On Using Standard Residuals as a Metric of Kriging Model Quality

Christopher D. Congdon*
Pennsylvania State University, University Park, PA 16802

Jay D. Martin†
The Applied Research Laboratory, State College, PA, 16804-0030

An interpolating kriging model, though it will always return the observations exactly, may not provide a good representation of the computer simulation at other values within the input domain. Without access to additional and potentially costly validation observations, it is difficult to determine if a kriging model is a good representation of the original computer model. One method to determine the predictive quality of a kriging model is to use leave-one-out cross-validation. A second difficulty with creating kriging models is a lack of diagnostic tests to determine how to improve the kriging model to result in a better estimation of the original computer model. This paper presents developments of diagnostic tools for creating kriging models. A computationally efficient form for the leave-one-out cross-validation residual and the variance at the left out location is presented. The standardized residuals can then be used to test if all of the observations appear to come from the Gaussian spatial process specified by the kriging model. This lack of fit may be the result of: 1) erroneous data, 2) the form of the kriging model is not sufficient to estimate the observations as a Gaussian process, 3) or the range of the model is not well represented by a single spatial random process. Two practical examples are provided to demonstrate how to interpret the results and make decisions on how to improve the predictive capability of the kriging model. The first example is a one-dimensional adiabatic flame temperature calculation. The second problem is a two-dimensional Branin test function.

Nomenclature

e	=	Residual error
σ^2	=	Process variance
d	=	Dimension of the input space
n	=	Number of observations included in input vector $\mathbf{x}$
k	=	Number of trend function coefficients
$\mathbf{x}$	=	Input vector (d x 1)
$\mathbf{X}$	=	Matrix of input vectors (n x d)
$\mathbf{E}_{ij}$	=	Elementary Matrix for Row/Column Interchange of i and j.
$\mathbf{I}$	=	Identity Matrix
$\boldsymbol{\beta}$	=	Vector of trend coefficients (k x 1)
$\boldsymbol{\theta}$	=	Vector of correlation coefficients (d x 1)
$\mathbf{r}(\mathbf{x})$	=	Correlation vector (n x 1)
$\mathbf{R}$	=	Correlation matrix (n x n)

* Graduate Assistant, Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802.
† Research Associate, Applied Research Laboratory, The Pennsylvania State University, P.O. Box 30, State College, PA, 16804-0030 Email: jdm111@psu.edu, and AIAA Member.

$R(\mathbf{x}_1, \mathbf{x}_2)$ = Spatial Correlation Function

$\mathbf{f}(\mathbf{x})$ = Vector of the set of regression functions of input vector $\mathbf{x}$ (k x 1)

$\mathbf{F}$ = Matrix of regression functions of input matrix $\mathbf{X}$ (n x k)

y = Response variable

q_i = Reciprocal of the i^{th} diagonal of the Inverse of the Correlation Matrix

$\mathrm{Z}(\mathbf{x})$ = Model of stationary Gaussian random process

subscripts

i = i^{th} row of vector/matrix

(i) = i^{th} row removed from vector/matrix

superscripts

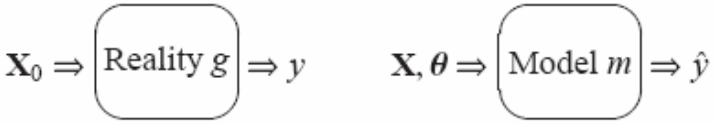
^{T} = Transpose operator

$\hat{}$ = Predicted value or function

I. Introduction

This paper investigates the use of standard residuals as a metric for evaluating the quality of a kriging model. The use of kriging models in system design is increasing in popularity due to its ability to provide a computationally efficient model that can interpolate a set of observations from a computationally expensive computer simulation. A kriging model is a metamodel, i.e. a model of a model. It is often used as a Design and Analysis of Computer Experiments (DACE) [2] tool to fit observed data with an interpolating model. The kriging model is a Gaussian Process model, capable of providing an expected value and variance at any location within the domain of the model's input space. This work provides methods to evaluate the predictive capabilities of the kriging model by using standardized leave-one-out cross-validation residuals to test if all of the observations appear to come from the Gaussian spatial process specified by the kriging model. This lack of fit may be the result of: 1) erroneous data, 2) the form of the kriging model is not sufficient to estimate the observations as a Gaussian process, 3) or the range of the model is not well represented by a single spatial random process.

Reality consists of an unknown deterministic relationship g that maps an unknown dimensioned input vector $\mathbf{X}_0$ to the observed output y as shown in Figure 1. A model m is designed to approximate reality by using an observable subset of inputs $\mathbf{X}$ and a set of constant (but unknown) θ which are invariant to the input. Sacks et al. [2] state that a computer model is deterministic because measurement error is non-existent. The same inputs to the model produce identical results and therefore the traditional experimental techniques of physical experimentation, which are designed to deal with measurement error, are lacking.



(a) Basic Form of Reality (b) Basic Form of a Model

Figure 1. Representations of Reality and a Model. [1]

One concern in using any model is that it provides the best representation of the observations as possible. An interpolating kriging model, though it will always return the observations exactly, may still not provide a good representation of the computer simulation at other values within the input domain. Without the use of additional observations, it is difficult to quantify the predictive capability of an interpolating kriging model [3]. In most cases it is better to include these additional observations in the kriging model to improve its accuracy rather than use them to test the quality of the model. The focus of this study is to use cross-validation methods to create a test for the quality of a kriging model by calculating standardized residuals at each left out observation.

There are typically three reasons a kriging model is unable to accurately estimate the output of a computer simulation. The first and easiest to resolve is that the model parameters values currently used are not the best values. This can be resolved through using either a better global optimization algorithm with Maximum Likelihood Estimation or integration when using Bayesian Analysis. The second reason is that insufficient observations are available to sample the variability that exists in the computer simulation. This can be recognized by a large process variance in the model form. The best method to resolve this issue is to use entropy-based iterative sampling [2]. The

last reason is that the best form of the kriging model has not been chosen to represent the mapping of the input space to the output values of the computer simulation.

This paper will present background material on the kriging model and on model adequately diagnostics. Specifically, this paper focuses on the standardized cross-validation residual (SCVR) as a diagnostic statistic, which is a leave-one-out cross-validation technique. A computationally-efficient form of the leave-one-out cross-validation residual and the variance at the left out location is then presented. This form allows for the calculation of every SCVR for a kriging model with only a single matrix inversion. Finally, two practical examples are provided to demonstrate the use of the developed algorithms and to provide examples of how to interpret the results and make decisions on how to improve the predictive capability of the kriging model.

