AIAA JOURNAL
Vol. 43, No. 4, April 2005

Use of Kriging Models to Approximate Deterministic Computer Models

Jay D. Martin*
Applied Research Laboratory, State College, Pennsylvania 16804-0030
and
Timothy W. Simpson†
Pennsylvania State University, University Park, Pennsylvania 16802

The use of kriging models for approximation and metamodel-based design and optimization has been steadily on the rise in the past decade. The widespread use of kriging models appears to be hampered by 1) computationally efficient algorithms for accurately estimating the model's parameters, 2) an effective method to assess the resulting model's quality, and 3) the lack of guidance in selecting the appropriate form of the kriging model. We attempt to address these issues by comparing 1) maximum likelihood estimation and cross validation parameter estimation methods for selecting a kriging model's parameters given its form and 2) an R^2 of prediction and the corrected Akaike information criterion assessment methods for quantifying the quality of the created kriging model. These methods are demonstrated with six test problems. Finally, different forms of kriging models are examined to determine if more complex forms are more accurate and easier to fit than simple forms of kriging models for approximating computer models.

I. Introduction

KRIGING models have become a popular method for approximating deterministic computer models.[1–7] They have been used in a variety of applications including conceptual design,[8] structural optimization,[9] multidisciplinary design optimization,[10] aerospace engineering,[11] and mechanical engineering.[12] Kriging models offer a good choice for these types of applications because of their flexibility to approximate many different and complex response functions. They are also a good choice for approximating deterministic computer models because they can interpolate the observed or known data points.[13,14] The widespread use of kriging models appears to be hampered by 1) the lack of guidance in selecting the appropriate form of the kriging model (see Sec. II.A), 2) a computationally efficient algorithm for estimating the model's parameters (see Sec. II.B), and 3) a method to assess effectively the model's quality (see Sec. II.C). In this work we investigate these three aspects of using kriging models to approximate deterministic computer models and attempt to draw conclusions based upon the results of creating kriging models for six test problems: a one-dimensional, four two-dimensional, and one five-dimensional functions (see Sec. III). For each test problem, many different forms of kriging models are fit using two parameter selection methods (maximum likelihood estimation and cross validation) for each model's form. All of the resulting models are then assessed for their ability to approximate the original deterministic computer model. Two assessment methods that employ only the data used to fit the models are compared. All of the models are also assessed with large validation data sets to provide an approximation of the true errors. The results are analyzed in Sec. IV, and conclusions are given in Sec. V.

Received 28 February 2004; revision received 13 September 2004; accepted for publication 14 September 2004.

*Research Assistant, Applied Research Laboratory. Member AIAA.
†Associate Professor, Departments of Mechanical and Nuclear Engineering and Industrial and Manufacturing Engineering. Member AIAA.

II. Overview of Kriging

Kriging was initially developed by geologists to estimate mineral concentrations over an area of interest given a set of sampled sites from the area[15]; it was also introduced about the same time in the field of statistics to include the correlations that exist in the residuals of a linear estimator.[16] There are many texts in geostatistics[17,18] and in spatial statistics[19–21] that provide many details on the development and use of kriging models in their respective disciplines. This section will cover the details that are important for using kriging models to approximate deterministic computer models.

A kriging model is a generalized linear regression model that accounts for the correlation in the residuals between the regression model and the observations.[16] Given the mathematical form of kriging (see Sec. II.A), the process of using kriging first requires the estimation of the "best" model parameters (see Sec. II.B), and an assessment of the resulting kriging model's accuracy (see Sec. II.C) before it can be used as an approximation to a deterministic computer model.

A. Mathematical Form

The mathematical form of a kriging model has two parts as shown in Eq. (1). The first part is a linear regression of the data with k regressors modeling the drift of the process mean, also called the "trend," over the domain. Most previous engineering applications use a constant trend model over the domain[14] and rely on the second part of the model to "pull" the response surface through the observed data by quantifying the correlation of nearby points.

$$\hat{\boldsymbol{y}}(\boldsymbol{X}) = \sum_{j=1}^{k} \beta_j f_j(\boldsymbol{X}) + Z(\boldsymbol{X}) \tag{1}$$

The second part $Z(\boldsymbol{X})$ is a model of a Gaussian and stationary random process with zero mean and covariance:

$$V(x_1, x_2) = \sigma^2 \boldsymbol{R}(x_1, x_2) \tag{2}$$

The process variance σ^2 is a scalar of the spatial correlation function (SCF) $\boldsymbol{R}(x_1, x_2)$, which controls the smoothness of the resulting kriging model, the influence of nearby points, and the differentiability of the surface by quantifying the correlation between two observations. Koehler and Owen[22] provide an overview of four common SCFs used for approximating a deterministic computer model and

detail the impact of selecting different parameter values for these functions.

There is little guidance in the literature on the selection of the best form of the SCF. Statisticians suggest the use of the Matérn family of SCFs (Ref. 21, p. 12). The Gaussian function is the most commonly used SCF in engineering design[14] as it provides a relatively smooth and infinitely differentiable surface, making it a better choice when used with gradient-based optimization algorithms. It is the SCF used in this work and is defined with only one parameter θ, which controls the range of influence of nearby points[11,22,23] as follows:

$$\boldsymbol{R}(x_1 - x_2) = e^{-\theta|x_2 - x_1|^2}, \qquad \text{where} \qquad \theta > 0 \tag{3}$$

The correlation function range parameter θ used in Eq. (3) has little meaning in a physical sense. An alternative form of the range parameter for the Gaussian SCF is (Ref. 24, p. 15)

$$\boldsymbol{R}(x_1 - x_2) = \exp\left[-(|x_2 - x_1|/d)^2\right], \qquad \text{where} \qquad d > 0 \tag{4}$$

The range parameter d indicates the distance at which the influence is $e^{-1} = 0.3679$ or approximately 37%. To support a multivariate correlation function, a univariate correlation function is used for each of the p input dimensions[6,25] rather than using the Euclidean norm of the space (Refs. 1 and 24, p. 5). A product correlation rule is used for mathematical convenience:

$$\boldsymbol{R}(\boldsymbol{X}_1, \boldsymbol{X}_2) = \prod_{j=1}^{p} \boldsymbol{R}(|x_{2,j} - x_{1,j}|) \tag{5}$$

This formulation improves the flexibility of modeling the correlation of each input dimension at the expense of requiring the selection of additional model parameters (p parameters instead of just one).

Let the set of n known inputs to the computer model be $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\} \subset \Omega$, where Ω is the set of all possible inputs to the model that result in an output, that is, the domain of the computer model. The resulting outputs are $\boldsymbol{Y} = \{\boldsymbol{y}(x_1), \boldsymbol{y}(x_2), \ldots, \boldsymbol{y}(x_n)\}$. Given these sampled outputs of the computer model, consider a linear predictor of the output:

$$\hat{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{\lambda}^T(\boldsymbol{x})\boldsymbol{Y} \tag{6}$$

at any point $\boldsymbol{x} \in \Omega$. The kriging approach treats $\hat{\boldsymbol{y}}(\boldsymbol{x})$ as a random function and finds the best linear unbiased predictor $\boldsymbol{\lambda}^T(\boldsymbol{x})\boldsymbol{Y}$, which minimizes the mean square error (MSE) of the prediction,

$$\text{MSE}[\hat{\boldsymbol{y}}(\boldsymbol{x})] = E[\boldsymbol{\lambda}^T(\boldsymbol{x})\boldsymbol{Y} - \boldsymbol{y}(\boldsymbol{x})]^2 \tag{7}$$

subject to the unbiasedness constraint,

$$E[\boldsymbol{\lambda}^T(\boldsymbol{x})\boldsymbol{Y} - \boldsymbol{y}(\boldsymbol{x})] = 0 \tag{8}$$

The general form of kriging, universal kriging, is defined with a set of regression functions

$$\boldsymbol{f}(\boldsymbol{x}) = \{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x})\}^T \tag{9}$$

A second type of kriging, ordinary kriging, is a special case of universal kriging, where

$$\boldsymbol{f}(\boldsymbol{x}) = \{1\} \tag{10}$$

Ordinary kriging is the most commonly used form of kriging employed to approximate computer models.[13,14,23] This work compares the accuracy of the kriging models created using the simple regressors of Eq. (10) to the more complex regressors of Eq. (9).

A vector $\boldsymbol{F}$ is constructed by evaluating $\boldsymbol{f}(\boldsymbol{x})$ at each of the n known observations:

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{f}(x_1) \\ \boldsymbol{f}(x_2) \\ \vdots \\ \boldsymbol{f}(x_n) \end{bmatrix} \tag{11}$$

The next definition needed is for the correlation matrix $\boldsymbol{R}$, which is composed of spatial correlation functions evaluated at each possible combination of the known points,

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}(x_1, x_1) & \boldsymbol{R}(x_1, x_2) & \cdots & \boldsymbol{R}(x_1, x_n) \\ \boldsymbol{R}(x_2, x_1) & \boldsymbol{R}(x_2, x_2) & \cdots & \boldsymbol{R}(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{R}(x_n, x_1) & \boldsymbol{R}(x_n, x_2) & \cdots & \boldsymbol{R}(x_n, x_n) \end{bmatrix} \tag{12}$$

This matrix $\boldsymbol{R}$ is a positive semidefinite matrix because the SCF defining each element is positive semidefinite. It is also symmetric because $\boldsymbol{R}(x_i, x_j) = \boldsymbol{R}(x_j, x_i)$, and the diagonal consists of all ones because $\boldsymbol{R}(x_i, x_i) = 1$.

