Robust Kriging Models

Jay D. Martin *
Applied Research Laboratory, State College, PA, 16804-0030, USA

Kriging models have proven useful in estimating complex and computationally expensive analyses. They are capable of interpolating a set of observations by quantifying both longer range variations with parametric trends and shorter range variations with spatial correlations. Kriging models have had some difficulty with robustness in situations when there are a larger number of input dimensions and few observations as well as a larger number of observations with few dimensions. This paper will detail how to add a parameter to the kriging model that will account for random or measurement errors. The result is a model that will no longer interpolate all of the observations and may be a function of fewer of input dimensions. The resulting model may be better able to approximate the orignal function as demonstrated in four two-dimensional examples.

I. Introduction

Kriging models provide a very flexible and computationally efficient metamodel form that can be adapted to reproduce many complex response surfaces. It is useful for the approximation of computer analyses since it is capable of interpolating all of the observations used to create it. The kriging model is a statistics-based model that can incorporate the trend model properties from more typical linear regression with the spatial correlation properties of kernel-based approximation methods such as radial basis functions. The kriging model parameters can be objectively estimated given the set of observations from the process it is intended to approximate.

With the potential strengths of the kriging model form at approximating complex responses also come some weaknesses. These possible weaknesses include computational difficulty and expense at estimating the optimal model parameters. These weaknesses are the result of a lack of robustness with the kriging model. The robustness of a kriging model can be compromised in three practical situations. In the first situation, a kriging model is desired to fit a set of observations in which there are few observations (n) relative to the number of input dimensions (d)) (e.g. $n < 10d$). In this situation, there is not enough information to adequately estimate the spatial correlation that may exist between the observations. A second situation may arrise when there are a larger number of observations to interpolate relative to the number of input dimensions. In this situation there may be additional information present in the observations that can't be adequately quantified by the kriging model form. The last situation may arrise when a number of observations are located very close to each other, as may occur when using a kriging model as part of an iterative optimization process. In this situation, the kriging model can become numerically unstable. This work presents a method to adresses these three situations of robustness by demonstrating the impact on including a *nugget* parameter in the model that relaxes the interpolation constraint and gives the *universal* kriging model the ability to smoothly transition from the purely parametric trend model of linear regression to the strict interpolation model of *universal* kriging.

This article presents a possible option for dealing with the lack of robustness that has been experienced by many users of the interpolating kriging model form in the area of design and design automation. The next section of this article details the background of the interpolating kriging model formulation and provides a small addition of that form to loosen the requirement to interpolate all of the observations, the *nugget* parameter. One difficulty with this non-interpolating form may be the requirement to estimate at least one additional parameter to define the (spatial) covariance of the observed data used in the resulting model form. The Discussion of Formulation section provides a more general explanation of the formulation to give users a more intuitive feel for the mechanisms being used in the altered model form. Four demonstration problems are described and used to demonstrate the effectiveness of the new model form, highlighting the impact of the new *nugget* parameter on the resulting kriging model. Finally, the article closes with some conclusions about this robust kriging model form and provides a list of future developments required for the further use of this kriging model form.

II. Background

Kriging models were originally developed for the estimation of spatially distributed geological attributes given a set of observations [1]. The models were later used as computational tools to permit the increased performance of many

*Research Associate, PO Box 30, N. Atherton St. Member AIAA

different optimization algorithms against computationally expensive analysis models[2,3]. This line of development started with the design optimization being performed on the metamodel instead of the original computer analysis. Due to the inherent Bayesian nature of the kriging model, the ability to take an initial belief and iteratively update it with additional observations, the next line of development were adaptive metamodeling techniques. These techniques included EGO[4] and its many variations[5,6]. The basis of these methods were to create a multi-objective problem that included both the original objective function and a measurement of uncertainty in the kriging response surface. The effectiveness of these methods lie mostly in the scaling parameters used to weight the uncertainty in the output to the actual value being returned. The accuracy of the estimated uncertainty was not directly important to the success of the method. As such, little attention was needed to validate the accuracy of the resulting uncertainty estimates.

A. Kriging Model Form

The kriging model is a best linear unbiased predictor (BLUP) much like a linear regression model. The form of a kriging model can be represented as

$$\hat{y}(\mathbf{x}) = \sum_{i=1}^{k} \beta_i f_i(\mathbf{x}) + Z(\mathbf{x}) + \epsilon(\mathbf{x}) \tag{1}$$

where the first term on the right-hand side of the equation is a trend function that quantifies any long-range parametric trends that exist between the model inputs and output, the second term is a spatial stochastic process model that quantifies the shorter range spatial correlations that exist between the residuals of the observations and the trend function, and the third part of the model is a noise or measurement error that is not directly quantified by the model with the previous two terms. The interpolating kriging model does not include this last error term. By including the error term, the requirement to include all of the input dimensions to uniquely identify each observation for interpolation is removed. By removing this requirement, an analysis of the data can be performed to determine the most important input dimensions for estimating the output and leave the remaining dimensions out of the model and treat them as small noise sources rather than try to estimate the most appropriate trend and spatial correlation parameters, a process that can be both numerically difficult and may induce more errors into the predicting model. Finally, by including a small noise term in the model, the situation of where both short and long range spatial correlations exist in the observed data can be more easily accommodated by quantifying the shortest ranges with the noise term.