II. Background

The term kriging represents a family of generalized least-square regression algorithms named after D. Kringe, who introduced the original model for analyzing mining data[4]. Currin et al. [5] provide one of the first references regarding the use of kriging to approximate a computer model, and Sacks et al. [2] popularized the technique for use in DACE. There are many texts in geostatistics [4,6] and in spatial statistics [7] that provide many details on the development and use of kriging models in their respective disciplines.

This section is broken up into 3 subsections in order to provide the framework for the proposed SCVR diagnostic approach. First, an overview of the kriging model specifically for use in DACE is provided. The second section provides discussion of residual diagnostics as used in linear regression. Finally, some background regarding cross-validation residuals is provided.

A. Kriging Model

A kriging model is a Best Linear Unbiased Estimator model that explicitly accounts for the correlation in the residuals between the regression model and the observations [8]. The kriging model, denoted by,

$$\hat{y}(\mathbf{x}) = \sum_{i=1}^{k} \beta_i \, \mathrm{f}_i(\mathbf{x}) + \mathrm{Z}(\mathbf{x}) \,. \tag{1}$$

consists of two components. The global trend component of the kriging model is a linear regression of the data with *k* regressors. The best form of the trend model used in the kriging model can be chosen by using traditional information criteria for model form selection [9]. The second component, $\mathrm{Z}(\mathbf{x})$, is a model of a stationary Gaussian spatial random process with zero mean and covariance is

$$\mathrm{Cov}(\mathrm{Z}(\mathbf{x}_1), \mathrm{Z}(\mathbf{x}_2)) = \sigma^2 R(\mathbf{x}_1, \mathbf{x}_2) \,, \tag{2}$$

where σ^2 is the process variance and $R(\mathbf{x}_1, \mathbf{x}_2)$ is a spatial correlation function (SCF) between data points. The kriging model is defined by selecting the trend function regressors $\mathrm{f}(\mathbf{x})$ and the SCF $R(\mathbf{x}_1, \mathbf{x}_2)$ where

$$R(\mathbf{x}_1, \mathbf{x}_2) = \prod_{i=1}^{d} R(x_{2,i}, x_{1,i}) \,. \tag{3}$$

The multivariate correlation function shown in Eq. (3) [10,11]is used to combine each of the *d* input dimensions rather than using the Euclidean norm of the space. This formulation improves the flexibility of modeling the correlation of each input dimension at the expense of requiring the selection of additional correlation function parameters (*d* parameters instead of just one).

The minimum requirements to consider when choosing a potential SCF are that it must be positive semi-definite and finite. The desired properties of an SCF are the ability to control: (1) the range of influence of nearby points, (2) the smoothness of the resulting surface, and (3) the differentiability of that surface. Koehler and Owen [49] provide an insightful overview of four commonly used SCFs used with kriging models: Gaussian, exponential, cubic spline, and Mat´ern functions. The Gaussian function is the most commonly used SCF in engineering design [12] as it

provides a relatively smooth and infinitely differentiable surface, making it a better choice when used with gradient-based optimization algorithms and only requires the selection of a single correlation parameter. The form of the Gaussian SCF used in this paper is

$$R(\mathrm{x}_1, \mathrm{x}_2) = \exp\left[-\left(\frac{\left| \mathrm{x}_2 - x_1 \right|}{\theta} \right)^2 \right] \tag{4}$$

where θ is the range parameter or correlation coefficient, which much be specified for each dimension *d*. With this form of the Gaussian SCF, θ can be thought of as the distance at which the influence is e^{-1} or approximately 37% [3]. The best linear unbiased estimator (BLUE) of $\hat{y}(\mathbf{x})$ is then given by

$$\hat{y}(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta} + \mathbf{r}^{\mathrm{T}}(\mathbf{x})\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}). \tag{5}$$

$\boldsymbol{\beta}$ is the vector of trend parameters β, $\mathbf{F}$ is the matrix of trend functions $\mathbf{f}^{\mathrm{T}}(\mathbf{x})$, $\mathbf{r}^{\mathrm{T}}(\mathbf{x})$ is a spatial correlation vector given by

$$\mathbf{r}^{\mathrm{T}}(\mathbf{x}) = \left[R(\mathbf{x}, \mathbf{x}_1) \quad R(\mathbf{x}, \mathbf{x}_2) \quad \cdots \quad R(\mathbf{x}, \mathbf{x}_n) \right] \tag{6}$$

and $\mathbf{R}$ is a spatial correlation matrix that is related to $\mathbf{r}^{\mathrm{T}}(\mathbf{x})$ by

$$\mathbf{R} = \left[\mathbf{r}(\mathbf{x}_1) \quad \mathbf{r}(\mathbf{x}_2) \quad \cdots \quad \mathbf{r}(\mathbf{x}_n) \right]. \tag{7}$$

The kriging model, in its traditional form, assumes that the variability of the spatial process is constant across the input domain, and is often referred to as being homoscedastic. This research explores using residual diagnostics to determine if the homoscedastic requirement of the kriging model has been violated.

B. Residual Diagnostics

A metamodel's quality can be assessed with two measurements [9]: 1) accuracy reproducing the observed data and 2) accuracy in predicting the original model in unobserved locations. Since kriging (and other interpolating models), by definition, reproduce observed data exactly, this traditional definition of a residual error is useless for evaluating an interpolating kriging.