The last definition needed is a vector to represent the correlation between an unknown point $\boldsymbol{x} \in \Omega$ and the n known sample points:

$$\boldsymbol{r}(\boldsymbol{x}) = \{\boldsymbol{R}(\boldsymbol{x}, x_1), \boldsymbol{R}(\boldsymbol{x}, x_2), \ldots, \boldsymbol{R}(\boldsymbol{x}, x_n)\}^T \tag{13}$$

If $\boldsymbol{\lambda}(\boldsymbol{x})$ solves the minimization problem of Eq. (7) subject to the unbiasedness constraint of Eq. (8), then $\boldsymbol{\lambda}^T(\boldsymbol{x})\boldsymbol{Y}$ is called the best linear unbiased predictor (BLUP) of $\hat{\boldsymbol{y}}(\boldsymbol{x})$. By solving for $\boldsymbol{\lambda}(\boldsymbol{x})$ and substituting into Eq. (6), the BLUP of $\hat{\boldsymbol{y}}(\boldsymbol{x})$ is given by

$$\hat{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{f}^T(\boldsymbol{x})\hat{\boldsymbol{\beta}} + \boldsymbol{r}^T(\boldsymbol{x})\boldsymbol{R}^{-1}(\boldsymbol{Y} - \boldsymbol{F}\hat{\boldsymbol{\beta}}) \tag{14}$$

where the least-squares estimate of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^T\boldsymbol{R}^{-1}\boldsymbol{Y} \tag{15}$$

and the MSE or variance of the estimate $\hat{\boldsymbol{y}}(\boldsymbol{x})$ is as follows:

$$\text{MSE}[\hat{\boldsymbol{y}}(\boldsymbol{x})] = \sigma^2 \left\{ 1 - [\boldsymbol{f}^T(\boldsymbol{x}) \quad \boldsymbol{r}^T(\boldsymbol{x})] \begin{bmatrix} \boldsymbol{0} & \boldsymbol{F}^T \\ \boldsymbol{F} & \boldsymbol{R} \end{bmatrix} \begin{bmatrix} \boldsymbol{f}(\boldsymbol{x}) \\ \boldsymbol{r}(\boldsymbol{x}) \end{bmatrix} \right\} \tag{16}$$

The first component of Eq. (14) is the generalized least-squares estimate of a point $\boldsymbol{x} \in \Omega$ given the correlation matrix $\boldsymbol{R}$; meanwhile, the second component pulls the generalized least-squares estimate through the observed data points. The estimates at the observations are returned exactly; the MSE at these points is zero because there is no uncertainty in the results of a deterministic computer model. As an unobserved point, $\boldsymbol{x}$ moves away from the observations, the second component of Eq. (14) approaches zero, yielding the generalized least-squares estimate, and the uncertainty in the estimate approaches its maximum value, the process variance σ^2.

B. Parameter Estimation Methods

Parameter estimation is the process of selecting the regression function coefficients, process variance, and SCF parameters $\boldsymbol{\gamma} = \{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\}$ for the kriging model that shows the best possible predictive performance. In this work, two methods that require optimization to estimate the kriging model parameters, maximum likelihood estimation (MLE) and cross validation (CV), are compared based on two criteria: 1) accuracy of the resulting kriging model as determined by using a large validation set of data for each test problem and calculating the rms error for the validation data, and 2) accuracy of the selected model parameters as determined by comparing the values of the correlation function parameters found with the two methods to those found by minimizing the rms error (RMSE) of the validation data.

One of the core assumptions of MLE is that the observations come from a Gaussian process; unfortunately, this is seldom the case with observations of computer models used in engineering design.[26] In this work, the residuals used for MLE are tested for the normality (Gaussian shape) using the modified Shapiro–Wilk's test[27–29] to determine if the actual distribution of the residuals has any impact on the ability of MLE to select the best kriging model parameters. The Shapiro–Wilk's test for normality provides a p-value (a probability) that the given samples (in this case the residuals) came from a normal or Gaussian distribution.

1. Fitting Kriging Models Using MLE

The design and analysis of computer experiments[13] primarily uses the statistics-based method of MLE as an objective estimator of the best kriging model parameters γ that are most consistent with the observed data.[6,30,31] MLE assumes the residuals have a known probability distribution shape, which in most cases is the Gaussian probability distribution. The logarithm of the multivariate Gaussian likelihood function is

$$\ell[\gamma|\boldsymbol{Y}] = -(n/2)\,\ell_n[2\pi\sigma^2] - \tfrac{1}{2}\,\ell_n[|\boldsymbol{R}|] - (1/2\sigma^2)(\boldsymbol{Y}-\boldsymbol{F}\boldsymbol{\beta})^T\boldsymbol{R}^{-1}(\boldsymbol{Y}-\boldsymbol{F}\boldsymbol{\beta}) \tag{17}$$

By taking the derivative of this log-likelihood equation with respect to $\boldsymbol{\beta}$ and σ^2 and solving for zero, the closed-form solution for the optimal value of $\boldsymbol{\beta}$ is found to match Eq. (15), and σ^2 is

$$\hat{\sigma}^2 = (1/n)(\boldsymbol{Y}-\boldsymbol{F}\hat{\boldsymbol{\beta}})^T\boldsymbol{R}^{-1}(\boldsymbol{Y}-\boldsymbol{F}\hat{\boldsymbol{\beta}}) \tag{18}$$

A closed-form solution does not exist for the optimal parameters of most common SCFs, requiring numerical optimization. To reduce the number of model parameters determined with the numerical optimization, the profile log-likelihood is optimized. The profile log-likelihood substitutes[32] the known optimal values of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ from Eqs. (15) and (18) back into Eq. (17) and optimizes for the unknown correlation parameters $\boldsymbol{\theta}$.

Three computational problems often exist when using MLE to select model parameters for kriging models: 1) multimodality of the log-likelihood function, 2) long ridges in the log-likelihood function, and 3) ill-conditioned correlation matrices. The first problem was initially identified by Warnes and Ripley[33] and Ripley (Ref. 24, pp. 15–21) while using a spherical correlation function.[30] They determined the problem resulted from their test problem having both short- and long-range correlations. When a second-order trend function was used to model the long-range effects, the multimodality was no longer observed. Mardia and Watkins[32] determined multimodality was caused by discontinuities in the second derivative of the SCF (a property of the spherical SCF) and by small numbers of observations used in the likelihood function. Stein (Ref. 21, p. 173) states that "when using the Matérn model, I am unaware of any examples of likelihood function with more than one local maximum." The Matérn model is a general form of SCF of which the Gaussian and exponential are special cases. The second issue, long ridges in the likelihood function, was acknowledged by Stein[21] and shown by Warnes and Ripley.[33] A long ridge of nearly constant and optimal values can lead to numerical difficulties for gradient-based and simplex optimization algorithms.

The last issue of using MLE is ill-conditioned correlation matrices. The correlation number of a matrix is defined as the ratio of the absolute values of the largest and smallest eigenvalues of the correlation matrix.[34] It is a worst-case loss in precision when inverting the correlation matrix. A matrix is considered ill-conditioned if the logarithm of the condition number is greater than the computer's precision (15.9546 for a Pentium M computer running Windows XP). A large condition number will result in a matrix that may have significant numerical inaccuracies when inverted. A study on the factors affecting the condition number was performed by Davis and Morris.[35] They determined that larger correlation ranges and closer located observations tend to result in larger condition numbers. The same result was found during the investigations of this work.

The issue of ill-conditioned matrices was noticed by Booker et al.[9] and Sasena[36] in the context of optimization of a deterministic computer model using kriging models. As their global optimization algorithms converged, observations began to cluster around the optimal value, resulting in rows and columns of the correlation matrix that were almost identical. Booker et al.[9] improved the condition number by adding a small value (10^{-6}) to the diagonal elements of the correlation matrix, and Sasena[36] used a small nugget effect (e.g., 10^{-12}). These techniques are very similar and result in a kriging model that no longer interpolates the observations. More recently, Booker[37] introduced a second Gaussian process to model the short-range variability present in the local vicinity of the optimal point, allowing the kriging model to still interpolate all of the observations. A fixed set of evenly distributed observations is used in this work (see Sec. III) to avoid ill-conditioned correlation matrices.

2. Fitting Kriging Models Using CV

An alternative to MLE for parameter estimation is CV.[19] Currin et al.[6] indicated the problem of finding the best parameters for a predictive kriging model "suggests cross validation." Currin et al.[38] provided three definitions of cross validation: 1) "leave-one-out" predictive density,[39] 2) leave-one-out squared bias, and 3) MLE as a form of cross validation. In this work, leave-one-out squared bias cross validation is used because only the expected value of the response is of interest[38] and not the probability distribution of the kriging model output.[40,41]

Cross validation of a kriging model is determined by holding all of the model parameters $\gamma = \{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}\}$ constant while creating n kriging models using each subset of the remaining $n-1$ points and calculating the error at each omitted location in turn. Currin et al.[38] and Mitchell and Morris[42] provided a computationally efficient formula for the leave-one-out CV error of prediction at each deleted site for a constant trend function kriging model as

$$e_i = q_i(g_i - \beta w_i) \tag{19}$$

where

$$g = \boldsymbol{R}^{-1}\boldsymbol{Y}, \qquad w = \boldsymbol{R}^{-1}\boldsymbol{F}$$

and q is the inverse of the diagonal of $\boldsymbol{R}^{-1}$. If the elements of q are placed on the diagonal of a matrix $\boldsymbol{Q}$, a more general form of Eq. (19), allowing the cross validation of a more complex trend function, can be restated as

$$e = \boldsymbol{Q}(g - w\beta) \tag{20}$$

By minimizing the average squared bias,

$$\Phi = e^T e/n \tag{21}$$

the estimated values of the trend function coefficients are

$$\hat{\boldsymbol{\beta}} = (w^T\boldsymbol{Q}^2 w)^{-1} w^T\boldsymbol{Q}^2 g \tag{22}$$

where $\boldsymbol{Q}^2$ is $\boldsymbol{Q}^T\boldsymbol{Q}$, or $\boldsymbol{Q}$ with each element squared. This is very similar to results Currin et al.[37] found when solving for the constant trend function,

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} q_1^2 w_i g_i \Big/ \sum_{i=1}^{n} q_i^2 w_i^2$$

Equation (21) is minimized over the correlation function parameter space by substituting the optimal $\hat{\boldsymbol{\beta}}$ from Eq. (22), which is also a function of the correlation function parameters. This method is similar to the profile log-likelihood method to reduce the number of variables being optimized.