1. Covariance Model

The variance of the residuals from the linear regression model (the first term in Eq. 1) can be defined as $\sigma^2 = {\sigma_z}^2 + {\sigma_e}^2$ where ${\sigma_z}^2$ is the variance of the spatial process $Z(\mathbf{x})$ and ${\sigma_e}^2$ is the variance of the noise term or the measurement errors $\epsilon(\mathbf{x})$. The covariance of the spatial process $Z(\mathbf{x})$ from Eq 1 is given by

$$\begin{aligned} \operatorname{cov}(\mathrm{Z}(\mathbf{t}), \mathrm{Z}(\mathbf{u})) &= V(\mathbf{t}, \mathbf{u}) \\ &= {\sigma_z}^2 R(\mathbf{t}, \mathbf{u}) \end{aligned} \tag{2}$$

where the spatial correlation function $R(\mathbf{t}, \mathbf{u})$ is defined for multiple dimensions using the product of the correlations in each dimension as follows

$$R(\mathbf{t}, \mathbf{u}) = \prod_{i=1}^{d} \mathrm{e}^{-\left(\frac{\mathbf{u_i} - \mathbf{t_i}}{\theta_i}\right)^2}. \tag{3}$$

In general, there is a significant amount of correlation between the spatial process variance and the random noise or error variance. It may prove beneficial to define a new term that quantifies the ratio of variances from the spatial process and the noise term. This term is frequently called the *nugget* from geostatistics[7]. It is that element of variance that still exists in the observations as the distance between them approaches zero. The *nugget* is defined as

$$k = \frac{{\sigma_e}^2}{{\sigma_z}^2 + {\sigma_e}^2} \tag{4}$$

The nugget term, k, is the ratio of the variance of the noise term to the overall variance σ^2. The variance of the noise term is then ${\sigma_e}^2 = k\sigma^2$ and the variance of the spatial process is ${\sigma_z}^2 = (1-k)\sigma^2$, the remainder of the overall variance.

2. General Definitions

The locations of a set of n observations of the computer model are defined as $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \Omega$, where Ω is the set of all possible inputs to the model that result in an output, i.e., the domain of the computer model. The resulting outputs are $\mathbf{y} = \{y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_n)\}$. The kriging approach treats $\hat{y}(\mathbf{x})$ as a random function and finds the best linear unbiased predictor, $\boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{x})\,\mathbf{y}$, which minimizes the mean square error of the prediction subject to an unbiasedness constraint.

A few definitions are needed before the solution to the optimization problem subject to the constraint can be given. First, a matrix $\mathbf{F}$ is constructed by evaluating the vector of regressors, $\mathbf{f}(\mathbf{x})$, at each of the n known observations,

$$\mathbf{F} = \{\mathbf{f}(\mathbf{x_1}), \mathbf{f}(\mathbf{x_2}), \ldots, \mathbf{f}(\mathbf{x_n})\}^{\mathrm{T}}. \tag{5}$$

A vector $\mathbf{r}(\mathbf{x})$ that represents the correlation between an unknown point, $\mathbf{x} \in \Omega$, and the n known sample points is defined as:

$$\mathbf{r}(\mathbf{x}) = \{R(\mathbf{x}, \mathbf{x}_1), R(\mathbf{x}, \mathbf{x}_2), \ldots, R(\mathbf{x}, \mathbf{x}_n)\}^{\mathrm{T}}. \tag{6}$$

Given the set of observations of the computer model $\mathbf{X}$, the kth element, $k = 1, \ldots, n$ of the vector function $\mathbf{r}(\mathbf{x})$ is $\mathbf{r}_k(\mathbf{x})$. The correlation matrix $\mathbf{R}$ quantifies the correlation of the observations to themselves. It a square $(n \times n)$ matrix that is symmetric about its diagonal. Its kth row (or column) is defined as $\mathbf{R}_k = \mathbf{r}(\mathbf{X}_k)$ where $\mathbf{X}_k$ is the kth observation in $\mathbf{X}$. The covariance function $\mathbf{v}(\mathbf{x})$ quantifies the covariance of an unobserved location $\mathbf{x} \in \Omega$ to the set of observations $\mathbf{X}$. It scales the correlation function $\mathbf{r}(\mathbf{x})$ with the spatial process variance ${\sigma_z}^2$ and is defined as $\mathbf{v}(\mathbf{x}) = {\sigma_z}^2\mathbf{r}(\mathbf{x}) = (1-k)\sigma^2\mathbf{r}(\mathbf{x})$.