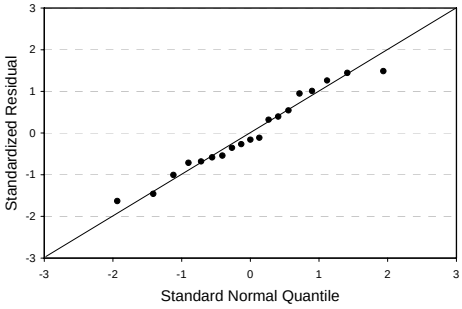
The key assumptions for using a kriging model as an estimator are identical to those used in linear regression and response surface methodology [13] In linear regression, the errors are typically required to be normally, independently and identically distributed. The concept of model adequacy checking is used to ensure that the model adequately approximates reality and to ensure that the assumptions used to generate the model parameters were not violated. This framework leads to a series of residual plot tests which can be used to test the model assumptions. The residual error e is the difference between reality $y(\mathbf{x})$ and the predicted value $\hat{y}(\mathbf{x})$ and is described by

$$e(\mathbf{x}) = y(\mathbf{x}) - \hat{y}(\mathbf{x}). \tag{8}$$

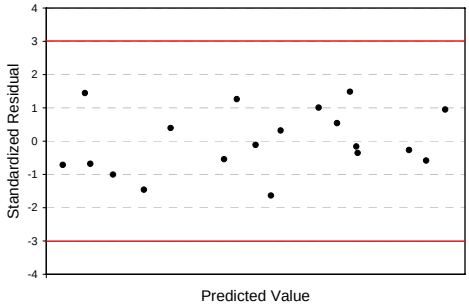
The residuals are often standardized by dividing by the process variance in order to scale them to a unit variance as shown by

$$e(\mathbf{x})_S = \frac{y(\mathbf{x}) - \hat{y}(\mathbf{x})}{\hat{\sigma}(\mathbf{x})}. \tag{9}$$

Section III discusses a strategy to produce standardized cross-validation residuals (SCVR) for a kriging model which will be useful for kriging model diagnostics.



(a) Normal Probability Plot of Standardized Residuals.



(b) Standardized Residuals vs. Predictions.

Figure 2. Typical Residual Diagnostic Plots for Linear Regression.

Figure 2 demonstrates two different diagnostic plots that are used for linear regression. The normal probability plot shown in Figure 2a is useful to check that the normality assumption of the residuals is appropriate. If there are strong deviations from a straight line, it is likely that the model assumptions should be revisited. It is possible that an incorrect trend function was chosen or that the error variance is not constant. A second test is to plot the standardized residuals vs. predicted value $\hat{y}(\mathbf{x})$ (see Figure 2b). This plot is useful for detecting outliers and can assist in validating the constant variance assumption. Any observation with a standardized residual outside of +/- 3 may indicate a region where the fitted model is a poor approximation of reality [13]. An ordered-structure to the residuals in this plot could indicate an incorrect trend function, missing parameters, or a correlation structure that has not been accounted for in the model. [14]

A limitation of graphical-based residual diagnostics is the ability to represent higher dimensions. As d>2, it becomes increasingly difficult to make sense of the various plots. Cook [15] presents a methodology to generate partial residual plots when d>2 in an attempt to make sense of multivariate data. However, it would be more objective to utilize quantitative methods to provide insight to the validity of the model assumptions. The Shapiro-Wilk test can be used to test the null hypothesis that the residuals come from a normally distributed population. [16] A rejection of the null hypothesis means a high confidence that the normality assumption is invalid. The Cliff-Ord test [17] can be used to test for spatial correlation in residuals. Cook and Weisberg present a test for heteroscedasticity, i.e. the spatial covariance is not constant over input domain, based on the score statistic.[18]

C. Cross-Validation Residuals

An interpolating model such as kriging exactly reproduces every observed point therefore the traditional definition of residuals is useless. A modification to the residual error calculation that is meaningful to kriging is Predicted Error Sum of Squares (PRESS) or cross-validation residual analysis [19,20]. The i^{th} PRESS residual is calculated by generating a kriging model with observation i removed, while the model parameters are held constant (i.e. they are not re-estimated for each subset of observations). An in-depth description of the PRESS method for linear regression can be found in Meyers and Montgomery [13]. This PRESS residual error $e_{(i)}$ is the difference between the predicted response at the i^{th} left-out point $\hat{y}_{(i)}(\mathbf{x}_i)$ and observed $y(\mathbf{x}_i)$ defined as

$$e_{(i)} = y(\mathbf{x}_i) - \hat{y}_{(i)}(\mathbf{x}_i). \tag{10}$$

Therefore, the predicted error sum of squares is

$$\text{PRESS} = \sum_{i=1}^{n} \left(e_{(i)}\right)^2. \tag{11}$$

The standardized cross-validation residual (SCVR) at each point *i* is

$$\text{SCVR} = \frac{e_{(i)}}{\sqrt{\text{VAR}\left[e_{(i)}\right]}}, \tag{12}$$

where $\text{VAR}\left[e_{(i)}\right]$ is the variance of the left-one-out residual.

The SCVR test can provide insight into the fit of the individual point to the overall model. A properly-defined model should be able to predict the location of individual points after those points are removed. This insight is useful in gauging the predictive capability of a regression model [13]. Similar diagnostic strategies of linear regression can be employed for kriging models. Typically, a kriging model assumes that the spatial correlation function is constant across the model's domain (homoscedastic), i.e. the spatial covariance is independent of the location in space. Additionally it must be assumed that the residual errors are also normally, independently and identically distributed to use MLE for parameter selection and to estimate the probability distribution of the output. Independence can be determined by checking the spatial correlation of the kriging model's SCVRs.

Jones et al. [19] provide several methodologies to validate stochastic processes models (kriging) prior to using them for optimization. A similar SCVR concept to that presented here for validation was utilized. They suggest that the SCVR should be contained to +/- 3. Just as for the linear regression models, Jones et al. [19] show that an analysis of these SCVRs can provide an indication of the kriging model's ability to fit the data. If there are a number of points that don't fit very well (i.e. their SCVR is over +/- 3), they found that this could be an indication that the homoscedastic requirement has been violated.

III. Standardized Cross-Validation Residuals

In general, the calculation of the leave-one-out cross-validation errors could require the calculation of *n*, (*n*-1) x (*n*-1) correlation matrices, a possibly computationally expensive calculation. Currin et al. [5] produced equations for efficiently calculating the leave-one-out cross-validation residuals for a kriging model with a constant trend model. Their equations only require the inverse of the original correlation matrix $\mathbf{R}$. These equations were slightly modified to incorporate a non-constant trend function by Martin and Simpson [21]. The PRESS residual for a non-constant trend model for each observation *i* can be calculated as follows. The predicted leave-one-out response at location *i* is derived from Eq. (5) and denoted by

$$\hat{y}_{(i)}(\mathbf{x}_i) = \mathbf{f}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta} + \mathbf{r}_{(i)}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{R}_{(i)}^{-1}\left(\mathbf{y}_{(i)} - \mathbf{F}_{(i)}\boldsymbol{\beta}\right). \tag{13}$$