An advantage to CV over MLE is that it does not assume any probability distribution shape for the residuals. This can result in a better estimate of the optimal model parameters than MLE in cases where the probability distribution of the output is not Gaussian.[43] Currin et al.[6] state, "Of the various kinds of cross validation we have tried, maximum likelihood seems the most reliable." Sacks et al.[13] felt that "our experience is that even crude MLEs can lead to useful predictions and quantification of uncertainty." Wahba[44] and Stein[45] compared CV to MLE as methods for choosing model parameters. They both considered the situation when the correct stochastic process was used, that is, the observations followed the assumed distribution. Wahba[44] found that at relatively small numbers of observations CV selected model parameters better than MLE based on a predictive mean-square-error criterion for relatively smooth observations, and CV always performed better for larger numbers of observations. Stein[45] looked at the resulting uncertainty in the model parameter estimates and found "both estimates are asymptotically

normal with the GCV [CV] estimates having twice the asymptotic variance of the MML [MLE] estimate." It is unknown if the performance of the two differs greatly as the observations depart from the assumed probability distribution, but Wahba[44] believes that CV is more robust in this case.

After selecting the best kriging model parameters, the resulting kriging model must be assessed to provide a measurement of its capability to reproduce the output of the computer model or simulation. In the next section, two methods, a cross validation and an information-based method, are presented to assess the quality of metamodels, permitting comparisons of accuracy between the different forms of kriging models and of the methods used to select the parameters for the kriging model forms.

C. Metamodel Assessment

A metamodel's quality can be assessed with two measurements: 1) accuracy when reproducing the observed data and 2) accuracy in predicting the original model at unobserved locations. The first measurement is important for regression models, but the measurement is meaningless for interpolating models like kriging because they reproduce the observations exactly for any values of the model parameters and is not considered in this work. The second measurement is most accurately assessed by measuring the error in a model's prediction at a large set of m additional observations. These errors can then be summarized with the RMSE or with the R^2_{actual} measurement, which is defined as

$$R^2_{\text{actual}} = 1 - \text{SSE}/SS_T$$

where SSE is the sum of the squared errors of prediction at the m additional observation sites and the total sum of the squares is

$$SS_T = Y^T Y - \left(\sum_{i=1}^{m} y_i\right)^2 \Big/ m \tag{23}$$

This method requires a large number ($m \gg n$) of potentially computationally expensive computations to obtain the additional observations of the computer model that are not used to fit the metamodel. Larger values of the R^2_{actual} indicate the metamodel is a better approximation of the original model, and the best values of R^2_{actual} approach one. As a more computationally efficient alternative, CV[2,46,47] or Akaike's information criterion (AIC) can be used.[48,49]

The two methods, CV and AIC, are compared in this work because they do not require the use of additional observations to assess the predictive capability of the model, which is particularly important when working with computationally expensive analyses. The most common measurement of the accuracy for predicting the original model is cross validation. This procedure is identical to that used during CV parameter estimation. Cross validation leaves out a fixed number of the current observations, frequently just one, from the metamodel and uses the remaining observations to estimate the value of the computer model at those points.[47] The second method, AIC, quantifies the accuracy (actually the uncertainty) of the metamodel's parameters. AIC is a function of the log-likelihood of the model parameters exactly as used for the MLE parameter estimation method. The CV and AIC measurements are compared to R^2_{actual} in Sec. IV for their ability to assess the quality of a kriging model and to provide a criterion for selecting between different forms of a kriging model.

In the context of generalized linear regression models, cross validation error assessment is often termed predictive error sum of squares (PRESS).[47,50] To calculate the PRESS statistic, each observation is omitted in turn while the remaining sites are used to fit the regression model. This new model is then used to predict the withheld observation and its resulting error or residual. The PRESS statistic is defined as the sum of squares of the leave-one-out cross validation errors and can be used in place of the sum of square residuals to compute an R^2 for prediction,

$$R^2_{\text{prediction}} = 1 - \text{PRESS}/SS_T \tag{24}$$

where the total sum of the squares is calculated with Eq. (23) with the n observations. This statistic gives an indication of the predictive capability of the metamodel that is scale independent and can be used for interpolating models such as kriging models.

The PRESS statistic can be improved by adding more degrees of freedom (i.e, more parameters) to the metamodel being assessed, which can lead to overfitting the data. The $R^2_{\text{prediction}}$ can be adjusted accordingly to reduce this tendency to overfit the data as

$$R^2_{\text{prediction adjusted}} = 1 - [(n-1)/(n-q)]\left(1 - R^2_{\text{prediction}}\right) \tag{25}$$

where n is the number of observations and q is the dimension of $\boldsymbol{\beta}+$ the dimension of $\boldsymbol{\theta}+1$ for the variance.

AIC[48] can be used to evaluate the quality of a model based upon the value of the log-likelihood function from Eq. (17) and the number of parameters q used to fit the model. It is defined as

$$\text{AIC} = -2\ell[\boldsymbol{\gamma}|\boldsymbol{y}] + 2(q) \tag{26}$$

When the number of observations n is small ($n/q < 40$), as is most often the case with metamodels, the AIC tends to be too small. This value has been corrected[51] with

$$\text{AIC}_c = \text{AIC} + 2q(q+1)/(n-q-1) \tag{27}$$

where n is the number of observations used to estimate the model parameters. From a given set of potential models, the one that results in the smallest value of AIC_c should be selected. The CV and AIC measurements are compared to R^2_{actual} in Sec. IV for their capability to 1) assess the quality of and 2) provide a criterion for selecting between different forms of a kriging model.

III. Experimental Setup

To demonstrate the different parameter estimation and model assessment methods, six test problems are used: one one-dimensional (see Sec. III.A), four two-dimensional (see Sec. III.B), and one five-dimensional (see Sec. III.C) problems. All of these test problems were implemented using Mathematica 5.0 running on Windows XP sp1. Only the one- and five-dimensional problems are relatively computationally expensive, taking 20 s to execute. The two-dimensional problems are common test problems from the literature. The experimental procedure used is outlined in Sec. III.D, and the results are analyzed in Sec. IV.

A. One-Dimensional Problem

The one-dimensional test problem calculates the output temperature of a chemical reaction.[52] The ratio of oxidant to the fuel being burned is increased from no oxidant to an excess of oxidant. In this process, the reaction increases temperature to a maximum and then decreases as excess oxidant is added (Fig. 1). The sample points, indicated by dots in Fig. 1, are evenly spaced at 0.1 increments from 0 to 1 for a total of 11 points. A set of 500 evenly spaced validation points is used to evaluate the RMSE for each kriging model.

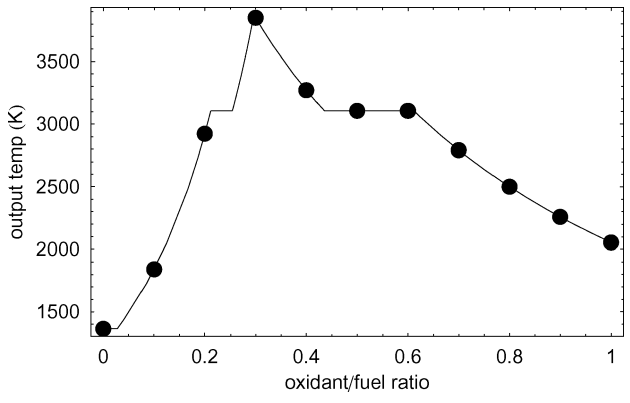


Fig. 1 Output temperature vs oxidant fuel ratio.

