46th AIAA Aerospace Sciences Meeting and Exhibit
7 - 10 January 2008, Reno, Nevada

AIAA 2008-901

Error Measures for Noise-free Surrogate Approximations

Tushar Goel[1], Raphael T Haftka[2*], Wei Shyy[3†]
[1]*Livermore Software Technology Corporation, Livermore CA, 94551*
[2]*University of Florida, Gainesville FL, US 32611*
[3]*University of Michigan, Ann Arbor CA, 48109*

Keywords: Error measures, Surrogate models, Ensemble strategy, PRESS, Prediction variance.

Abstract

Error measures are important for assessing uncertainty in surrogate predictions. We use a suite of test problems to appraise a few error estimations for polynomial response surface and kriging. In addition, we study the performance of cross-validation error measures that can be used with any surrogate. We use a large number of experimental designs to obtain the variability of error estimates with respect to experimental designs for each problem. We find that the (actual) errors for polynomial response surfaces are less sensitive to the choice of experimental designs than kriging errors. This is attributed to the variability in the maximum likelihood estimates of the kriging parameters. We find that no single error measure outperforms other measures on all test problems. Computationally expensive integrated local error measures (standard error for polynomials and mean square error for kriging) estimate the actual root mean square error very well. The distribution-free cross-validation error characterized actual errors reasonably well. While estimated root mean square error for polynomial response surface is a good estimate of the actual errors, process variance for kriging is not. We explore methods of simultaneously using multiple error measures and demonstrate that the geometric means of combinations of multiple error measures improve the assessment of the actual errors compared to the individual error measures.

I. Introduction

Surrogate based optimization is very attractive for computationally expensive problems. We construct surrogate models to evaluate the performance of systems using a limited amount of data and apply the optimization to the surrogates. Different error measures are used to assess the accuracy of surrogates (Queipo et al. [1] and the reference within). These measures are also used for determining sampling locations in many surrogate-based optimization methods like EGO ([2], [3]) and for adaptive sampling. The success of these optimization methods depends on the accuracy of the error estimation.

Error measures can be broadly classified as parametric (model based) or distribution-free (model independent). Parametric error measures are typically based on some statistical assumptions. For example, prediction variance for polynomial response surface approximation is developed assuming that the data contains normally distributed noise with zero mean and same variance σ^2 in all data and no correlation. When these statistical assumptions are not satisfied, as is usually the case, the accuracy of the resulting error estimates is questionable. On the other hand, model independent error measures are not based on any statistical assumption. However, they may have higher computational cost.

Error measures can further be characterized as global or local error estimates. Global error measures provide a single number for the entire design space, for example, process variance (parametric) for kriging or predicted residual sum of squares (model independent) estimate for polynomial response surface approximation. Local or pointwise error measures estimate error at some point in the entire design space. Examples of the parametric pointwise error measures are standard error for polynomial response surfaces, and mean square errors for kriging. Recently, Goel et al. [4] proposed using standard deviation of responses from multiple surrogates as a model independent pointwise error measure. While global error measures are practical for assessing the overall quality of surrogates, local error measures assess the quality in different regions.

While there have been some earlier efforts in appraising the efficiency of leave-one-out global error estimation measures by, e.g., Meckesheimer et al. [5], there is a lack of systematic study to assess the accuracy of different error measures. The primary goal of this paper is to appraise the global performance of different error measures with the help of a variety of test problems. Specifically, we focus on estimated variance of noise, estimated standard error for polynomial response surfaces, process variance and mean squared error for kriging, and predicted residual sum of squares (PRESS) for both surrogates. Among these error measures, estimated standard error and mean squared error are local error measures and all other measures are global error measures. We chose these error measures because of two reasons. Firstly, polynomial response surface and kriging are the

[*] Distinguished Professor, Fellow AIAA
[†] Clarence L "Kelly" Johnson Professor and Head, Fellow AIAA

most popular surrogates in industry. Secondly, mostly practitioners use these error measures to assess the quality of the surrogates. We account for the influence of experimental designs on error estimates by considering a large number of experimental designs obtained using a combination of Latin hypercube sampling and D-optimality criterion. Noting the difficulties in identifying regions of high uncertainty using a single error estimation measure, we explore the idea of simultaneously using multiple error measures that is motivated from our previous work on ensemble of surrogates [4]. Here, we examine the idea of combining different error measures to better estimate the actual errors.

The paper is organized as follows. We briefly describe relevant error estimation measures in next section. Description of different test problems and the numerical procedure followed are delineated in Section III. We compare different error measures and demonstrate the benefits of combining multiple error measures in the Section IV. Finally, we recapitulate the major conclusions.

II. Error Estimation Measures

Error in approximation at any design point $e(\mathbf{x})$ is defined as the difference between the actual function $y(\mathbf{x})$ and the predicted response $\hat{y}(\mathbf{x})$. If the actual function is known, the accuracy of different surrogates can be compared using the global prediction metrics, for example, maximum absolute error, or the integrated local error measures like root mean square error in the entire design space. The choice of appropriate measure depends on the application. The maximum absolute error may be more important for design of critical components but for most applications the root mean squared error (RMSE) in the design space (integrated local error measure) serves as a good measure of accuracy of approximation.

We integrate the pointwise (or local) error measures to estimate the root mean square error in desired space, as follows,

$$RMSE = \sqrt{\frac{1}{V}\int_V e^2(\mathbf{x})dV}, \tag{1}$$

where $e(\mathbf{x})$ is the estimated error at the design point $\mathbf{x}$ in the considered design space V. We evaluate Equation (1) using a numerical integration technique [6] as,

$$RMSE = \sqrt{\frac{1}{N_{test}}\sum_{j=1}^{N_{test}} \xi_j e_j^2}, \tag{2}$$

where ξ_j is the weight used for integration, and N_{test} is the number of test points. Note that the above equation can be used with any pointwise error measure.

Since the actual response in the design space is unknown (that is the primary reason of developing surrogates), we cannot compute the actual errors. So we appraise the quality of surrogates using appropriate error estimation measures like prediction variance or mean square errors. These error measures are usually based on certain assumptions on the actual functions, for example noise in the data for polynomial response surface approximations, Gaussian process for kriging, etc. The model independent (distribution-free) error measures that do not make any assumption on the data or the surrogate model are also being developed. In this section, we briefly describe the formulation of relevant parametric and model independent error measures for two popular surrogates.