Substituting Eq. (13) into Eq. (10) and collecting the $\boldsymbol{\beta}$ terms results in

$$e_{(\mathrm{i})} = \left(y(\mathbf{x}_i) - \mathbf{r}_{(i)}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{R}_{(i)}^{-1}\mathbf{y}_{(i)}\right) - \left(\mathbf{f}^{\mathrm{T}}(\mathbf{x}_i) - \mathbf{r}_{(i)}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{R}_{(i)}^{-1}\mathbf{F}_{(i)}\right)\boldsymbol{\beta}. \tag{14}$$

It can be shown that the operation of swapping the i^{th} and j^{th} observations of both $\mathbf{R}$ and $\mathbf{R}^{-1}$ via elementary matrix operations preserves their relationship

$$\mathbf{E}_{ij}\mathbf{R}\mathbf{E}_{ij} = \left[\mathbf{E}_{ij}\mathbf{R}^{-1}\mathbf{E}_{ij}\right]^{-1}, \tag{15}$$

where $\mathbf{E}_{ij}$ is the elementary matrix for the interchange of row/column *i* with row/column *j*. Therefore, a partitioned matrix of both $\mathbf{R}$ and $\mathbf{R}^{-1}$ can be generated with any observation *i* placed in the last row/column so that:

$$\mathbf{R}\mathbf{R}^{-1} = \begin{bmatrix} \mathbf{R}_{(i)} & \mathbf{r}_{(i)}(\mathbf{x}_i) \\ \mathbf{r}_{(i)}^{\mathrm{T}}(\mathbf{x}_i) & 1 \end{bmatrix}\begin{bmatrix} \mathbf{P} & \mathbf{u}_{(i)} \\ \mathbf{v}_{(i)} & \frac{1}{q_i} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \tag{16}$$

where q_i is the reciprocal of the i^{th} diagonal element of $\mathbf{R}^{-1}$, $\mathbf{u}_{(i)}$ and $\mathbf{v}_{(i)}$ are the i^{th} column and row of $\mathbf{R}^{-1}$ respectively. Note that for a symmetric $\mathbf{R}^{-1}$, $\mathbf{v}_{(i)} = \mathbf{u}_{(i)}^{\mathrm{T}}$. The first row of the partitioned $\mathbf{R}$ multiplied by the last column of the partitioned $\mathbf{R}^{-1}$ by definition equals 0. Solving for $\mathbf{u}_{(i)}$ obtains

$$\mathbf{u}_{(i)} = \frac{\mathbf{R}_{(i)}^{-1}\mathbf{r}_{(i)}(\mathbf{x}_i)}{q_i}, \tag{17}$$

and the transpose of it is

$$\mathbf{u}_{(i)}^{\mathrm{T}} = -\frac{\mathbf{r}_{(i)}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{R}_{(i)}^{-1}}{q_i}. \tag{18}$$

Substituting Eq. (18) into Eq. (14) we get

$$e_{(\mathrm{i})} = \left(y(\mathbf{x}_i) + q_i\mathbf{u}_{(i)}^{\mathrm{T}}\mathbf{y}_{(i)}\right) - \left(\mathbf{f}^{\mathrm{T}}(\mathbf{x}_i) + q_i\mathbf{u}_{(i)}^{\mathrm{T}}\mathbf{F}_{(i)}\right)\boldsymbol{\beta}. \tag{19}$$

The next step is to recognize that the first component of Eq. (19) is the last column of the partitioned $\mathbf{R}^{-1}\mathbf{y}$ where

$$\mathbf{R}^{-1}\mathbf{y} = \begin{bmatrix} \mathbf{P}\mathbf{y}_{(i)} + \mathbf{u}_{(i)}y(\mathbf{x}_i) \\ \mathbf{u}_{(i)}^{\mathrm{T}}\mathbf{y}_{(i)} + y(\mathbf{x}_i)/q_i \end{bmatrix} \tag{20}$$

and the second component of Eq. (19) is the last column of the partitioned $\mathbf{R}^{-1}\mathbf{F}$ where

$$\mathbf{R}^{-1}\mathbf{F} = \begin{bmatrix} \mathbf{P}\mathbf{F}_{(i)} + \mathbf{u}_{(i)}\mathbf{f}^{\mathrm{T}}(\mathbf{x}_i) \\ \mathbf{u}_{(i)}^{\mathrm{T}}\mathbf{F}_{(i)} + \mathbf{f}^{\mathrm{T}}(\mathbf{x}_i)/q_i \end{bmatrix}. \tag{21}$$

Since we only care about the last terms, we can substitute Eq. (20) and Eq. (21) into Eq. (19) to get

$$e_{(\mathrm{i})} = q_i\left(g_i - w_i\boldsymbol{\beta}\right) \tag{22}$$

where $g = \mathbf{R}^{-1}\mathbf{y}$ and $w = \mathbf{R}^{-1}\mathbf{F}$. If the elements of q are placed on the diagonal of a matrix $\mathbf{Q}$, the more general form of the PRESS for a non-constant trend function is

$$e_{\mathrm{PRESS}} = \mathbf{Q}\left(g - w\boldsymbol{\beta}\right). \tag{23}$$

Equation (23) provides a general form for the leave-one-out cross-validation residual required to calculate the SCVR. The expected variance of the each leave-one-out cross-validation location can be calculated without needing to invert n $(n-1)\times(n-1)$ correlation matrices. The variance of the prediction at the left-out location i is defined as [22]:

$$\mathrm{Var}\left[e_{(i)}\right] = \sigma^2\left(1 - \mathbf{r}_{(i)}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{R}_{(i)}^{-1}\mathbf{r}_{(i)}(\mathbf{x}_i)\right) \tag{24}$$

where σ^2 is the process variance. Using Eq. (16), it can be shown

$$\mathbf{r}_{(i)}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{u}_{(i)} + \frac{1}{q_i} = 1. \tag{25}$$

Substituting Eq. (17) into Eq. (25) and solving for q_i obtains

$$-\mathbf{r}_{(i)}^{\mathrm{T}}(\mathbf{x}_i)\frac{\mathbf{R}_{(i)}^{-1}\mathbf{r}_{(i)}(\mathbf{x}_i)}{q_i} + \frac{1}{q_i} = 1 \Leftrightarrow q_i = 1 - \mathbf{r}_{(i)}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{R}_{(i)}^{-1}\mathbf{r}_{(i)}(\mathbf{x}_i). \tag{26}$$

Eq. (26) can be seen as the scaling parameter to the process variance in the leave-one-out variance shown previously and is defined identically as used for the leave-one-out residual error. The result of this simplification is that the leave-one-out variance from Eq. (24) is given by

$$\mathrm{Var}\left[e_{(i)}\right] = \sigma^2 q_i. \tag{27}$$

Now, in a similar form as Eq. (12), an equation for the SCVR for point *i* of a kriging model is now given by

$$\mathrm{SCVR} = \frac{\sqrt{q_i}\left(g_i - w_i\boldsymbol{\beta}\right)}{\sigma}. \tag{28}$$

IV. Examples

In order to demonstrate the interpretation of the SCVR-based diagnostics, the technique will be applied to two different models. First, a one dimensional adiabatic flame temperature model will be demonstrated. Secondly, the technique will be demonstrated on a two-dimensional test function. Traditional graphical methods of residual analysis will be demonstrated for each of the examples.