B. Two-Dimensional Problems

The first two-dimensional test problem is highly nonlinear in one dimension and is linear in the second dimension. This was introduced by Osio and Amon[7] and used by Jin et al.[53] as a test problem for sequential experimentation. The function is

$$y(\boldsymbol{X}) = \cos[6(x_1 - 0.5)] + 3.1(|x_1 - 0.7|) + 2(x_1 - 0.5) + 7\sin[1/(|x_1 - 0.5| + 0.31)] + 0.5x_2 \qquad \boldsymbol{X} \in [0, 1]^2 \quad (28)$$

The second two-dimensional test problem is a "mystery" multimodal function in two dimensions that comes from Sasena[36]

$$y(\boldsymbol{X}) = 2 + 0.01\left(x_2 - x_1^2\right)^2 - (1 - x_1) + 2(2 - x_2)^2 + 7\sin(0.5x_1)\sin(0.7x_1x_2) \qquad \boldsymbol{X} \in [0, 5]^2 \quad (29)$$

The third two-dimensional test problem is the Branin test function,[53] which has less oscillation over its domain; it is defined as

$$y(\boldsymbol{X}) = \left[x_2 - (5.1/4\pi^2)x_1^2 + (5/\pi)x_1 - 6\right]^2 + 10(1 - 1/8\pi)\cos(x_1) + 10 \qquad x_1 \in [-5, 10], \qquad x_2 \in [0, 15] \quad (30)$$

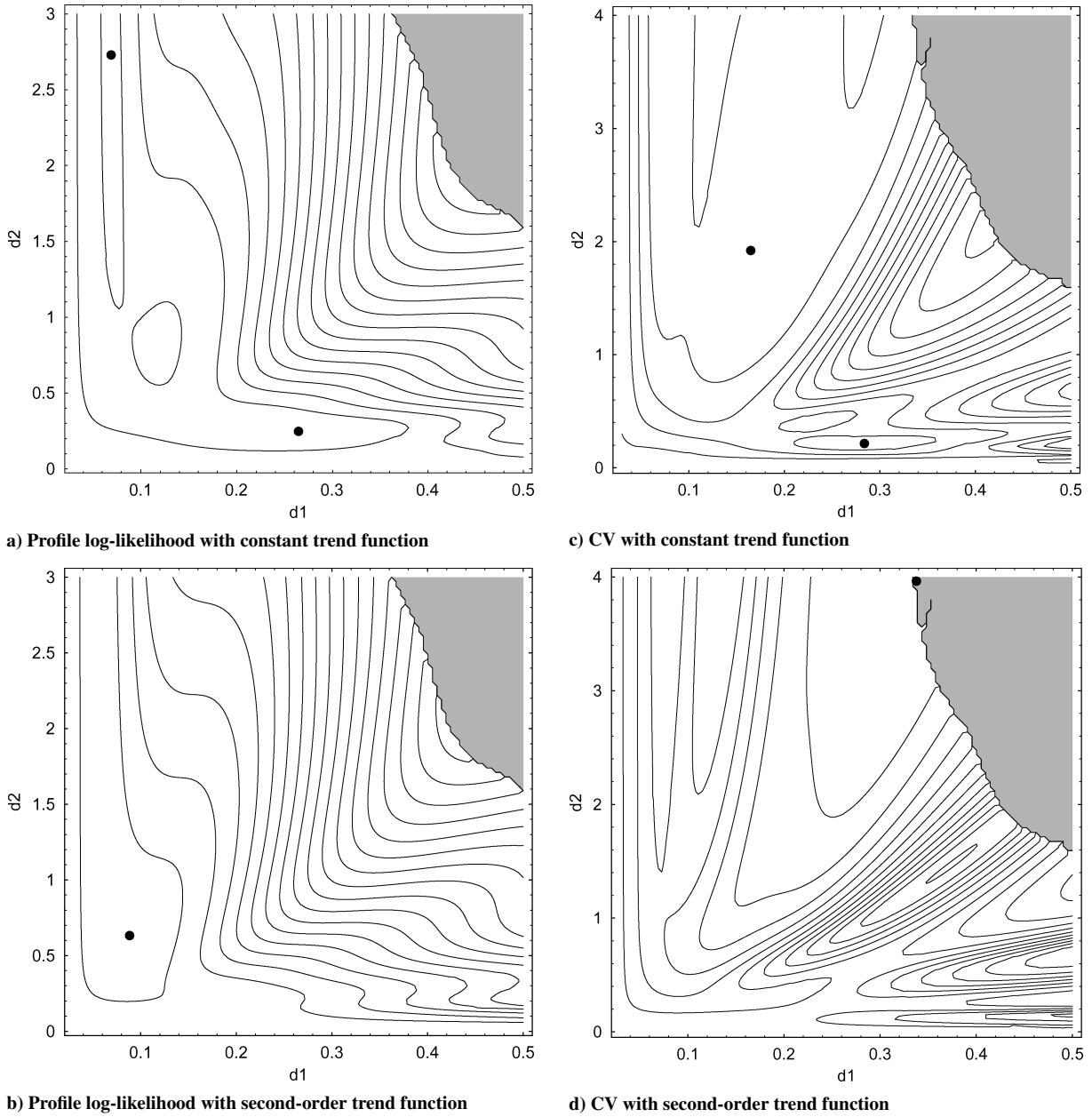
The final two-dimensional problem is the six-hump camelback function,[36] which is defined as

$$y(\boldsymbol{X}) = \left(4 - 2.1x_1^2 + x_1^4/3\right)x_1^2 + x_1x_2 + \left(-4 + 4x_2^2\right)x_2^2 \qquad x_1 \in [-2, 2], \qquad x_2 \in [-1, 1] \quad (31)$$

For all of these two-dimensional problems, the same set of 21 sample points is used to fit each kriging model.[42] These points were generated using a Latin Hypercube,[54] which is similar to the design used by Jones et al.[55] The domain of each example is mapped to $[0, 1]^2$. A set of 900 validation points placed on a 30×30 grid is used to evaluate the RMSE for the metamodels.

C. Five-Dimensional Problem

The last test problem is an analysis of a gas generation system.[52] It has five inputs that describe the geometry of the system and its operating conditions, and it returns the volume of gas generated by the system. The inputs to the problem are scaled and nondimensionalized to map the feasible domain to $[0, 1]^5$. For this problem, a 40-point Latin Hypercube is used for sampling, and a set of 500 uniformly random validation points is used to evaluate the RMSE for each kriging model.



a) Profile log-likelihood with constant trend function

c) CV with constant trend function

b) Profile log-likelihood with second-order trend function

d) CV with second-order trend function

Fig. 2 Contour plots of MLE and CV parameter estimation methods vs d1 and d2 for six-hump camelback function.

D. Experimental Procedure

Six steps are used to compare the parameter estimation and metamodel assessment methods for each test problem:

1) Evaluate each computer model for the six test problems at the selected sample points.

2) Fit an ordinary kriging model using both MLE and CV methods to select the range parameters [see Eq. (4)] for the Gaussian SCF for each input dimension. The trend function coefficient values are chosen using Eqs. (15) and (22) for MLE and CV, respectively.

3) Assess each of the resulting kriging models using the $R^2_{\text{prediction}}$, $R^2_{\text{prediction adjusted}}$, and AIC_c. Use Eq. (20) to calculate the leave-one-out CV errors for all models. Compare these results with the R^2_{actual} found using the validation data. Use Eq. (18) as suggested by Currin et al.[38] to estimate the process variance σ^2 to calculate AIC_c.

4) Perform the Shapiro–Wilk's test for normality (Sec. II.B) on the observations using the results of the MLE method to remove the bias in the observations; this test is not used on the CV-fitted models because they do not require a specific probability distribution of the modeled residuals.

5) Determine the "best" kriging model parameters for a set of observations and model form used in step 2 by using the validation data to find the kriging model parameters that minimize the RMSE (maximize R^2_{actual}). The resulting estimated models should have parameter values that are close to the actual values found in this step.

6) Repeat steps 2–5 for other sets of trend functions as listed in Table 1.

Results from these experiments are discussed next.

IV. Analysis of Results

The analysis of results is divided into three sections. The first section describes some of the computational issues encountered while determining the kriging model parameters. The last two sections compare the trend function (see Sec. IV.B) and metamodel assessment methods (see Sec. IV.C).

A. Computational Issues

The most difficult computational issue with using kriging models is the optimization process required to determine the optimal model parameters. During optimization, correlation range parameter values were constrained to be 1) larger than half the minimum distance between observations (e.g., 0.024 for the two-dimensional test problems), 2) smaller than five times the largest distance between observations (e.g., five for all test problems), and 3) the values must result in a well-conditioned correlation matrix (see Sec. II.B.1). The first two limits were chosen because there is no evidence available from the set of observations to identify correlation ranges outside of these limits.

Contour plots of the MLE and CV parameter estimation methods for constant and second-order trend functions are shown in Fig. 2 as an example of the difficulties that can arise during kriging model parameter estimation. The gray regions in all of the plots of Fig. 2 indicate values for the correlation function parameters, d1 and d2, that yield an ill-conditioned correlation matrix. The ill-conditioned correlation matrix constraint was similar in all of the two-dimensional test problems. The multiple local optima were seen in the six-hump camelback and the mystery function test problems, but they did not appear to occur in the other test problems.

The three issues of using MLE discussed in Sec. II.B.1 of long ridges, multiple local optima, and ill-conditioned matrices can be seen in the contour plot of Fig. 2a, which shows the profile log-likelihood plot with a constant trend function. The point in the upper left-hand portion of the plot indicates the global maximum, and the lower central point is that found by a Broyden, Fletcher, Goldfarb, and Shanno (BFGS) quasi-Newton method in Mathematica. To resolve the issue of multiple local optima, a Nelder–Mead simplex algorithm[56] and a differential evolution algorithm[57] were used for comparison of results. The algorithms used are available within Mathematica 5.0. In the remaining results, the optimal values presented were the best results of the three algorithms with no one algorithm always providing the best optimal values. The profile log-likelihood plot in Fig. 2b, using a second-order trend function, appears to have a single local maximum, affirming the results observed by Warnes and Ripley[33] of an improved trend function often resulting in an easier to optimize profile log-likelihood function.