A. Error Measures for Polynomial Response Surface Approximation

Polynomial response surface approximation (PRS, [7]) is still the most popular surrogate model among practitioners. The observed response $y(\mathbf{x})$ of a function at a point $\mathbf{x}$ is represented as a linear combination of the basis functions vector $\mathbf{f}(\mathbf{x})$ (mostly monomials are selected as basis functions) and the true coefficient vector $\boldsymbol{\beta}$, and the error ε. The error in the approximation ε is assumed to be uncorrelated, and normally distributed with zero mean and σ^2 variance. That is,

$$y(\mathbf{x}) = \mathbf{f}^T\boldsymbol{\beta} + \varepsilon; \qquad \varepsilon \in N(0,\sigma^2), \tag{3}$$

where N represents the normal distribution. The polynomial response surface approximation of the observed response $y(\mathbf{x})$ is,

$$\hat{y}(\mathbf{x}) = \mathbf{f}^T\hat{\boldsymbol{\beta}}, \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector of size $N_\beta \times 1$, where N_β is the number of coefficients in the polynomial response surface model. The error in approximation at i^{th} data point is given as

$$e_i = y_i - \hat{y}_i. \tag{5}$$

The coefficient vector $\hat{\boldsymbol{\beta}}$ can be obtained by minimizing a loss function L, defined as

$$L = \sum_{i=1}^{N_S} |e_i|^p, \tag{6}$$

where p is the order of loss function, and N_s is the number of sampled design points. We can minimize the absolute error or the maximum absolute error by using $p=1$ or $p = \infty$, respectively. However, here we use the conventional quadratic loss function ($p = 2$) that minimizes the sum of square of the residuals in approximation, unless specified otherwise. This loss function is the most popular because we can obtain the coefficient vector $\hat{\boldsymbol{\beta}}$ from the solution of a linear system of algebraic equations,

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}, \tag{7}$$

where X is the matrix of linear equations constructed using the basis functions $\mathbf{f}(\mathbf{x})$, and $\mathbf{y}$ is the vector of observed responses at N_s data points. The estimated variance of noise (or root mean square error) in approximation is given by,

$$(\hat{\sigma}_{prs})^2 = (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}) \Big/ (N_s - N_\beta). \tag{8}$$

For more details on polynomial response surface approximation, we refer the reader to the texts by Myers and Montgomery [7] and Khuri and Cornell [8]. The square root of estimated variance of noise $\hat{\sigma}_{prs}$ is often used as a rough estimate of the average actual root mean square error over the entire design space.

The actual absolute error in PRS approximation at a point $\mathbf{x}$ is given as

$$e^{prs}(\mathbf{x}) = |y(\mathbf{x}) - \hat{y}(\mathbf{x})|. \tag{9}$$

The standard error $e_{es}(\mathbf{x})$ (square root of prediction variance, [7]) is used to characterize the pointwise actual error in approximation as,

$$e_{es}(\mathbf{x}) = \sqrt{Var[\hat{y}(\mathbf{x})]} = \hat{\sigma}_{prs} \sqrt{\mathbf{f}^T(\mathbf{x}) \left(X^T X\right)^{-1} \mathbf{f}(\mathbf{x})}, \tag{10}$$

where $(\hat{\sigma}_{prs})^2$ is the estimated variance of noise, $\mathbf{f}(\mathbf{x})$ is the vector of basis functions used in approximation, and X is the matrix of linear equations constructed using $\mathbf{f}(\mathbf{x})$. This error estimate is based on the statistical assumption of uncorrelated noise in the data. While standard error is a local error measure, we use Equation (1) with e_{es} to obtain the integrated local error measure that is compared with the relevant actual root mean square error.

B. Error Measures for Universal Kriging

Kriging (KRG) is named after the pioneering work of D.G. Krige (a South African mining engineer) and is formally developed by Matheron [9]. Universal kriging estimates the response $y(\mathbf{x})$ at a design point $\mathbf{x}$ as the sum of a polynomial trend model $\sum_j \beta_j f_j(\mathbf{x})$, and a systematic departure term $Z(\mathbf{x})$ representing low (large scale) and high frequency (small scale) variations around the trend model,

$$y(\mathbf{x}) = \hat{y}(\mathbf{x}) = \sum_{j=1}^{N_\beta} \beta_j f_j + Z(\mathbf{x}). \tag{11}$$

The systematic departure components are assumed to be correlated as a function of distance between the locations under consideration. The Gaussian function is the most commonly used correlation function. The Gaussian correlation between some point $\mathbf{x}$ and the i^{th} design point $\mathbf{x}^{(i)}$ is given below.

$$r\left(Z(\mathbf{x}), Z(\mathbf{x}^{(i)}), \boldsymbol{\theta}\right) = \prod_{j=1}^{N_v} \exp\left(-\theta_j (x_j - x_j^{(i)})^2\right), \tag{12}$$

where N_v is the number of variables in the problem.

The predicted response at some point **x** is

$$\hat{y}(\mathbf{x}) = (\mathbf{r}(\mathbf{x}))^T R^{-1}\mathbf{y} - (X^T R^{-1}\mathbf{r}(\mathbf{x}) - \mathbf{f}(\mathbf{x}))^T (X^T R^{-1} X)^{-1} X^T R^{-1}\mathbf{y}, \quad (13)$$

where $\mathbf{r}(\mathbf{x})$ is the vector of correlations between the point **x** and the design points (whose components are given by Equation (12)), R is the matrix of correlations among design points, **y** is the vector of observed responses at data points, X is the Gramian design matrix constructed using basis functions in the trend model at the design points, and $\mathbf{f}(\mathbf{x})$ is the vector of basis functions in the trend model at point **x**. Note that, kriging interpolates data but may yield errors at unsampled locations.

Martin and Simpson [10] compared the maximum likelihood approach and a cross-validation based approach to estimate kriging parameters β_j, θ_j and found that the maximum likelihood approach to estimate kriging parameters was the best. We adopt the maximum likelihood approach to estimate kriging parameters β_j, θ_j [11].

The estimated process variance associated with the kriging approximation is given as,

$$(\hat{\sigma}_{krg})^2 = \frac{1}{N_s}(\mathbf{y} - X\hat{\boldsymbol{\beta}})^T R^{-1}(\mathbf{y} - X\hat{\boldsymbol{\beta}}), \quad (14)$$

where $\hat{\boldsymbol{\beta}}$ is the approximation of the coefficient vector $\boldsymbol{\beta}$ in Equation (11), and is given as $\hat{\boldsymbol{\beta}} = (X^T R^{-1} X)^{-1} X^T R^{-1}\mathbf{y}$. We note that the Equation (14) underestimates the actual process variance because this equation does not account for the fact that correlations R and **r** are estimated (Den Hertog et al. [12]). We explore if the square root of process variance $\hat{\sigma}_{krg}$ is a global estimate of root mean square of actual error in the entire design space.