A. One Dimensional Adiabatic Flame Temperature Calculation

This 1-D example calculates the adiabatic flame temperature of methane in a constant-volume combustion process with respect to the equivalence ratio. The equivalence ratio of a fuel is the ratio of the stoichiometric oxidizer to fuel ratio to the actual oxidizer to fuel ratio [23]. In many combustion applications the equivalence ratio is the single most important factor in determining a system's performance. The model used for this example, the software AFTP[‡] , incorporates the effects of dissociation due to incomplete combustion and generates the adiabatic flame temperature of various fuels under numerous conditions. The sample points, indicated by the dots in Figure 3, are evenly spaced at 0.2 increments from 0 to 2 for a total of 10 points. The line in Figure 3 represents a set of 400 evenly spaced validation points.

[‡] Adiabatic Flame Temperature Program (AFTP) Software Program Developed by Christopher Depcik at University of Michigan, Version 1.1.1.1. http://www.depcik.com.

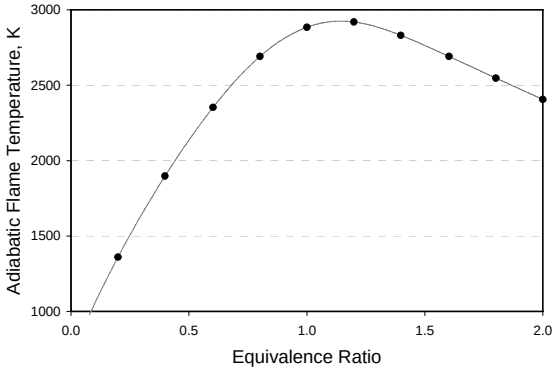
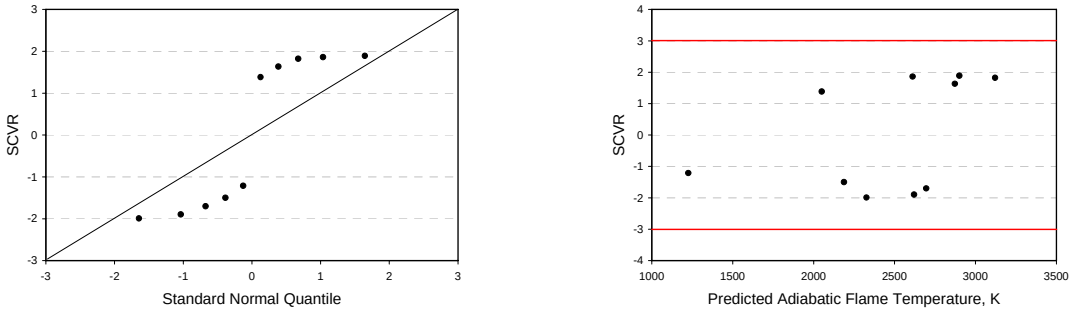


Figure 3. Adiabatic Flame Temperature vs. Equivalence Ratio for Methane in Constant Volume Combustion Process.

The kriging model used for this analysis consists of a 2nd order trend model and a Gaussian SCF defined in Eq. (4). The input to this problem was scaled to [0,1]. The model parameters were determined via an iterative process in which the trend coefficients β were calculated by Maximum Likelihood Estimation (MLE). The most appropriate trend model was found via traditional linear regression hypothesis testing methods for significant trend coefficients [24]. This process led to a full second order model where $\mathrm{f}(x) = \{1, x, x^2\}$. The diagnostic plots (see Section IIb) for the 1-D Adiabatic Flame Temperature Example are shown in Figure 4.



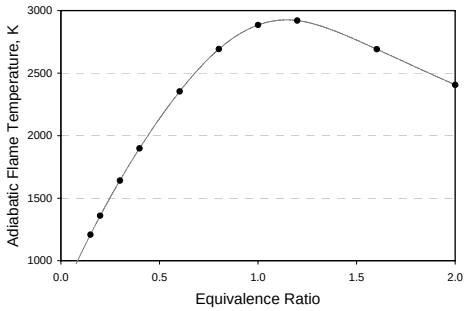
(a) Normal Probability Plot of SCVR. **(b)** SCVR vs. Cross-Validated Predictions.

Figure 4. Diagnostic tests for 1-D Adiabatic Flame Temperature Model:

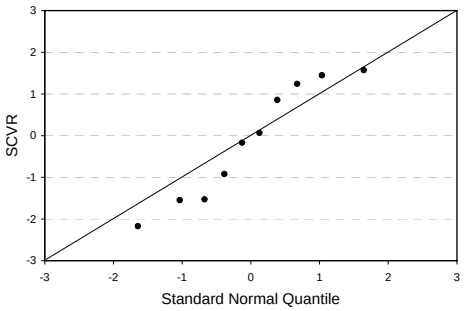
The normal probability plot, shown in Figure 4a, demonstrates a deviation from the normality assumption. The Shapiro-Wilk test confirms the graphical interpretation, with a p-value of 0.02. This violation of the normality assumption reduces the accuracy of the estimated parameters (trend and correlation coefficients) possibly reducing the prediction capability of the model.