The cross validation parameter estimation method using the constant and second-order trend functions for the six-hump camelback test problem is shown in Figs. 2c and 2d. The CV method does not appear to be as dominated by ridges (valleys) consisting of nearly optimal values as the MLE method, but the resulting surfaces still appear to have multiple local optima. The source of the "crack" in the ill-conditioned correlation matrix constraint region is not known. The global minimum in Fig. 2c occurred at the left central point. The lower point was a local minimum found using the gradient-based, BFGS quasi-Newton method. All three optimization methods found very similar optimal points as shown in Fig. 2d, which were found to lie on the ill-conditioned correlation matrix constraint.

B. Parameter Estimation and Trend Function Importance

The first aspect investigated in this work is the most accurate method to estimate the kriging model parameters. To evaluate the most accurate method, 1) the validation errors of the resulting model and 2) the accuracy of the correlation function parameters selected by each method are compared. The abbreviations used for the trend functions of the test problems are summarized in Table 1.

The R^2_{actual} values for the validation data for each trend function are shown in Fig. 3. The "best" method has the largest R^2_{actual} possible for a given form of the kriging model as determined in step 5 of the experimental procedure. This is not a truly fair comparison because

Table 1 Abbreviations for trend functions

Abbreviation	One-dimensional trend functions	Two-dimensional trend functions	Five-dimensional trend functions
0	1	1	1
1	$\{1, x\}$	——	$\{1, x_1, x_2, x_3, x_4, x_5\}$
01	——	$\{1, x_2\}$	——
10	——	$\{1, x_1\}$	——
11	——	$\{1, x_1, x_2\}$	——
111	——	$\{1, x_1, x_2, x_1x_2\}$	——
12	——	$\{1, x_1, x_2, x_2^2\}$	——
21	——	$\{1, x_1, x_2, x_1^2\}$	$\{1, x_1, \ldots, x_5, x_1^2, \ldots, x_5^2, x_1x_2, \ldots, x_4x_5\}$
2	$\{1, x, x^2\}$	——	$\{1, x_1, \ldots, x_5, x_1^2, \ldots, x_5^2\}$
22	——	$\{1, x_1, x_2, x_1^2, x_2^2\}$	——
221	——	$\{1, x_1, x_2, x_1^2, x_2^2, x_1x_2\}$	——
3	$\{1, x, x^2, x^3\}$	——	——

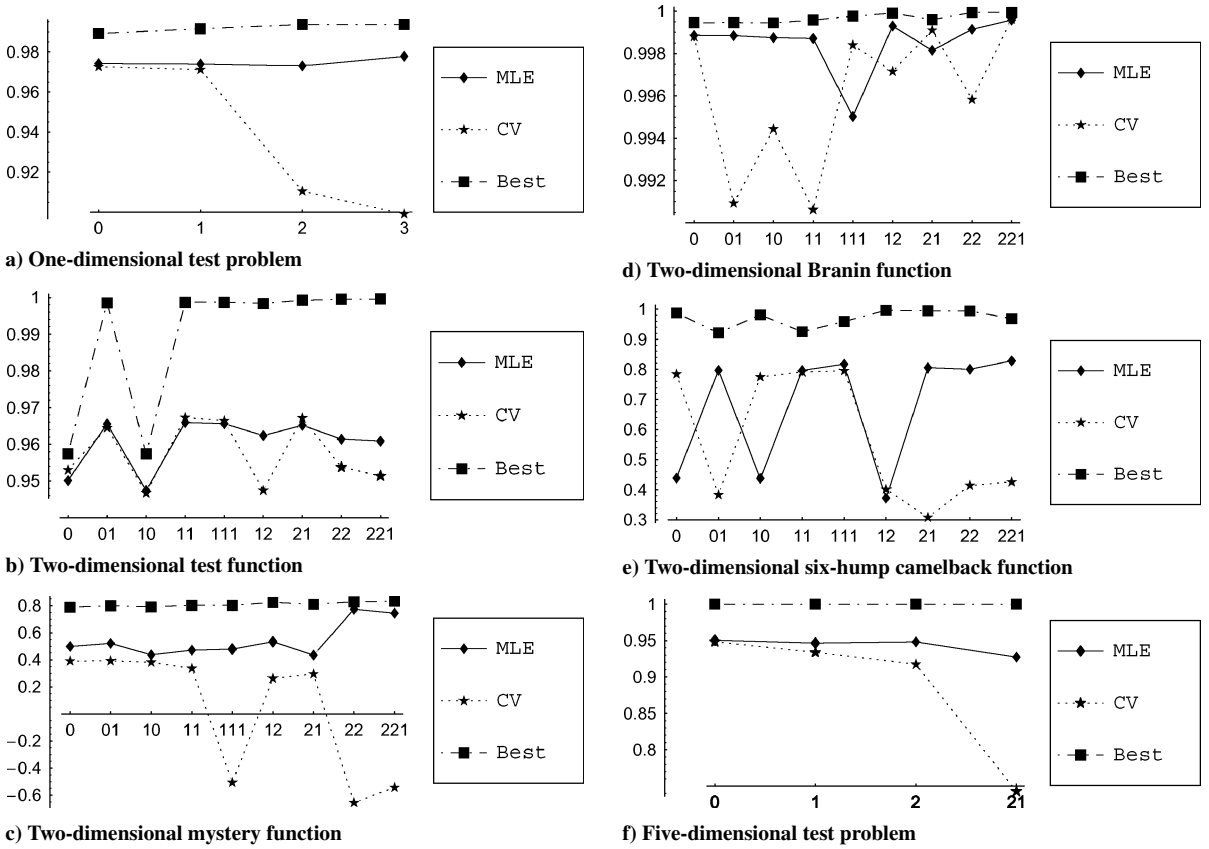


Fig. 3 Plots of R^2_{actual} vs trend function for the six test problems.

the "best" models were determined from a much richer set of observations (many more observations of the approximated model). The MLE and CV methods were given only a small set of observations to estimate the model parameters.

The six test problems provide examples of situations when sufficient observations are available to create a good metamodel (one-dimensional test problem, first two-dimensional test function, two-dimensional Branin function, and five-dimensional test problem) and when insufficient observations are available to model the original deterministic computer model (two-dimensional mystery function and two-dimensional six-hump camelback function). In the case of the two-dimensional mystery function, none of the forms of the kriging model that we investigated were capable of producing a good metamodel with the 21 interpolation points. The two-dimensional six-hump camelback problem can be reproduced with the 21 interpolation points, but the points did not contain the information needed to properly select the correct correlation distances. The results from the two examples of insufficient observations are included for completeness and to allow the reader to see the results from poor metamodels, but conclusions about the accuracy of parameter estimation methods and the importance of the trend function from these examples should be considered carefully.

In general, as more parameters are used to define the kriging model, the "best" R^2_{actual} is improved. The same also appears to be true for models fit with the MLE method. Models fitted with the CV method tend to perform worse as more parameters are added. Given the results of Fig. 3, the ability of kriging models to approximate deterministic computer models is improved by using a more complex trend function if using the MLE parameter estimation method. The second result drawn from Fig. 3 is MLE consistently performs better then CV as a kriging model parameter estimation method. This supports the reported experience of many other researchers (see Sec. II.B.2).

The correlation parameters chosen by each kriging model parameter estimation method are presented in Fig. 4. To compare the results for all of the test problems, the geometric means of the correlation distances are calculated. The geometric means of the "best" correlation distances tend to become smaller as more parameters are added to the trend function. The results found using MLE tend to follow the same trend. These results support the conclusion from Warnes and Ripley[33] and Ripley[24] that an improved trend function will tend to model the long-range effects in the model while the correlations will model the short-range effects. In general, CV selects larger correlation distances than MLE, and both methods tend to overestimate the best correlation distances. The overestimation of the correlation distances by CV tends to increase with more parameters in the trend function. This overestimation of the correlation parameters by CV can be attributed to including the edge points in the leave-one-out cross validation, which tends to increase the range of the correlation in order to more accurately extrapolate the edge points. Future studies should examine the impact of excluding the edge points from the RMSE calculation during CV to improve its effectiveness as a kriging model parameter estimation method.

As stated in Sec. II.B, the MLE method may perform poorly if the residuals from the trend function do not have a Gaussian probability distribution. The Shapiro–Wilk's test of normality provides a p-value (probability) that the sampled distribution is Gaussian. A p-value over 0.5 indicates moderate evidence that the data comes from a Gaussian distribution, whereas a p-value over 0.75 indicates strong evidence.[58] Figure 5 plots standard error vs p-value where standard error is defined here as

$$\text{standard error} = \frac{\text{MLE RMSE} - \text{Best RMSE}}{\text{Best RMSE}} \tag{32}$$

The five-dimensional problem is not included because of scaling issues; the p-values were very small (<0.029), and the standard errors large (>30.8). Figure 5 indicates that large p-values tend to yield smaller standard errors. Small p-values yield inconclusive results.