The pointwise (local) estimate of actual error in kriging approximation is given by computing the mean squared error $\varphi(\mathbf{x})$ as follows [11].

$$\varphi(\mathbf{x}) = (\hat{\sigma}_{krg})^2 \left(\mathbf{1} + \mathbf{u}(\mathbf{x})^T (X^T R^{-1} X)^{-1} \mathbf{u}(\mathbf{x}) - \mathbf{r}(\mathbf{x})^T R^{-1}\mathbf{r}(\mathbf{x})\right), \quad (15)$$

$$\mathbf{u}(\mathbf{x}) = X^T R^{-1}\mathbf{r}(\mathbf{x}) - \mathbf{f}(\mathbf{x}), \quad (16)$$

where $(\hat{\sigma}_{krg})^2$ is the process variance (Equation (14)), and **1** is the vector of ones. Similar to the standard error for polynomial response surface approximation, the mean squared error also is a pointwise mean square estimate of the actual error. We use the square root of mean square error (MSE) to compare with the actual absolute error.

$$e_{mse}(\mathbf{x}) = \sqrt{\varphi(\mathbf{x})}. \quad (17)$$

As before, we integrate the mean square error $e_{mse}(\mathbf{x})$ using Equation (1) to compare with the actual root mean square error.

Note that, the integrated local error measures can characterize the actual root mean square error in an arbitrarily selected design domain whereas the global error measures are independent of the domain size (depend only on the data).

C. Model Independent Error Estimation Measure – Generalized Cross-validation Error

The error estimation measures discussed so far are suitable for particular surrogates. Next we discuss error measures that can be used with any surrogate (model independent or distribution-free error measures).

Generalized cross-validation error (GCV), also known as PRESS (predicted residual sum of squares) in the polynomial response surface approximation terminology ($GCV = PRESS / N_s$), is estimated by using the data at N_s points as follows. We fit surrogate models to $N_s - 1$ points by leaving one design point at a time and predict response at the left out point. Then GCV is defined as,

$$GCV = \frac{1}{N_s}\sum_{i=1}^{N_s}(y_i - \hat{y}_i^{(-i)})^2, \quad (18)$$

where $\hat{y}_i^{(-i)}$ represents the prediction at the left-out point $\mathbf{x}^{(i)}$ using the surrogate constructed from all the design points except ($\mathbf{x}^{(i)}$, y_i). We use the square root of generalized cross-validation error to estimate actual root mean square error in the approximation and compare its performance with other global error estimation measures. Though we have shown a global GCV, a local counterpart of the GCV can also be developed.

An analytical estimate of GCV is available for polynomial response surface approximation (Myers and Montgomery [7]). Mitchell and Morris [13] and Currin et al. [14] provided computationally inexpensive expressions to evaluate cross-validation error for kriging using a constant trend model while holding other model parameters constant. Martin [15], and Congdon and Martin [16] extended this analytical estimate of cross-validation error to account for more complex trend functions while keeping the model parameters constant. However, here we estimate GCV using Equation (18). While we study leave-one-out GCV, Meckesheimer et al. [5] studied the influence of number of points in sub-samples to compute cross-validation error and gave their recommendations on sub-sample size to estimate GCV for different surrogates. They recommended using leave-one-out GCV for both polynomial and kriging models as is used in this study.

D. Ensemble of Error Estimation Measures

Finally, we explore the concept of an ensemble of error measures that is inspired by our previous work on the ensemble of surrogates [4]. Since there are many error estimation measures for different surrogates, we can simultaneously use the information to better estimate the actual errors. In this context, a few ideas are as follows.

1. Simultaneously use multiple error measures to identify regions of high errors. This idea is useful to reduce the probability of failing to detect high error regions using a single error measure.
2. Identify a suitable error measure using a cross-validation based approach, that is, we can compare predicted errors and actual errors at data points while computing the cross-validation errors using a suitable criterion and select the best error measure.
3. Get a better estimate of the actual errors by averaging different predicted error measures. This idea can be applied to different surrogates and error measures.

A detailed study of different approaches to use an ensemble of error measures is given by Goel [17]. We only demonstrate results for the averaging of multiple error measures in this study for the sake of brevity.

III. Test Problems and Testing Procedure

We illustrate the numerical procedure adopted to evaluate the capabilities of different error measures in Figure 1. Firstly, we describe the test problems used in this study, followed by a stepwise description of the numerical procedure used to compare different error measures.

A. Test Problems

To test the predictive capabilities of different error measures, we employ two types of problems ranging from two to six variables: (i) analytical functions (Dixon-Szegö [18]) that are often used to test the global optimization methods, and (ii) engineering problem: a cantilever beam design problem (Wu et al. [19]) that is extensively used as a test problem in reliability analysis. The details of each test problem are given as follows.

III.A.1. Branin-Hoo function (BH)

$$f(x_1,x_2)=\left(x_2-\frac{5.1x_1^2}{4\pi^2}+\frac{5x_1}{\pi}-6\right)^2+10\left(1-\frac{1}{8\pi}\right)\cos(x_1)+10, \tag{19}$$

$$x_1\in[-5,10], \qquad x_2\in[0,15].$$

III.A.2. Camelback function (CB)

$$f(x_1,x_2)=\left(4-2.1x_1^2+\frac{x_1^4}{3}\right)x_1^2+x_1x_2+\left(-4+4x_2^2\right)x_2^2, \tag{20}$$

$$x_1\in[-3,3], \qquad x_2\in[-2,2].$$

III.A.3. Goldstein-Price function (GPR)

$$f(x_1,x_2)=\left[1+(x_1+x_2+1)^2\left(19-4x_1+3x_1^2-14x_2+6x_1x_2+3x_2^2\right)\right]\times$$

$$\left[30+(2x_1-3x_2)^2\left(18-32x_1+12x_1^2+48x_2-36x_1x_2+27x_2^2\right)\right], \tag{21}$$

$$x_1,x_2\in[-2,2].$$

III.A.4. Hartman functions

$$f(\mathbf{x}) = -\sum_{i=1}^{m} c_i \exp\left\{-\sum_{j=1}^{N_v} a_{ij}\left(x_j - p_{ij}\right)^2\right\},$$
$$\text{where } \mathbf{x} = \left(x_1, x_2, \ldots, x_{N_v}\right) \quad x_i \in [0,1]. \tag{22}$$

Two instances of this problem are considered based on the number of design variables. For the chosen examples, $m = 4$.

a. Hartman-3 (HM3)

This problem has three variables. The choice of parameters is given in Table 1 [18].

b. Hartman-6 (HM6)

This instance of the problem has six design variables and the parameters used in the function are tabulated in Table 2 [18]. For this case, all design variables were in the range [0, 0.5] instead of [0, 1].