Figure 4b presents the SCVR vs. cross-validation predictions. While the primary interpretation of this plot is in looking for outliers (+/- 3 SCVR) and in looking for trends in the residuals, a key observation is in the grouping of the sample points. The uniformly-distributed sampling strategy inadvertently generated a clustering of points which predict from 2000-3000K, with a single observation below 2000K (at 1300K). Lin [25] discusses pitfalls in using a cross-validation residual for kriging models when the function is highly non-linear and there are not enough points spread throughout the design space. Uniform sampling strategies on the decision space $\mathbf{X}$ does not guarantee uniform sampling on the objective space $\mathbf{f(X)}$. With *a priori* knowledge of this mapping, as in the case of this example, the clustering issue can be remedied by generating additional samples at the low-sampling density. Figure 5 demonstrates the effect of reducing the objective space clustering, by reapportioning 2 samples from the high-density values near 2500K to the low-density values near 1500K. These two points were reapportioned instead of

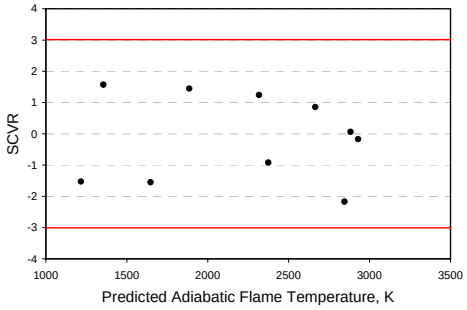
added to the model to separate the effect of additional samples from the desired effect of improved dispersion across the objective space.



(a) Sampling Strategy.



(b) Normal Probability Plot of SCVR.



(c) SCVR vs. Cross-Validated Predictions.

Figure 5. Effect of Reapportioning Samples on 1-D Adiabatic Flame Temperature Model.

This change significantly improved the normal probability plot of the SCVR (see Figure 5b) . This observation of improvement is collaborated by the Shapiro-Wilk test, which now has a p-value of 0.50. Figure 5c shows how the new sampling strategy is more balanced across predicted values.

While it may be easy to visualize and/or quantify the mapping of $\mathbf{X}$ onto the objective space for low-dimensional problems, it is not as obvious how to properly sample observations without *a priori* knowledge of mappings of more complex functions. There is active research on the process of sequential sampling and how it relates to kriging[26,27]. The impact of such strategies on improving model adequacy, specifically pertaining to SCVR analysis, will be studied in future research.

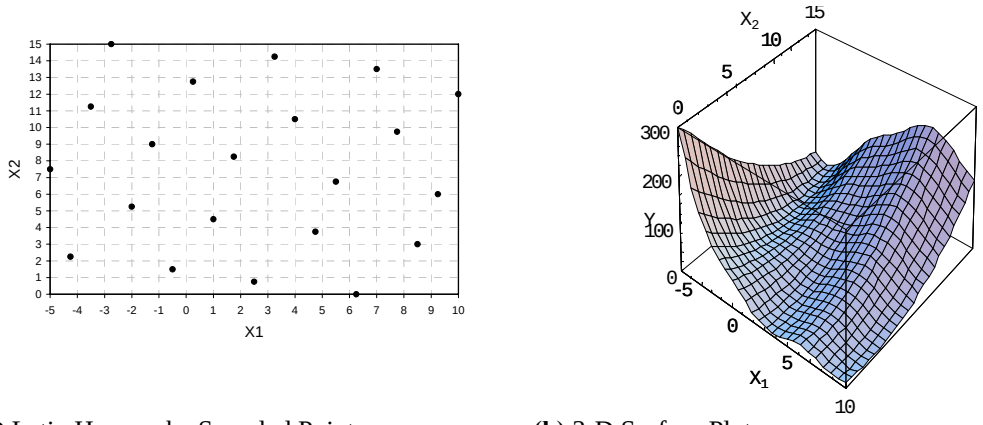
B. Two-Dimensional Branin Function

This 2-D example is the Branin test function, Eq. (29), which is a well known test function for global optimization. [28]. For this problem a 21-point Latin Hypercube is used for sampling, shown in Figure 6.

$$f(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2} x_1^2 + \frac{5}{\pi} x_1 - 6 \right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\mathrm{Cos}(x_1) + 10, \qquad x_1 \in [-5,10], x_2 \in [0,15] \tag{29}$$

For this example, two forms of the kriging model are examined in order to demonstrate the effectiveness of the SCVR diagnostic approach in detecting deviations from the model assumptions. As described in the background, the effectiveness of the kriging model is dependant on the selection of the correlation parameters for the SCF. Therefore, two separate cases are examined: a) where all parameters were selected via MLE and b) where one of the correlation coefficients was deviated from the MLE estimate in order to demonstrate the effect of a model with an

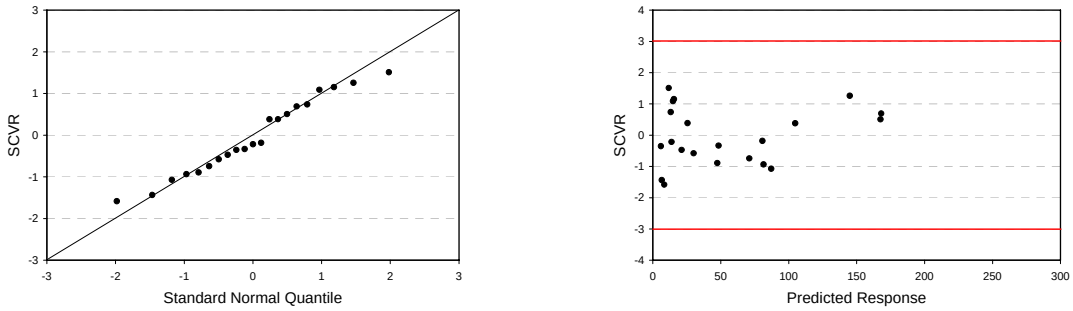
inadequate spatial component. Both models used the same second-order trend model where $\mathbf{f}(x) = \{1, x_2, x_1x_2, x_2^2\}$. The trend model form was selected by using the AIC [24], and the trend coefficients were selected by using MLE.