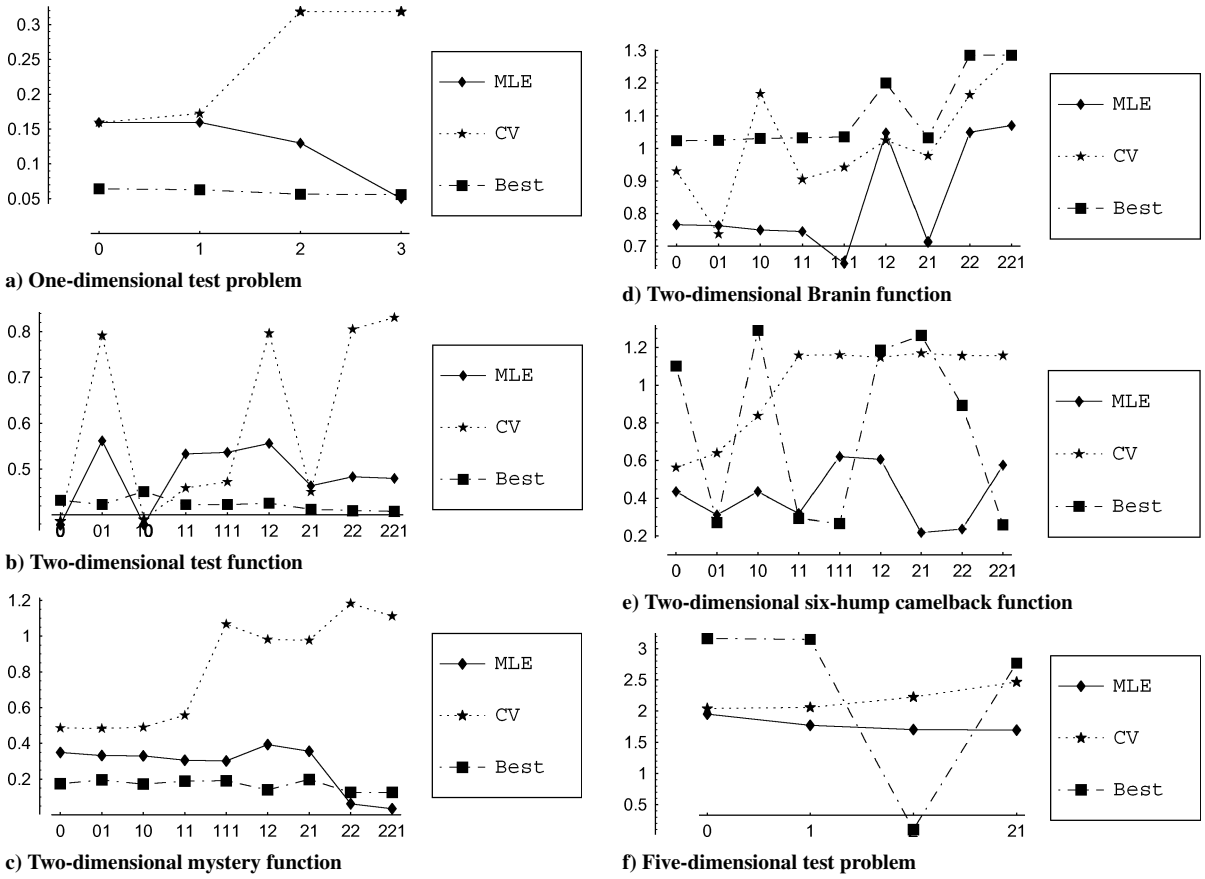


Fig. 4 Plots of geometric mean of the correlations vs trend function for the six test problems.

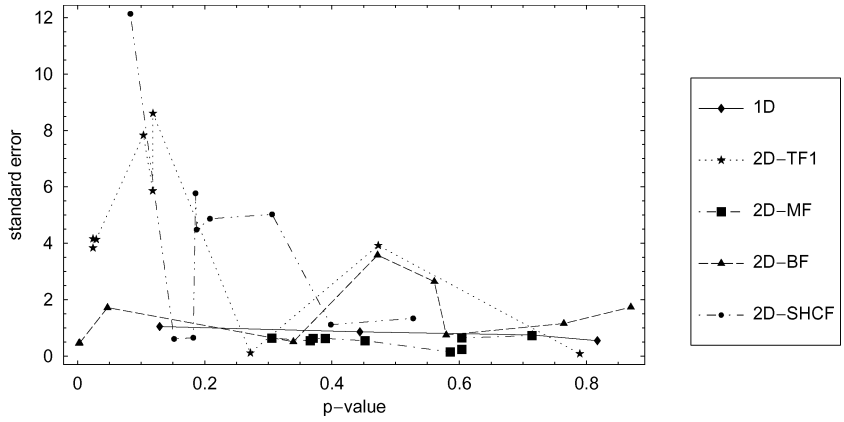


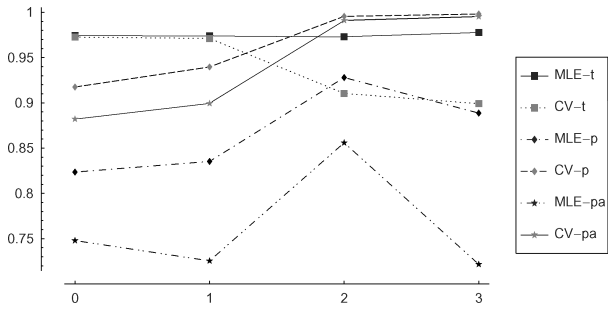
Fig. 5 Plot of standard error vs *p*-value of the Shapiro–Wilk's test for normality.

C. Metamodel Assessment

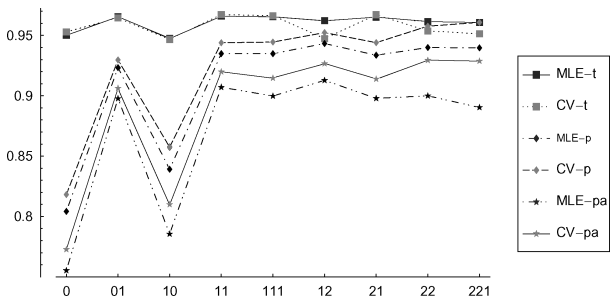
After the selection of the optimal model parameters for the observed data, it is important to quantify the accuracy of resulting kriging model. There are two decisions that need to be made: 1) is the kriging model a good approximation of the deterministic computer model? and 2) which kriging model is the best approximation of the deterministic computer model?

The $R^2_{\text{prediction}}$ and the $R^2_{\text{prediction adjusted}}$ measurements are compared to the R^2_{actual} for all six test problems in Fig. 6. The abbreviations in the legend identify first the parameter estimation method used and second the measurement method with -t (true) for the R^2_{actual} as determined using the validation data, -p for the $R^2_{\text{prediction}}$, and -pa for the $R^2_{\text{prediction adjusted}}$. The x axis in all of the plots indicates the trend function used (see Table 1).

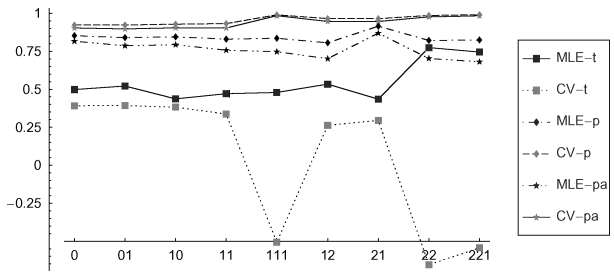
The CV parameter estimation method results in a better $R^2_{\text{prediction}}$ and $R^2_{\text{prediction adjusted}}$ measurement than MLE for all kriging models created. This is expected because the CV method finds the model parameters that minimize the leave-one-out cross validation error. The results of the test problems shown in Figs. 6a, 6b, 6d, and 6f indicate that a kriging model can provide a good approximation of the deterministic computer model ($R^2_{\text{actual}} > 0.95$). For these four test problems, $R^2_{\text{prediction}}$ appears to more closely approximate R^2_{actual} than $R^2_{\text{prediction adjusted}}$, underestimating the R^2_{actual} in most cases. It is disturbing to observe such a high $R^2_{\text{prediction}}$ and low R^2_{actual} for CV fitted models in Figs. 6c and 6e, though the values of $R^2_{\text{prediction}}$ for MLE fitted models appear to more closely estimate the R^2_{actual} for the resulting kriging models. One conclusion to draw from the results shown in Fig. 6 is if $R^2_{\text{prediction}} > 0.9$, then it is expected that