III.A.5. Cantilever beam design problem [19] (CT5)

The cantilever beam design problem (Figure 2) that is widely used for reliability analysis is illustrated as follows. The displacement of the beam G_d, is defined as:

$$G_d = D_0 - D = D_0 - \frac{4L^3}{Ewt}\sqrt{\left(\left(\frac{F_y}{t^2}\right)^2 + \left(\frac{F_x}{w^2}\right)^2\right)}, \tag{23}$$

where D_o is the allowable initial deflection of the beam (taken as *2.25"*), L is the length of the beam (fixed at *100"*). We develop surrogate models of G_d as a function of the Young's modulus (E), the horizontal loading (F_x), the vertical loading (F_y), the width (w), and the thickness (t) of the beam. The ranges of the design variables are given in Table 3.

B. Testing Procedure

For each example, we appraise the capabilities of different error measures as shown in the loop in Figure 1. A detailed description of each step is as follows.

III.B.1. Create experimental designs (ED)

We select an experimental design to construct surrogate models of the example problem. There are different experimental designs that cater to bias and noise errors, for example Latin hypercube sampling [20] is a stratified sampling plan that reduces the bias error, or D-optimal design [7] minimizes the D-efficiency defined as,

$$D_{eff} = \left(|M| \Big/ \max|M|\right)^{1/N_\beta}, \quad |M| = \left|X^T X\right| \Big/ N_s^{N_\beta}, \tag{24}$$

where M is the moment matrix. Goel et al. [21] demonstrated that a combination of the Latin hypercube sampling (LHS) and the D-optimality criterion to construct experimental designs is helpful in reducing noise and bias errors simultaneously. In this procedure, we generate a large sample of LHS points using MATLAB [22] routine 'lhsdesign' with 'maximin' criterion that maximizes the minimum distance between points. We allocate a maximum of 100 iterations for optimization. We use this LHS sample as the initial grid of points for MATLAB routine 'candexch' to select N_s design points without duplication, using D-optimality criterion. We allow a maximum of 40 iterations to find the D-optimal design from the large LHS sample. The number of design points (N_s), order of the polynomial to estimate D-optimality, and the number of points used in the large LHS sample (N_{lhs}) are given in Table 4.

We note that there is an uncertainty in the results due to the random components in experimental designs and the possibility of convergence to local optima for both LHS and D-optimal design. To reduce this uncertainty due to the choice of experimental design, we present results based on multiple instances of experimental designs for all problems. As shown in Appendix, a small number (100) of experimental designs yields high variability in the predicted accuracy of error estimates. Hence, we estimate results using 1000 experimental designs.

III.B.2. Surrogate construction

Next, we construct surrogate models using the data evaluated at ED sampled points. The order of polynomial used for polynomial response surface approximation is given in Table 4. We used Equation (7) for PRS approximation in the least squares sense. We used the kriging software developed by Lophaven et al. [11] with a linear trend model and the Gaussian correlation function. Lophaven et al. [11] require bounds on the parameters governing the Gaussian distribution that were taken as [0.01, 200]. To reduce the risk of getting

stuck into local optima, while estimating kriging parameters using the maximum likelihood function, we use five initial guess for the optimizer used by Lophaven et al. [11]. We select the model that maximizes the likelihood function. The surrogate models were constructed in the normalized design space for all the problems, such that the minimum value of any design variable was translated to zero and the maximum value was translated and scaled to one.

III.B.3. Global error estimates

We estimated global error measures, namely, estimated root mean square error for PRS (Equation (8)), process variance for kriging (Equation (14)), and generalized cross-validation errors for each surrogate using Equation (18). These error measures characterize the actual RMS errors in approximation in the entire design domain.

III.B.4. Test points and pointwise error estimates

To evaluate the actual RMS errors in the entire design domain, we identify the location of test points. For two- and three- dimensional spaces, the surrogate models were tested using a uniform grid of p^{N_v} points where N_v is the number of variables, and p is the number of points along each direction (given in Table 4). We used 2000 points[‡], selected *via* optimized LHS, to assess the performance of different error estimation measures for the cantilever beam design problem and the Hartman function with six variables. We computed the pointwise responses and the actual absolute errors at these test points using different surrogates. We also evaluated the pointwise errors namely, estimated standard error (Equation (10)) for PRS, and mean square error (Equation (17)) for kriging.

III.B.5. Integrated pointwise error estimates

To have a fair comparison of the global error measures that characterize the actual RMS errors with the local error measures, we use the integrated local error measures. We numerically integrate predicted and actual errors in the entire design domain using Equation (2) to get the root mean square errors. In low dimensional spaces, we use a uniform grid of test points so the weights (ξ_i) are determined according to the trapezoidal rule. For the cantilever beam design problem and the Hartman function with six variables, we use a quasi-random set of test points so we choose $\xi_i = 1$ in Equation (2). As discussed earlier, the integrated local error measures may characterize the integrated actual errors better in the selected domain of interest.

III.B.6. Error ratios

To compare the global error prediction capabilities of different error measures, we estimate the ratio of predicted global error or the root mean square local error measure with the actual root mean square error in the entire design space. The desired value of this ratio is one. A value less than one indicates that the actual RMS error is underestimated by the predictor, and a value greater than one indicates that the global error measure overestimates the magnitude of actual RMS error. Finally, we summarize the results for 1000 experimental designs using its mean and coefficient of variation.

IV. Results and Discussion

A. Errors in Approximations

We normalized the actual RMS error by the range of any individual function in the entire design space, and show the mean and coefficient of variation of the normalized actual RMS errors based on 1000 EDs in Table 5. A low value of the normalized actual RMS error indicates good approximation of the actual function. Polynomial response surface approximation was a good surrogate for all problems, except the Hartman function with three variables (10% or more errors in approximation). Kriging performed reasonably well for all the problems, except the Camelback function and the Hartman function with six variables. Notably, the PRS model yielded much lower coefficient of variation compared to other surrogate models. This suggests that the PRS model is a more reliable approximation. The higher variability in the kriging approximation is attributed to the variability in maximum likelihood estimates of the model parameters.

