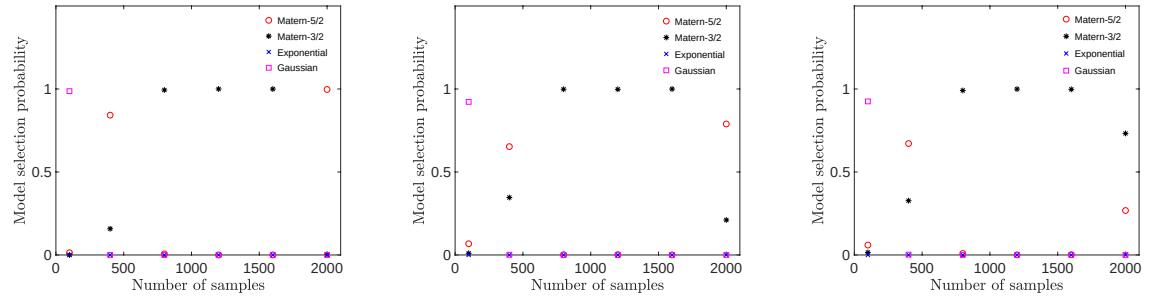
From the results, we can conclude that the performance of all three designs in most of the cases are very close to each other and none of them have a



(a) Error convergence plot with Sobol design

(b) Error convergence plot with HLHS design

(c) Error convergence plot with RLSS design



(d) Model selection probability plot with Sobol design

(e) Model selection probability plot with HLHS design

(f) Model selection probability plot with RLSS design

Figure 7: 20-dimensional Rosenbrock function

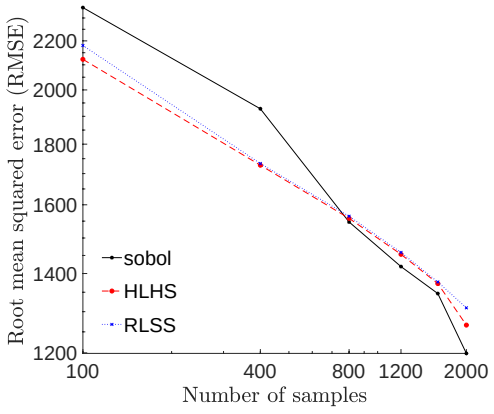


Figure 8: 20-dimensional Rosenbrock function

distinct advantage over the other. The only notable difference in performance is in the small sample 20-dimensional case where the selected model for the Sobol design has a significantly higher RMS error than the selected models with the other two designs. In general, for the 2 and 5-dimensional case, the Gaussian models seems to perform the best but its performance degrades with increase in dimensions while Matern-5/2 and Matern-3/2 models shows better performance in higher dimensions. It is to be noted that these observations correspond to the Rosenbrock function and might change for any other arbitrary function.

5. REFERENCES

Agarwal, N. and Aluru, N. R. (2009). "A domain adaptive stochastic collocation approach for analysis of mems under uncertainties." *Journal of Computational Physics*, 228(20), 7662–7688.

Akaike, H. (1974). "A new look at the statistical model identification." *IEEE transactions on automatic control*, 19(6), 716–723.

Babuška, I., Nobile, F., and Tempone, R. (2007). "A stochastic collocation method for elliptic partial differential equations with random input data." *SIAM Journal on Numerical Analysis*, 45(3), 1005–1034.

Bhaduri, A. and Graham-Brady, L. (2018). "An efficient adaptive sparse grid collocation method through

derivative estimation." *Probabilistic Engineering Mechanics*, 51, 11–22.

Bhaduri, A., He, Y., Shields, M. D., Graham-Brady, L., and Kirby, R. M. (2018). "Stochastic collocation approach with adaptive mesh refinement for parametric uncertainty analysis." *Journal of Computational Physics*.

Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Byrd, R. H., Hribar, M. E., and Nocedal, J. (1999). "An interior point algorithm for large-scale nonlinear programming." *SIAM Journal on Optimization*, 9(4), 877–900.

Cavanaugh, J. E. et al. (1997). "Unifying the derivations for the akaike and corrected akaike information criteria." *Statistics & Probability Letters*, 33(2), 201–208.

Halton, J. H. (1960). "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals." *Numerische Mathematik*, 2(1), 84–90.

Helton, J. C. and Davis, F. J. (2003). "Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems." *Reliability Engineering & System Safety*, 81(1), 23–69.

Krige, D. G. (1951). "A statistical approach to some basic mine valuation problems on the witwatersrand." *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6), 119–139.

Lataniotis, C., Marelli, S., and Sudret, B. (2015). "Uqlab user manual–kriging (gaussian process modelling)." *Report UQLab-V0*, 9–105.

Marelli, S. and Sudret, B. (2014). "Uqlab: A framework for uncertainty quantification in matlab." *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*, 2554–2563.

Matheron, G. (1963). "Principles of geostatistics." *Economic geology*, 58(8), 1246–1266.

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979). "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code." *Technometrics*, 21(2), 239–245.

Nocedal, J. (1980). "Updating quasi-newton matrices with limited storage." *Mathematics of computation*, 35(151), 773–782.

Sallaberry, C. J., Helton, J. C., and Hora, S. C. (2008). "Extension of latin hypercube samples with correlated variables." *Reliability Engineering & System Safety*, 93(7), 1047–1059.

Santner, T. J., Williams, B. J., and Notz, W. I. (2013). *The design and analysis of computer experiments*. Springer Science & Business Media.

Shields, M. D. (2016). "Refined latinized stratified sampling: A robust sequential sample size extension methodology for high-dimensional latin hypercube and stratified designs." *International Journal for Uncertainty Quantification*, 6(1).

Shields, M. D. (2018). "Adaptive monte carlo analysis for strongly nonlinear stochastic systems." *Reliability Engineering & System Safety*, 175, 207–224.

Shields, M. D., Teferra, K., Hapij, A., and Daddazio, R. P. (2015). "Refined stratified sampling for efficient monte carlo based uncertainty quantification." *Reliability Engineering & System Safety*, 142, 310–325.

Shields, M. D. and Zhang, J. (2016). "The generalization of latin hypercube sampling." *Reliability Engineering & System Safety*, 148, 96–108.

Soboĺ, I. (1976). "Uniformly distributed sequences with additional uniformity properties." *USSR Comput. Math. and Math. Phy*, 16, 236–242.

Vořechovskỳ, M. (2015). "Hierarchical refinement of latin hypercube samples." *Computer-Aided Civil and Infrastructure Engineering*, 30(5), 394–411.

Zhang, J., Chowdhury, S., Zhang, J., Messac, A., and Castillo, L. (2013). "Adaptive hybrid surrogate modeling for complex systems." *AIAA journal*, 51(3), 643–656.