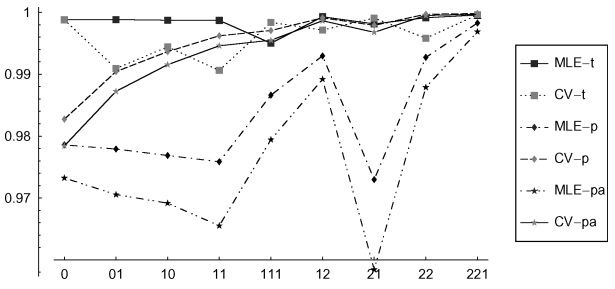
a) One-dimensional test problem



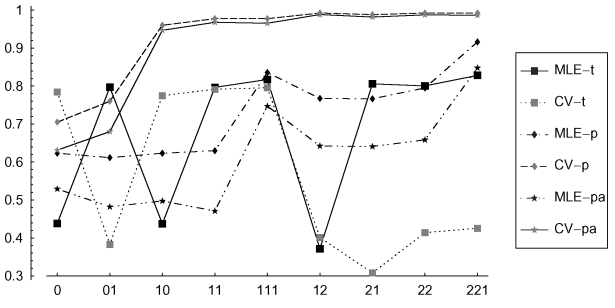
b) Two-dimensional test function



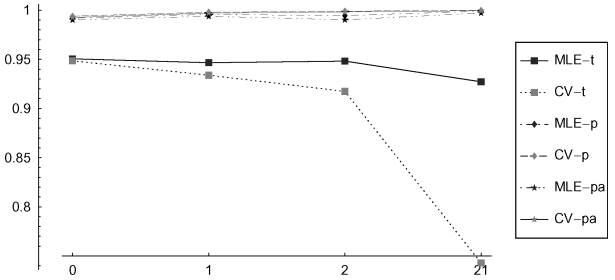
c) Two-dimensional mystery function



d) Two-dimensional Branin function

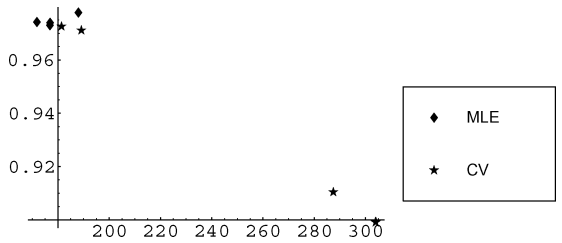


e) Two-dimensional six-hump camelback function

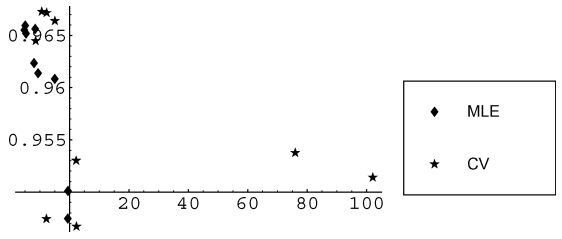


f) Five-dimensional test problem

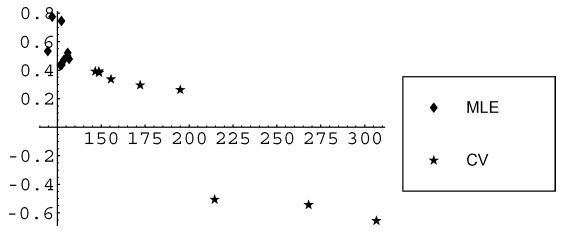
Fig. 6 Plots of all R^2 measurements vs trend function for the six test problems.



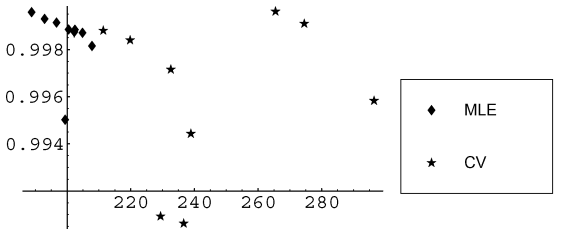
a) One-dimensional test problem



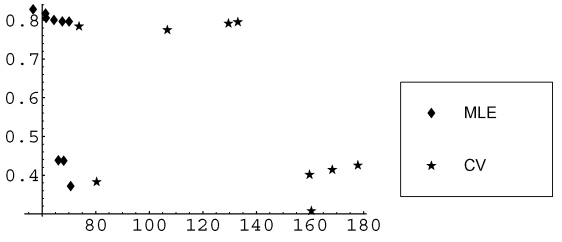
b) Two-dimensional test function



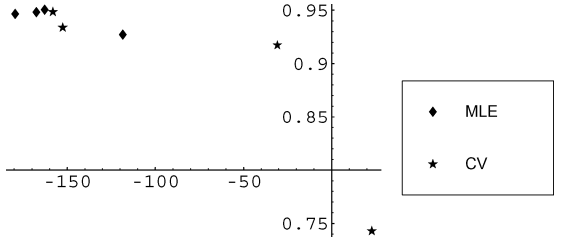
c) Two-dimensional mystery function



d) Two-dimensional Branin function



e) Two-dimensional six-hump camelback function



f) Five-dimensional test problem

Fig. 7 Plots of R^2_{actual} measurements vs AIC_C for the six test problems.

$R^2_{\text{actual}} > 0.9$ for kriging models estimated with MLE. The converse of this is not true as can be seen in Fig. 6a, where $R^2_{\text{prediction}} < 0.9$ and $R^2_{\text{actual}} > 0.95$.

The results shown in Fig. 6 indicate that neither $R^2_{\text{prediction}}$ nor $R^2_{\text{prediction adjusted}}$ appear to be a very precise metric. The results from Fig. 7 indicate that AIC_C is an effective criterion to distinguish between competing kriging model forms. The best (smallest) AIC_C measurements result from either the best model or nearly the best model. The poorest (largest) AIC_C values are the results of CV-fitted kriging models, where the correlation range has been overestimated. The results shown in Figs. 7c and 7e indicate that the AIC_C measurement is not effected by the overall quality of the kriging model. The negative R^2_{actual} in Fig. 7c results from the poor model parameters chosen by the CV method; a model consisting of the average of the observations would be a better predictor.

V. Conclusions

The use of kriging models to approximate deterministic computer models was investigated in this work. Specifically, three aspects on the predictive capability of the kriging model were examined: 1) model parameter estimation methods, 2) the importance of the trend function, and 3) metamodel error assessment techniques. Based upon the results presented in Sec. IV, the following conclusions can be drawn:

1) Maximum likelihood estimation (MLE) was the best method to select kriging model parameters even if the modeled observations do not have a Gaussian distribution. Cross validation (CV) has the potential of performing slightly better, especially for a constant trend function, but it also has the potential of performing much worse.

2) The corrected Akaike's information criterion (AIC_C) provided the best metamodel error assessment measurement to compare potential metamodels for a given set of observations, but it did not yield a means to estimate the actual model quality, namely, R^2_{actual}. The CV derived error assessment method $R^2_{\text{prediction}}$ appeared to provide a reasonable test for the metamodel's R^2_{actual}. It was observed that if the metamodel's $R^2_{\text{prediction}} > 0.90$, then the metamodel's $R^2_{\text{actual}} > 0.90$. The $R^2_{\text{prediction adjusted}}$ measurement was not useful.

3) A kriging model with a more complex trend function provided a better approximation of a deterministic computer model if the parameters were estimated using the MLE method. A more complex trend function also appeared to improve the shape of the profile log-likelihood function, making it easier to determine the global maximum using the MLE method. These results indicate that a trend function that is more complex than a constant value, as is typical in most current engineering applications, should be considered. Any other knowledge about the computer model being approximated, for example, if the output is a mass calculation and one of the input is a diameter then a second-order term for the diameter, should be incorporated into the trend function.

4) The widespread use of kriging models to approximate computer models will require more investigation into the best optimization algorithms to find the model parameters and improvement in the computational efficiency of the optimization algorithms. Both of the parameter estimation methods investigated (beyond the one-dimensional model) had multiple local optima, requiring the use of computationally expensive optimization methods. The parameter space is also plagued with long, steep ridges, further complicating the optimization.

There are still many unresolved issues that exist when using kriging models to approximate deterministic computer models. A few of them are listed here:

1) The results of the two accuracy measurements are dependent upon the observations used to estimate the kriging model parameters being a good representation of the original computer model being approximated. This "variable" is not quantified in this work, and the "noise" in the results makes it difficult to reach definitive conclusions.

2) The AIC_C measurement proved to be very useful in differentiating between potential kriging models, but it lacks the ability to draw absolute conclusions about the model. The scale of the AIC_C is primarily influenced by the process variance σ^2. Future studies should investigate normalizing the trend function residuals to a $\sigma^2 = 1$ and determining if better conclusions can be drawn from the AIC_C measurement for model accuracy.

3) There appears to be no accurate method to assess the RMSE of the resulting kriging model based on using cross validation of the set of observations used to fit the model. Future studies should investigate if there is a benefit to using a few observations, not used to fit the model, to calculate an RMSE value in a method similar to Meckesheimer et al.,[46] or if all observations should be included to provide the maximum amount of information to aid in parameter estimation.

Acknowledgment

The authors gratefully acknowledge support from Kam Ng of the Office of Naval Research, Code 333, Contract N00014-00-G-0058.

References

[1]Giunta, A., and Watson, L. T., "A Comparison of Approximation Modeling Technique: Polynomial Versus Interpolating Models," *7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Vol. 1, AIAA, Reston, VA, 1998, pp. 392–404.

[2]Simpson, T. W., Booker, A. J., Ghosh, S., Giunta, A., Koch, P. N., and Yang, R. J., "Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion," *Structural and Multidisciplinary Optimization*, Vol. 27, No. 5, 2004, pp. 302–313.

[3]Jin, R., Chen, W., and Simpson, T. W., "Comparative Studies of Metamodeling Techniques Under Multiple Modeling Criteria," *Journal of Structural and Multidisciplinary Optimization*, Vol. 23, No. 1, 2001, pp. 1–13.

[4]Batill, S. M., Renaud, J. E., and Gu, X., "Modeling and Simulation Uncertainty in Multidisciplinary Design Optimization," AIAA Paper 2000-4803, Sept. 2000.

[5]Barton, R. R., "Simulation Metamodels," *Proceedings of the 1998 Winter Simulation Conference*, edited by D. Medeiros, E. Watson, J. Carson, and M. Manivannan, Inst. of Electrical and Electronics Engineers, 1998, pp. 167–174.

[6]Currin, C., Mitchell, T. J., Morris, M. D., and Ylvisaker, D., "Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments," *Journal of the American Statistical Association*, Vol. 86, No. 416, 1991, pp. 953–963.

[7]Osio, I. G., and Amon, C. H., "An Engineering Design Methodology with Multistage Bayesian Surrogate and Optimal Sampling," *Research in Engineering Design*, Vol. 8, No. 4, 1996, pp. 189–206.

[8]Pacheco, J. E., Amon, C. H., and Finger, S., "Bayesian Surrogates Applied to Conceptual Stages of the Engineering Design Process," *Journal of Mechanical Design*, Vol. 125, No. 4, 2003, pp. 664–672.

[9]Booker, A. J., Dennis, J. E., Frank, P. D., Serafini, D. B., Torczon, V., and Trosset, M., "A Rigorous Framework for Optimization of Expensive Functions by Surrogates," *Structural Optimization*, Vol. 17, No. 1, 1999, pp. 1–13.