B. Accuracy of Error Measures

We summarize the mean and coefficient of variation (based on 1000 EDs) of the ratio of the predicted (global and integrated local) and the actual RMS errors for different problems in Table 6. The graphical representation of different error measures for polynomial response surface approximation and kriging, are presented in Figure 3 and Figure 4.

[‡] We observed that 2000 LHS points estimate the mean of the Hartman-6 function with reasonable accuracy. Using the trapezoidal rule the mean was estimated as -0.511 and the coefficient of variation of the mean was -0.019, and the mean estimated using the LHS points was -0.522 and the coefficient of variation of the mean was -0.018.

No single error measure outperforms all other error measures. We observe that the predicted (global and integrated local) error estimates for polynomial response surface approximation (PRS) provide more accurate (mean value is close to one) and robust (low coefficient of variation) estimates of the actual RMS error than their kriging counterparts. This is remarkable because the error in the examples considered here is bias error rather than noise, where kriging is supposed to perform better. For PRS, the parametric global (estimated RMSE) and the integrated local (standard error) measures underestimate the actual RMS errors but have the lowest variation (low coefficient of variation) with the choice of ED and the problem. The model independent global error measure (generalized cross validation error) overestimates the actual RMS error and yields low coefficient of variation with respect to the problem and the experimental design for PRS.

For kriging, the model based global error measure (process variance) significantly overestimates the actual RMS errors. This error measure is very sensitive to the choice of ED and the nature of the problem. The mean squared error that is a parametric local error measure provides the best estimate of the actual RMS error for kriging. It has the lowest variation with the experimental designs and the selection of the test problem. As was observed for PRS, the GCV serves as a good measure of the actual RMS error for kriging too. This measure overestimates the actual RMS error but has low sensitivity with respect to the choice of experimental design and the test problem compared to the global parametric error measure. Compared to the GCV for PRS, the GCV for kriging has higher overestimate and larger variation with the experimental design and the problem. There are some interesting observations from the results shown in Table 5 and Table 6.

1. The variation in the actual RMS error in PRS approximation with respect to the experimental designs is lower than the variation in its error estimates (estimated root mean square error, and standard error), possibly because the fitting process is not influenced by the statistical assumptions while the error estimate is governed by the assumptions that do not apply to the data,
2. Often the actual RMS error is underestimated by one error measure and overestimated by others such that we can benefit by averaging different error measures,
3. The integrated parametric local error measures (though computationally expensive) provide an accurate and reliable estimate of the actual RMS error in the design space. The other advantage is that these error measures can be directly tailored to account for the size of domain in consideration,
4. For PRS, the estimated RMSE is the best error measure, and
5. Though computationally expensive, model independent global error measure-the cross-validation error-performs well for both surrogates. This measure overestimates the actual RMS errors.

We conclude that despite relatively higher computational cost compared to the parametric global error measures, it is useful to estimate the parametric local errors especially when the cost of error prediction is significantly lower than the cost of acquiring the data.

C. Averaging Multiple Error Measures

As suggested in previous section, we may better estimate the magnitude of the actual RMS error by averaging different error estimates. To explore this possibility, we compared the global error estimates obtained using the geometric mean[§] of, i) the GCV and the estimated root mean square error (global measures) for PRS, ii) the GCV (global measure) and the integrated standard error (local measure) for PRS, and iii) the GCV (global measure) and the integrated mean squared error (local measure) for kriging, with the actual RMS errors in PRS or kriging, respectively. The mean and the coefficient of variation (based on 1000 experimental designs) of the ratios of the predicted and the actual RMS errors for different problems are summarized in Table 7.

We compare the performance of the averaged error measures with the performance of its constituent error measures, i.e., columns 2 and 3 in Table 7 is compared to columns 3 and 5 in Table 6; columns 4 and 5 in Table 7 with columns 4 and 5 in Table 6; and columns 6 and 7 in Table 7 with columns 8 and 9 in Table 6. We observe that compared to the individual error measures, the average of errors is consistently closer to one for different problems and results in similar coefficient of variation with respect to different experimental designs. The averaged error measure is significantly better than the worse of the two error measures and is at par or better than the better of the two. Thus, we can say that averaging of the error measures may be beneficial.

V. Conclusions and Future Work

We compared the accuracy of different error estimators using a suite of analytical and engineering test problems. The main conclusions of the paper are as follows:

1. We found that, though the error in approximation is due to model inadequacy instead of noise, the error measures for polynomial response surface approximations were better than those for kriging.
2. While the estimated root mean square error for PRS estimated the actual root mean squared errors accurately, the process variance for kriging performed poorly.

[§] Geometric mean g of two numbers x and y is defined as $g = \sqrt{xy}$.

3. On the other hand, computationally expensive integrated parametric pointwise error measures (standard error for polynomials and mean squared error for kriging) characterized the actual root mean square errors very well for both surrogates.
4. The model independent global error measures (PRESS) provided a robust estimate of the actual RMS errors but they overestimated the actual RMS errors for both PRS and kriging.
5. We observed improvement in prediction capabilities by simultaneously using multiple error measures. The geometric averaging helped reduce variability in the error predictions and yielded better error estimates compared to the individual error measures.

In this study, we appraised different error measures according to the global error estimation metrics. We observed that ensemble of error measures holds promise but more research is needed to explore the suitability of different error measures for averaging, and other methods of combining error measures. We also intend to study the local error estimation capabilities of different error measures i.e., how well different error measures characterize the actual error field?

VI. Acknowledgements

The present efforts have been supported by the Institute for Future Space Transport (under the NASA Constellation University Institute Program (CUIP) with Ms. Claudia Meyer as program monitor) and National Science Foundation (Grant # DMI-423280). Authors also thank Dr Nestor V. Queipo for many useful comments.