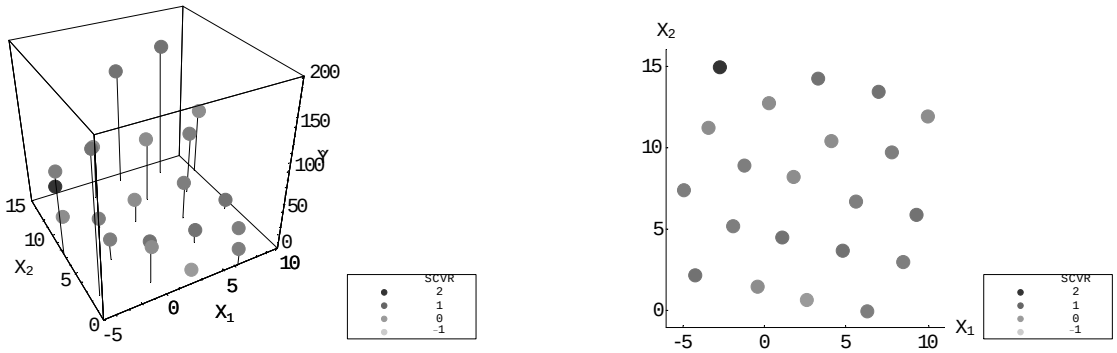
(a) 2-D Latin Hypercube Sampled Points. **(b)** 3-D Surface Plot.

Figure 6. 2 Dimensional Branin Function Sampling Strategy.

Figure 7 summarizes the key residual diagnostics for the optimal MLE parameter selection.



(a) Normal Probability Plot of SCVR. **(b)** SCVR vs. Cross-Validated Predictions.

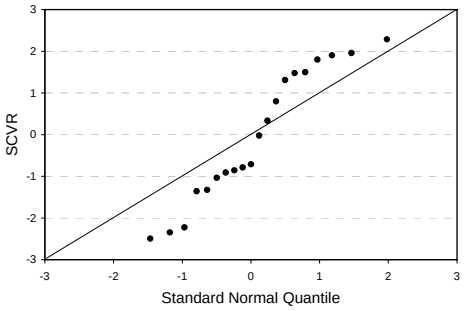


(c) 3-D projection of Cross-Validated Predictions vs. Location (Shaded by SCVR intensity). **(d)** Top-View of Cross-Validated Predictions vs. Location (Shaded by SCVR intensity).

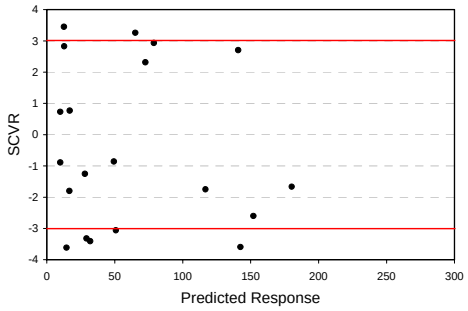
Figure 7. Diagnostic tests for 2-D Branin Test Function with Optimal Parameter Selection.

The Normal Probability Plot, shown in Figure 7a, indicates no issues with the normality assumption. The Shapiro-Wilk test fails to reject the null-hypothesis that the residuals come from a normal distribution, with a p-value of 0.54, confirming the visual inspection. The SCVR vs. cross-validation predictions plot in Figure 7b does not locate any outliers. There is a slight pattern in the residual plot, but it is likely due to variation sampling frequency over the range of predicted values. Additionally, there are no sampled points with a predicted response over 175, when the actual function achieves a maximum near 300 at x=[-5,0]. This is due to the LHS sampling not appropriately sampling the entire decision space. Figure 7c provides a 4-dimensional plot of the correlation structure of the SCVRs, with the 4th dimension representing the magnitude of the SCVR at each location. This can be a useful tool when *d* is small. Figure 7d is a top-down view of Figure 7c to help visualize the relationship of SCVR and the location in the decision space.

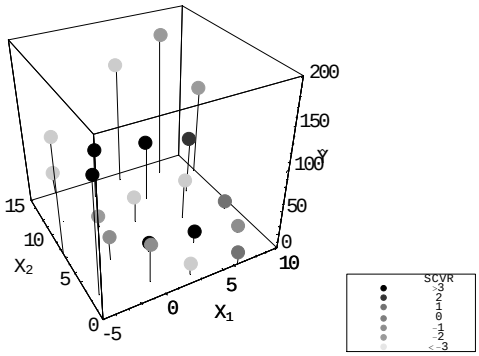
Figure 8 summarizes the key residual diagnostics for the second model, where the correlation parameter for x_1 was increased from 0.35 to 0.60. This modification was made in order to demonstrate the SCVR test's ability to evaluate model adequacy when the SCF does not adequately represent the process. The estimates of $\boldsymbol{\beta}$ and σ^2 were calculated independently for each model by MLE, given the prescribed correlation coefficients.



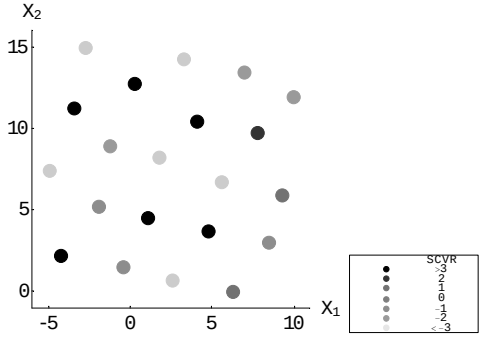
(a) Normal Probability Plot of SCVR.



(b) SCVR vs. Cross-Validated Predictions.



(c) 3-D projection of Cross-Validated Predictions vs. Location (Shaded by SCVR intensity).



(d) Top-View of Cross-Validated Predictions vs. Location (Shaded by SCVR intensity).

Figure 8. Diagnostic tests for 2-D Branin Test Function with Non-Optimal Correlation Coefficient.

The normal probability plot, shown in Figure 8a, demonstrates a deviation from the normality assumption. The Shapiro-Wilk test, with a p-value of 0.015, confirms the findings from the visual inspection. The SCVR vs. Predicted Response plot in Figure 8b demonstrates numerous outliners, which indicate a failure in the model to adequately predict the observed points. The 3-dimensional plots in Figure 8c and Figure 8d show a structure in the residuals, as there is evidence of clustering of SCVR by region.

V. Conclusion

This paper presented the Standardized Cross-Validation Residual (SCVR) as a possible method for determination of kriging model adequacy. A computationally-efficient form of the SCVR was presented to allow for its practical implementation as a tool to diagnose potential model-adequacy issues in kriging. The long-standing use of residuals in linear regression makes the translation to kriging a natural extension. The use of SCVRs for model

adequacy is relatively unexplored in the literature and the implementation of an efficient form for its calculation warrants additional research for proper implementation.