[10]Koch, P. N., Wujek, B. A., Golovidov, O., and Simpson, T. W., "Facilitating Probabilistic Multidisciplinary Design Optimization Using Kriging Approximation Models," AIAA Paper 2002-5415, Sept. 2002.

[11]Simpson, T. W., Maurey, T. M., Korte, J. J., and Mistree, F., "Kriging Metamodels for Global Approximation in Simulation-Based Multidisciplinary Design Optimization," *AIAA Journal*, Vol. 39, No. 12, 2001, pp. 2233–2241.

[12]Yang, R. J., Wang, N., Tho, C. H., Bobineau, J. P., and Wang, B. P., "Metamodeling Development for Vehicle Frontal Impact Simulation," American Society of Mechanical Engineers, ASME Design Engineering Technical Conf.—Design Automation Conf., DETC2001/DAC-21012, Sept. 2001.

[13]Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., "Design and Analysis of Computer Experiments," *Statistical Science*, Vol. 4, No. 4, 1989, pp. 409–435.

[14]Simpson, T. W., Peplinski, J., Koch, P. N., and Allen, J. K., "Metamodels for Computer-Based Engineering Design: Survey and Recommendations," *Engineering with Computers*, Vol. 17, No. 2, 2001, pp. 129–150.

[15]Matheron, G., "Principles of Geostatistics," *Economic Geology*, Vol. 58, 1963, pp. 1246–1266.

[16]Goldberger, A. S., "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, Vol. 57, No. 298, 1962, pp. 369–375.

[17]Journel, A. G., and Huijbregts, C. J., *Mining Geostatistics*, Academic Press, New York, 1978.

[18]Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Applied Geostatistics Series, Oxford Univ. Press, New York, 1997.

[19]Ripley, B. D., *Spatial Statistics*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Math Section, Wiley, New York, 1981.

[20]Cressie, N. A. C., *Statistics for Spatial Data*, Wiley, New York, 1991.

[21]Stein, M. L., *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Series in Statistics, Springer-Verlag, New York, 1999.

[22]Koehler, J. R., and Owen, A. B., "Computer Experiments," *Handbook of Statistics*, edited by S. Ghosh and C. R. Rao, Elsevier Science, New York, 1996, pp. 261–308.

[23]Booker, A. J., "Design and Analysis of Computer Experiments," *7th AIAA/USAF/NASA ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Vol. 1, AIAA, Reston, VA, 1998, pp. 118–128.

[24]Ripley, B. D., *Statistical Inference for Spatial Processes*, Cambridge Univ. Press, Cambridge, England, U.K., 1988.

[25]Booker, A. J., Conn, A. R., Dennis, J. E., Frank, P. D., Trosset, M., and Torczon, V., "Global Modeling for Optimization: Boeing/IBM/Rice Collaborative Project 1995 Final Report," The Boeing Co., Tech. Rep. ISSTECH-95-032, Bellevue, WA, Dec. 1995.

[26]Martin, J. D., and Simpson, T. W., "A Monte Carlo Simulation of the Kriging Model," AIAA Paper 2004-4483, 2004.

[27]Royston, P., "Remark AS R94: A Remark on Algorithm AS 181 the W-test for Normality," *Applied Statistics*, Vol. 44, No. 4, 1995, pp. 547–551.

[28]Shapiro, S. S., and Wilk, M. B., "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, Vol. 52, No. 3/4, 1965, pp. 591–611.

[29]Royston, P., "A Toolkit for Testing for Non-Normality in Complete and Censored Samples," *The Statistician*, Vol. 42, No. 1, 1993, pp. 37–43.

[30]Mardia, K., and Marshall, R., "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression," *Biometrika*, Vol. 71, No. 1, 1984, pp. 135–146.

[31]Kitanidis, P. K., "Parameter Uncertainty in Estimation of Spatial Functions: Bayesian Analysis," *Water Resources Research*, Vol. 22, No. 4, 1986, pp. 499–507.

[32]Mardia, K., and Watkins, A. J., "On Multimodality of the Likelihood in the Spatial Linear Model," *Biometrika*, Vol. 76, No. 2, 1989, pp. 289–295.

[33]Warnes, J. J., and Ripley, B. D., "Problems with Likelihood Estimation of Covariance Function of Spatial Gaussian Processes," *Biometrika*, Vol. 74, No. 3, 1987, pp. 640–642.

[34]Fadeev, D. K., and Fadeeva, V. N., *Computational Methods of Linear Algebra*, W. H. Freeman and Co., San Francisco, 1963, p. 126.

[35]Davis, G. J., and Morris, M. D., "Six Factors Which Affect the Condition Number of Matrices Associated with Kriging," *Mathematical Geology*, Vol. 29, No. 5, 1997, pp. 669–683.

[36]Sasena, M. J., "Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations," Ph.D. Dissertation, Dept. of Mechanical Engineering, Univ. of Michigan, Ann Arbor, MI, 2002.

[37]Booker, A. J., "Well-Conditioned Kriging Models for Optimization of Computer Simulations," The Boeing Co., Tech. Rept. M&CT-TECH-00-002, Bellevue, WA, Feb. 2000.

[38]Currin, C., Mitchell, T. J., Morris, M. D., and Ylvisaker, D., "A Bayesian Approach to the Design and Analysis of Computer Experiments," Oak Ridge National Lab., Tech. Rept. ORNL-6498, Oak Ridge, TN, Sept. 1988.

[39]Geisser, S., and Eddy, W. F., "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, Vol. 74, No. 365, 1979, pp. 153–160.

[40]Martin, J. D., and Simpson, T. W., "On Using Kriging Models as Probabilistic Models in Design," Society of Automotive Engineers, 2004 SAE World Congress, Paper 2004-01-0430, March 2004.

[41]Martin, J. D., and Simpson, T. W., "On the Use of Kriging Models to Approximate Deterministic Computer Models," American Society of Mechanical Engineers, ASME Design Engineering Technical Conf.—Design Automation Conf., DETC2004/DAC-57300, Sept.–Oct. 2004.

[42]Mitchell, T. J., and Morris, M. D., "Bayesian Design and Analysis of Computer Experiments: Two Examples," *Statistica Sinica*, Vol. 2, No. 2, 1992, pp. 359–379.

[43]Martin, J. D., and Simpson, T. W., "A Study on the Use of Kriging Models to Approximate Deterministic Computer Models," American Society of Mechanical Engineers, ASME Design Engineering Technical Conf.—Design Automation Conf., DETC2003/DAC48762, Sept. 2003.

[44]Wahba, G., "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem," *The Annals of Statistics*, Vol. 13, No. 4, 1985, pp. 1378–1402.

[45]Stein, M. L., "A Comparison of Generalized Cross Validation and Modified Maximum Likelihood for Estimating the Parameters of a Stochastic Process," *The Annals of Statistics*, Vol. 18, No. 3, 1990, pp. 1139–1157.

[46]Meckesheimer, M., Booker, A. J., Barton, R. R., and Simpson, T. W., "Computationally Inexpensive Metamodel Assessment Strategies," *AIAA Journal*, Vol. 40, No. 10, 2002, pp. 2053–2060.

[47]Laslett, G. M., "Kriging and Splines: An Empirical Comparison of Their Predictive Performance in Some Applications," *Journal of the American Statistical Association*, Vol. 89, No. 426, 1994, pp. 391–400.

[48]Akaike, H., "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, Vol. AC-19, No. 6, 1974, pp. 716–723.

[49]Vecchia, A. V., "Estimation and Model Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society, Series B*, Vol. 50, No. 2, 1988, pp. 297–312.

[50]Myers, R. H., and Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd ed., Wiley Series in Probability and Statistics, Wiley, New York, 2002, pp. 46–48.

[51]Anderson, D. R., Burnham, K. P., and White, G. C., "Comparison of Akaike Information Criterion and Consistent Akaike Information Criterion for Model Selection and Statistical Inference from Capture-Recapture Studies," *Journal of Applied Statistics*, Vol. 25, No. 2, 1998, pp. 263–282.

[52]Martin, J. D., and Simpson, T. W., "Use of Adaptive Metamodeling for Design Optimization," AIAA Paper 2002-5631, Sept. 2002.

[53]Jin, R., Chen, W., and Sudjianto, A., "On Sequential Sampling for Global Metamodeling in Engineering Design," American Society of Mechanical Engineers, ASME Design Engineering Technical Conf.—Design Automation Conf., DETC2002/DAC-34092, Sept.–Oct. 2002.

[54]McKay, M., Conover, W., and Beckman, R., "A Comparison of Three Methods for Selecting Values of Input in the Analysis of Output from a Computer Code," *Technometrics*, Vol. 21, No. 2, 1979, pp. 239–245.

[55]Jones, D. R., Schonlau, M., and Welch, W. J., "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, Vol. 13, No. 4, 1998, pp. 455–492.

[56]Nelder, J. A., and Mead, R., "A Simplex Method for Function Minimization," *Computer Journal*, Vol. 7, No. 4, 1965, pp. 308–313.

[57]Storn, R., and Price, K., "Differential Evolution—A Simple and Efficient Heuristic for Global Optimization over Continuous Space," *Journal of Global Optimization*, Vol. 11, No. 4, 1997, pp. 341–359.

[58]Abell, M. L., Braselton, J. P., and Rafter, J. A., *Statistics with Mathematica*, Academic Press, San Diego, CA, 1999, p. 340.

E. Livne
Associate Editor