VII. References

[1]. Queipo NV, Haftka RT, Shyy W, Goel T, Vaidyanathan R, Tucker PK (2005), Surrogate-based analysis and optimization, Progress in Aerospace Sciences, 41: 1-28.
[2]. Jones DR, Schonlau M, Welch WJ (1998), Efficient global optimization of expensive black-box functions, Journal of Global Optimization, 13 (4): 455-492.
[3]. Kleijnen JPC, van Beers WCM (2004), Application-driven sequential designs for simulation experiments: Kriging metamodeling, Journal of the Operational Research Society, 55 (9): 876-883.
[4]. Goel T, Haftka RT, Shyy W, Queipo NV (2007), Ensemble of surrogates, Structural and Multi-disciplinary Optimization Journal, 196 (4-6): 879-893.
[5]. Meckesheimer M, Barton RR, Simpson TW, Booker A (2002), Computationally inexpensive metamodel assessment strategies, AIAA Journal, 40(10): 2053-2060.
[6]. Ueberhuber CW (1997), Numerical computation 2: Methods, software and analysis, Springer-Verlag: Berlin, p.71.
[7]. Myers RH, Montgomery DC (1995), Response surface methodology-process and product optimization using designed experiments, John Wiley & Sons, Inc: New York.
[8]. Khuri AI, Cornell JA (1996), Response surfaces: Designs and analysis, Marcel Dekker, Inc: New York.
[9]. Matheron G (1963), Principles of geostatistics, Economic Geology, 58: 1246-1266.
[10]. Martin JD, Simpson TW (2005), Use of kriging models to approximate deterministic computer models, AIAA Journal, 43 (4): 853–863.
[11]. Lophaven SN, Nielsen HB, Sondergaard J, DACE: A MATLAB kriging toolbox, version 2.0, Information and Mathematical Modeling, Technical University of Denmark, 2002.
[12]. Den Hertog D, Kleijnen JPC, Siem AYD (2006), The correct kriging variance estimated by bootstrapping. Journal of the Operational Research Society, 57 (4): 400-409.
[13]. Mitchell TJ, Morris MD (1992), Bayesian design and analysis of computer experiments: Two examples, Statistica Sinica, 2 (2): 359-379.
[14]. Currin C, Mitchell TJ, Morris MD, Ylvisaker D (1998), A bayesian approach to the design and analysis of computer experiments, Oak Ridge National Laboratory, Oak Ridge TN, Technical Report ORNL-6498.
[15]. Martin JD (2005), A methodology for evaluating system-level uncertainty in the conceptual design of complex multidisciplinary systems, PhD Thesis, Dept. of Mechanical Engineering, The Pennsylvania State University, University Park PA.
[16]. Congdon CD, Martin JD (2007), On using standard residuals as a metric of kriging model quality, In: 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, Honolulu HI, 23 Apr – 26 Apr, AIAA-2007-1928.
[17]. Goel T (2007), Multiple surrogates and error modeling in optimization of liquid rocket propulsion components, PhD Thesis, Dept. of Mechanical and Aerospace Engineering, The University of Florida, Gainesville FL.
[18]. Dixon LCW, Szegö GP (1978), Towards global optimization 2, North-Holland, Amsterdam.
[19]. Wu YT, Shin Y, Sues R, Cesare M (2001), Safety factor based approach for probability-based design optimization, In: 42nd AIAA/ ASME/ ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference, Seattle WA, 2001, AIAA-2001-1522.

[20]. McKay M, Conover W, Beckman R (1979), A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21: 239–245.
[21]. Goel T, Haftka RT, Shyy W, Watson LT (2007), Pitfalls of using a single criterion for selecting experimental designs, accepted for publication in International Journal of Numerical Methods in Engineering.
[22]. MATLAB®, The Language of Technical Computing, Version 6.5 Release 13. © 1984-2002, The MathWorks Inc.

Table 1 Parameters used in the Hartman function with three variables

i	aij			ci	pij		
1	3.0	10.0	30.0	1.0	0.3689	0.1170	0.2673
2	0.1	10.0	35.0	1.2	0.4699	0.4387	0.7470
3	3.0	10.0	30.0	3.0	0.1091	0.8732	0.5547
4	0.1	10.0	35.0	3.2	0.03815	0.5743	0.8828

Table 2 Parameters used in the Hartman function with six variables

i	a_{ij}						c_i
1	10.0	3.0	17.0	3.5	1.7	8.0	1.0
2	0.05	10.0	17.0	0.1	8.0	14.0	1.2
3	3.0	3.5	1.7	10.0	17.0	8.0	3.0
4	17.0	8.0	0.05	10.0	0.1	14.0	3.2
i	p_{ij}						
1	0.1312	0.1696	0.5569	0.0124	0.8283		0.5886
2	0.2329	0.4135	0.8307	0.3736	0.1004		0.9991
3	0.2348	0.1451	0.3522	0.2883	0.3047		0.6650
4	0.4047	0.8828	0.8732	0.5743	0.1091		0.0381

Table 3 Ranges of variables for the cantilever beam design problem

Variable	Minimum	Maximum	Units
F_x	700	1300	lbs
F_y	900	1300	lbs
E	20e6	35e6	psi
w	2.0	3.0	inch
t	3.0	5.0	inch

Table 4 Numerical setup for different test problems. N_v: number of variables, N_s: number of training points, N_{test} is number of test points, for analytical problems $N_{test} = p^{N_v}$, where p is the number of points along each direction. For the Hartman-6, and the cantilever beam design problem, we specify the number of test points N_{test} in table. N_{lhs}: the number of points in the large LHS sample.

	N_v	N_s	p or N_{test}	Order of polynomial	N_{lhs}
Branin-Hoo	2	20	16	3	100
Camelback	2	30	16	4	150
GoldStein-Price	2	42	16	5	200
Hartman-3	3	70	11	4	250
Hartman-6	6	56	2000	2	200
Cantilever	5	42	2000	2	210

Table 5 Mean and coefficient of variation (CV) (based on 1000 experimental designs) of the normalized actual RMS error in the entire design space. We use the range of different functions to normalize the respective actual RMS errors. PRS: polynomial response surface approximation (the value in parenthesis is the order of loss function used to estimate coefficients), KRG: Kriging,

		PRS	KRG
Branin-Hoo	Mean	0.028	0.022
	CV	0.075	0.570
Camelback	Mean	0.041	0.112
	CV	0.050	0.271
Goldstein-Price	Mean	0.024	0.017
	CV	0.096	0.292
Hartman-3	Mean	0.109	0.035
	CV	0.065	0.301
Hartman-6	Mean	0.073	0.078
	CV	0.072	0.202
Cantilever	Mean	0.015	0.016
	CV	0.149	0.316

Table 6 Mean and coefficient of variation (CV) (based on 1000 experimental designs) of ratio of the predicted and the actual RMS error. $\hat{\sigma}_{prs}$: the estimated root mean square error for PRS, e_{se}^{prs} : the estimated standard error for PRS, GCV^{prs} : square root of the generalized cross-validation error for polynomial response surface approximation (PRS), $\hat{\sigma}_{krg}$: square root of the process variance for kriging, GCV^{krg} : square root of the generalized cross-validation error for kriging. (All-problems: indicate the mean and COV for all 6000 EDs). *Median is 3.09.