Two problems were provided as examples of diagnostics that can be performed using the SCVR methodology. The first example was a simple 1-D adiabatic flame temperature calculation which demonstrated the effect of clustering of data on the objective space and how analysis of the SCVRs was able to diagnose an inadequacy in the model. The second example explored the effect of improper parameter selection on the 2-D Branin Function. An analysis of the SCVRs showed a correlation structure in the residuals which corresponds to the selection of an inadequate spatial correlation function (due to improper parameter selection).

Future work relating to SCVRs can be broken down into 2 main areas. First, the implementation of quantitative model-adequacy measures should be studied, which can provide a more authoritative measure of the soundness of kriging model's assumptions. Possible candidates for research are the Cliff-Ord test[17] and Cook and Weisberg's test for heteroscedasticity [18]. Secondly, methods to address heteroscedasticity in the residuals require additional study.

There are three typical approaches to handle heteroscedastic data in a kriging model: 1) transformation on input space to make the data appear homoscedastic, 2) include a secondary spatial process to the kriging model, and 3) use a spatial variance that depends on location in the input's domain.

The first approach is similar to the approach used to address the residual normality requirement. Jones et al. [19] and Martin and Simpson [1] present methods for addressing this correlation between residuals. The second approach involves taking a region of points that violate the homoscedastic assumption and create a second kriging model for this secondary data. An encompassing function then combines both functions to create the metamodel. Booker [29] discusses combining two independent Gaussian processes to address ill-conditioned matrices caused by point "pile-up" as many optimization routines focus additional runs to a specific region, causing nearly identical row/columns in the resultant correlation matrix $\mathbf{R}$. Finally, the last approach attempts to use a more complex model to estimate the spatial covariance of the observed data. Higdon et al. [30] provide a SCF which allows the spatial dependence structure to vary as a function of location.

Acknowledgments

The authors would like to thank Dr. Kam Ng of the ONR333 for his generous support on Contract N00014-05-G-0106.

References

[1]Martin, J. D., Simpson, T. W., "On Using Kriging Models as Probabilistic Models in Design," *SAE Transactions Journal of Materials and Manufacturing*, Vol. 5, No. 2005, pp. 129-139.

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

[3]Martin, J. D., "A Methodology for Evaluating System-Level Uncertainty in the Conceptual Design of Complex Multidisciplinary Systems," Ph.D. Thesis, Mechanical Engineering, The Pennsylvania State University, University Park, PA, 2005.

[4]Goovaerts, P., *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York, 1997.

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

[6]Wackernagel, H., *Multivariate Geostatistcs*, Springer-Verlag Berlin Heidelberg, New York, 2003.

[7]Ripley, B. D., *Spatial Statistics*, John Wiley & Sons, New York, 1981.

[8]Goldberger, "Best Linear Unbiassed Prediction in the Generalized Linear Regression Model," *Journal of the Americal Statistical Association*, Vol. 57, No. 298, 1962, pp. 369.

[9]Martin, J. D., Simpson, T. W., "Use of Kriging Models to Approximate Deterministic Computer Models," *AIAA Journal*, Vol. 43, No. 4, 2005, pp. 853-863.

[10]Booker, A. J., Conn, A. R., Dennis, J. E. J., Frank, M., Trosset, M., Torczonm, V., "Global Modeling for Optimization," The Boeing Company 1995.

[11]Currin, C., Mitchell, T. J., Morris, M. D., 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.

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

[13]Myers, R. H., Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd ed., John Wiley & Sons, Inc., New York, 2002.

[14]Kutner, M. H., Nachtsheim, C., Neter, J., Li, W., *Applied Linear Statistical Models*, McGraw Hill, Homewood, Ill, 2005.

[15]Cook, R. D., "Exploring Partial Residual Plots," *Technometrics*, Vol. 35, No. 4, 1993, pp. 351-362.

[16]Shapiro, S. S., Wilk, M. B., "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, Vol. 52, No. 3/4, 1965, pp. 591-611.
[17]Cliff, A. D., Ord, J. K., *Spatial Autocorrelation*, Pion, London, 1973.
[18]Cook, R. D., Weisberg, S., "Diagnostics for Heteroscedasticity in Regression," *Biometrika*, Vol. 70, No. 1, 1983, pp. 1-10.
[19]Jones, D. R., Schonlau, M., Welch, W. J., "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, Vol. 13, No. 4, 1998, pp. 455-492.
[20]Booker, A. J., Dennis, J. E., Jr., Frank, P. D., Serafini, D. B., Torczon, V., Trosset, M., "A Rigorous Framework for Optimization of Expensive Functions by Surrogates," *Structural Optimization*, Vol. 17, No. 1, 1999, pp. 1-13.
[21]Martin, J. D., Simpson, T. W., "A Study on the Use of Kriging Models to Approximate Deterministic Computer Models," *2003 ASME Design Engineering Technical Conference*, Chicago, IL.
[22]Schonlau, M., "Computer Experiments and Global Optimization," Dissertation, Department of Statistics, University of Waterloo, Waterloo, Ontario, Canada, 1997.
[23]Turns, S., *An Introduction to Combustion- Concepts and Applications*, McGraw Hill, 2001.
[24]Martin, J. D., Simpson, T. W., "A Methodology to Manage System-Level Uncertainty During Conceptual Design," *ASME Journal of Mechanical Design*, Vol. 128, No. 4, 2006, pp. 959-968.
[25]Lin, Y., "An Efficient Robust Concept Exploration Method and Sequential Exploratory Experimental Design," Dissertation, Department of Mechanical Engineering, Georgia Institute of Technology, Atlanta, 2004.
[26]Jin, R., Chen, W., Sudjianto, A., "On Sequential Sampling for Global Meta-Modeling in Engineering Design," *2002 ASME Design Technical Conference*, Montreal, Canada, DETC2002/DAC-34092.
[27]van Beers, W. C. M., Kleijnen, J. P. C., "Kriging Interpolation in Simulation: A Survey," *Proceedings of the 2004 Winter Simulation Conference*, Washington, DC.
[28]Dixon, L. C. W., Szego, G. P., "The Global Optimization Problem; an Introduction," *Towards Global Optimization*, edited by Dixon, L. C. W., Szego, G. P., North Holland, Amsterdam, 1978, pp. 1-15.
[29]Booker, A. J., "Well-Conditioned Kriging Models for Optimization of Computer Simulations," The Boeing Company M&CT-TECH-00-002, 2000.
[30]Higdon, D., Swall, J., Kern, J., "Non-Stationary Spatial Modeling," *Bayesian Statistics*, Vol. 6, No. 1999, pp. 761-768.