		$\hat{\sigma}_{prs}$	e_{se}^{prs}	GCV^{prs}	$\hat{\sigma}_{krg}$	e_{mse}^{krg}	GCV^{krg}
Branin-Hoo	Mean	0.74	0.51	1.17	44.71	0.74	3.07
	CV	0.24	0.24	0.29	0.80	0.50	0.79
Camelback	Mean	0.75	0.54	1.22	2.88	0.99	1.29
	CV	0.17	0.17	0.21	0.52	0.25	0.30
Goldstein-Price	Mean	1.09	0.81	1.93	21.04	1.00	2.39
	CV	0.20	0.20	0.26	0.60	0.40	0.45
Hartman-3	Mean	0.82	0.64	1.30	5.53	1.20	1.51
	CV	0.15	0.15	0.18	0.32	0.33	0.34
Hartman-6	Mean	0.80	0.58	1.18	1.35	0.80	1.17
	CV	0.18	0.18	0.19	0.38	0.24	0.30
Cantilever	Mean	0.97	0.66	1.45	5.38	1.06	1.81
	CV	0.18	0.18	0.20	0.68	0.24	0.35
All problems	Mean	0.86	0.63	1.38	13.48*	0.97	1.87
	CV	0.24	0.25	0.31	1.62	0.37	0.71

Table 7 Mean and CV (based on 1000 experimental designs) of ratio of the average root mean square errors and the actual RMS error for different test problems. GCV^{prs} : square root of the generalized cross validation error for PRS, $\hat{\sigma}_{prs}$: square root of the root mean square error, e_{se}^{prs} : the estimated standard error for PRS , e_{mse}^{krg} : square root of the mean squared error. (All problems: indicate the mean and CV for all 6000 EDs).

	$\sqrt{GCV^{prs} \times \hat{\sigma}_{prs}}$		$\sqrt{GCV^{prs} \times e_{se}^{prs}}$		$\sqrt{GCV^{krg} \times e_{mse}^{krg}}$	
	Mean	CV	Mean	CV	Mean	CV
Branin-Hoo	0.93	0.25	0.77	0.25	1.46	0.58
Camelback	0.96	0.18	0.81	0.18	1.12	0.23
Goldstein-Price	1.45	0.21	1.25	0.21	1.50	0.33
Hartman-3	1.04	0.16	0.92	0.16	1.33	0.31
Hartman-6	0.97	0.18	0.82	0.18	0.97	0.25
Cantilever	1.18	0.19	0.98	0.19	1.36	0.24
All problems	1.09	0.26	0.93	0.27	1.29	0.40

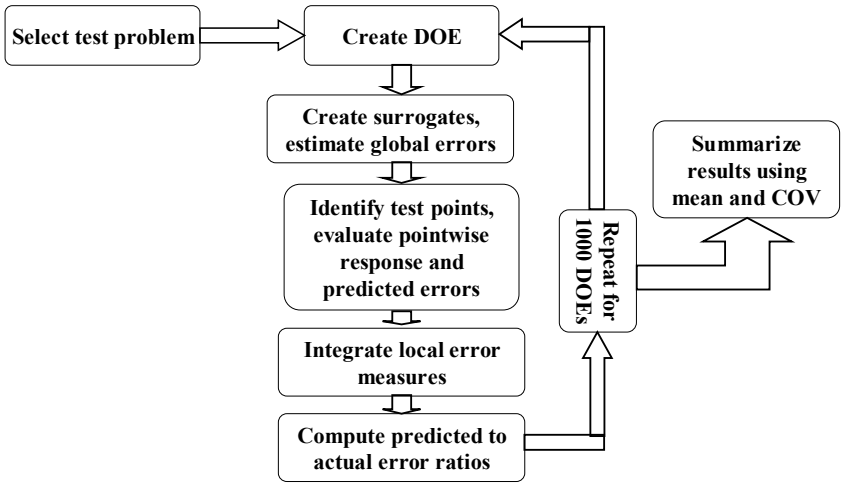


Figure 1 Flow chart of the test procedure.

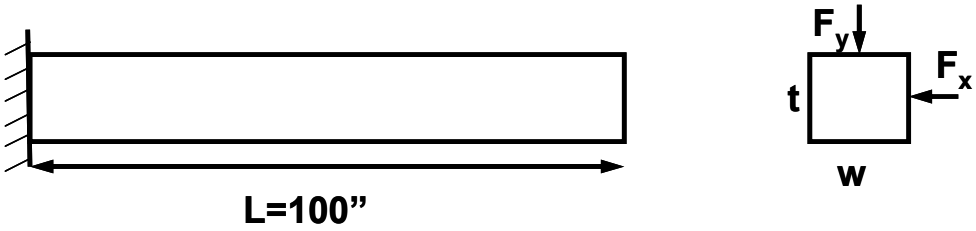
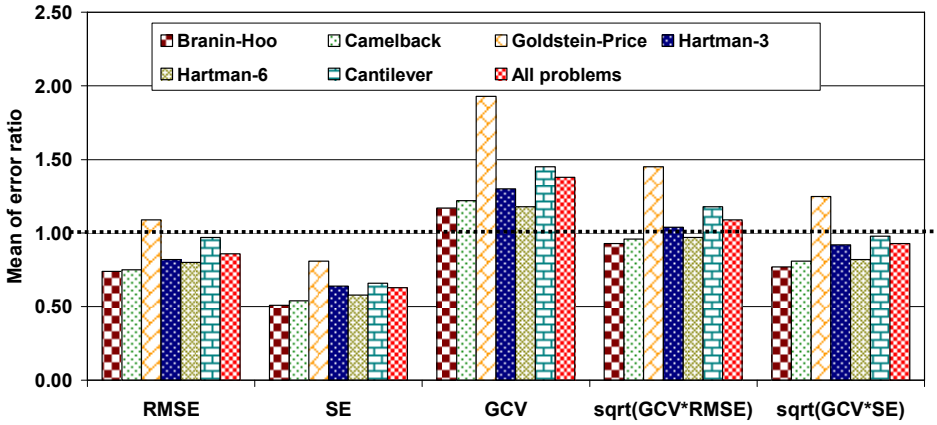
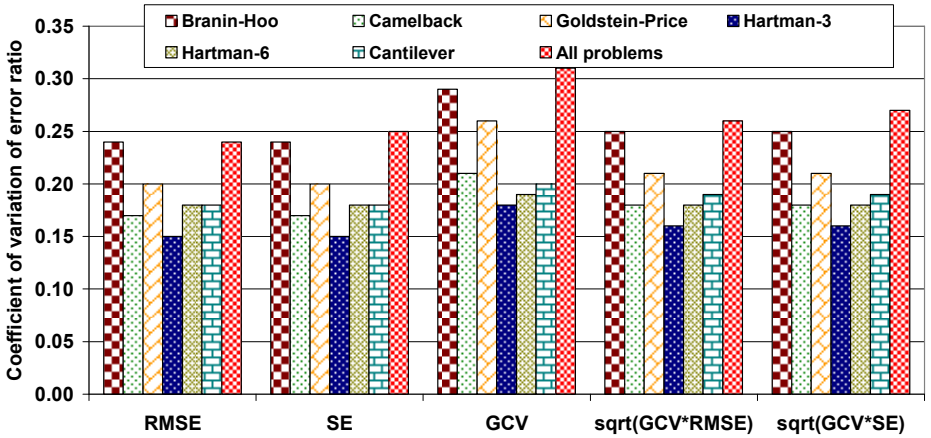


Figure 2 Cantilever beam subjected to horizontal and vertical random loads.

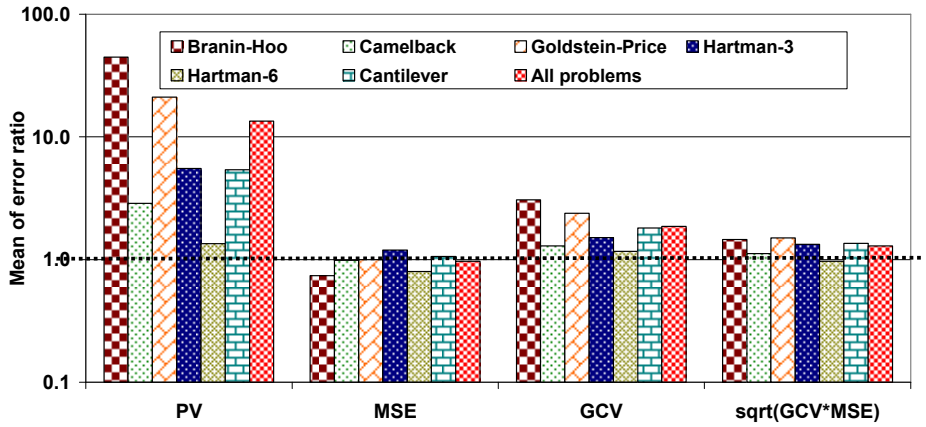


(A) Mean of the ratio of the predicted and the actual RMS errors

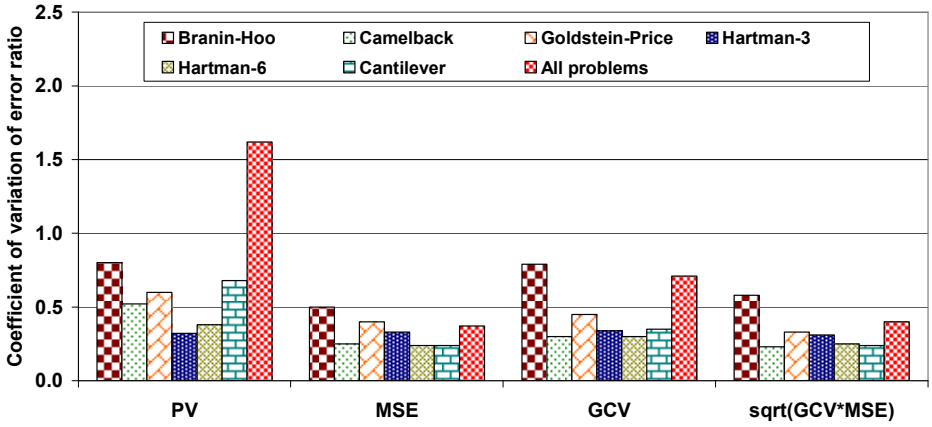


(B) Coefficient of variation of the ratio of the predicted and the actual RMS errors

Figure 3 Mean and coefficient of variation of different error measures for polynomial response surface approximation.



(A) Mean of the ratio of the predicted and the actual RMS errors



(B) Coefficient of variation of the ratio of the predicted and the actual RMS errors

Figure 4 Mean and coefficient of variation of different error measures for kriging approximation.

VIII. Appendix: Influence of Number of Experimental Designs

We show the improvement in the actual error characterization ability by using the geometric averaging of multiple error measures. We estimate the mean and coefficient of variation (CV) of the ratio of averaged predicted RMS errors and actual RMS errors based on 100 experimental designs (Table 8). We note that most results are not significantly different than those given in Table 7 (based on 1000 experimental designs) and the conclusions derived earlier are valid with fewer experimental designs too. However, the coefficient of variation of the mean values (CV/$\sqrt{N_{ED}}$) given in Table 8 (N_{ED} = 100) is significantly higher than the CV of the mean values given in Table 7 (N_{ED} = 1000). The consequence of this result is clear for Hartman-3 example problem where the predicted mean ratio of geometric average of kriging error measures and the actual RMS error obtained from 100 experimental designs (mean is 1.20, and the coefficient of variation of the mean is 0.021) is statistically different than the corresponding mean ratio obtained using 1000 experimental designs (mean is 1.33 and CV of the mean is 0.01). This result clearly demonstrates the importance of using 1000 experimental designs instead of 100 experimental designs.

Table 8 Mean and CV (based on 100 experimental designs) of ratio of the average root mean square errors and the actual RMS error for different test problems. GCV^{prs} : square root of the generalized cross validation error for PRS, $\hat{\sigma}_{prs}$: square root of the root mean square error, e_{se}^{prs} : the estimated standard error for PRS , e_{mse}^{krg} : square root of the mean squared error. (All problems: indicate the mean and CV for all 600 EDs).

	$\sqrt{GCV^{prs} \times \hat{\sigma}_{prs}}$		$\sqrt{GCV^{prs} \times e_{se}^{prs}}$		$\sqrt{GCV^{krg} \times e_{mse}^{krg}}$	
	Mean	CV	Mean	CV	Mean	CV
Branin-Hoo	0.91	0.25	0.76	0.25	1.44	0.64
Camelback	0.99	0.16	0.84	0.16	1.14	0.22
Goldstein-Price	1.49	0.22	1.28	0.22	1.46	0.31
Hartman-3	1.03	0.16	0.91	0.16	1.20	0.32
Hartman-6	0.99	0.19	0.84	0.19	1.00	0.26
Cantilever	1.17	0.20	0.97	0.20	1.32	0.21
All problems	1.10	0.27	0.93	0.27	1.26	0.41