The last element of the kriging model to be defined is the covariance matrix $\mathbf{V}$ that quantifies the covariance of the observations with themselves. The covariance matrix for the observations can then be defined as

$$\mathbf{V} = \sigma^2\left(k\mathbf{I} + (1-k)\mathbf{R}\right) \tag{7}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix. The error or noise term is added to the diagonal terms to properly quantify the covariance of an observation with *itself* as opposed to a second observation at the same location. The variance of a second observation at the same location would have a variance equal to the noise or error variance.

The best linear unbiased predictor of $\hat{y}(\mathbf{x})$ is given by

$$\hat{y}(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\,\hat{\boldsymbol{\beta}} + \mathbf{v}^{\mathrm{T}}(\mathbf{x})\,\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}\right), \tag{8}$$

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

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{F}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}. \tag{9}$$

The first component of Eq. (8) is the generalized least squares estimate of a location in the domain, $\mathbf{x} \in \boldsymbol{\Omega}$, given the covariance matrix, $\mathbf{V}$. The second component of Eq. (8) "pulls" the generalized least squares estimate toward the observed data points, providing a deterministic response surface that may interpolate the observations. The form of Eq. 8 presented here is slightly different than that used most frequently in the development of kriging models for approximating computer experiments. In most presentations, the covariance vector, $\mathbf{v}(\mathbf{x})$, and covariance matrix, $\mathbf{V}$, are simplified to be the correlation vector, $\mathbf{r}(\mathbf{x})$, and correlation matrix, $\mathbf{R}$ because they are both scaled by the same factor, σ^2. Using this form does not require the quantification of the process variance to calculate an estimate of the expected value of the spatial process, but this is no longer the case given the definition of the covariance matrix in Eq. 7.

The BLUP defined in Eq. (8) assumes the correlation parameters $\boldsymbol{\theta}$, used to define the spatial correlation of the observations, are known *a priori*. This is seldom the case and therefore the correlation parameters must also be estimated from the set of observations. The most common process of estimating the best model parameters, maximum likelihood estimation, is covered in the next section.

B. Estimation of Model Parameters

This definition of the nugget effect kriging model requires a means to estimate all of the model parameters needed in the model from a set of observations of the process being estimated by the model. The Maximum Likelihood Estimation (MLE) method is presented in this work. The MLE method assumes that the model paramters being estimated define a random model and as such, the parameters to the models are random variables. It is further assumed that the observations being used to fit the model come from a multivariate Gaussian random process given as $\mathbf{y} \sim \mathcal{N}(\mathbf{F}\beta, \mathbf{V})$.

In general, MLE methods attempt to select model parameter values that maximimize the likelihood of the observations. The likelihood of of the model parameters given the observations is given as

$$L = L\left(\beta, \mathbf{V}|\mathbf{y}\right) = \frac{e^{\frac{1}{2}(\mathbf{y}-\mathbf{F}\beta)^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{F}\beta)}}{(2\pi)^{\frac{1}{2}N}\left|\mathbf{V}\right|^{\frac{1}{2}}}. \tag{10}$$

The likelihood equation can be computationally expensive to evaluate due to the exponential function. The likelihood function is also, typically very near zero throughout the domain of possible model parameters resulting in a difficult function to optimize due to its "flatness". In order to generate more of a gradient the logarithm of the likelihood is often taken and is maximized. The parameter values that result in the maximum likelihood also result in the maximum logarithm of the likelihood so the optimization process can be simplified. The logarithm of the likelihood equation is

$$l = \log L = -\frac{1}{2}N\log 2\pi - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\left(\mathbf{y}-\mathbf{F}\beta\right)^T\mathbf{V}^{-1}\left(\mathbf{y}-\mathbf{F}\beta\right). \tag{11}$$

The optimal values of the trend function coefficients and the total variance can be determined in closed form by differentiating the log-likelihood function with respect to the desired parameter and solving for zero. The MLE optimal values of the trend function are the same as those determined by minimizing the squared error of the estimate from Eq. 9. The values of the best trend function parameters are a function of the spatial covariance parameters $\boldsymbol{\theta}, \sigma^2,$ and k that define the covariance matrix $\mathbf{V}$. By defining a new spatial correlation matrix,

$$\mathbf{R}' = k\mathbf{I} + (1-k)\mathbf{R}, \tag{12}$$

the covariance matrix can be represented as $\mathbf{V} = \sigma^2\mathbf{R}'$. By substituting this expression into Eq. 9, the following expression can be used to estimate the optimal trend function coefficients that is independent of the variance

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{F}^{\mathrm{T}}{\mathbf{R}'_{\boldsymbol{\theta},\mathbf{k}}}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^{\mathrm{T}}{\mathbf{R}'_{\boldsymbol{\theta},\mathbf{k}}}^{-1}\mathbf{y}. \tag{13}$$

where the subscripts of $\mathbf{R}'$ have been added to emphasize its dependence on $\boldsymbol{\theta}$ and k. Using this same definition, the estimate of the variance can also be calculated in closed form as

$$\hat{\sigma}^2 = \frac{1}{n}\left(\mathbf{y}-\mathbf{F}\hat{\boldsymbol{\beta}}\right)^{\mathrm{T}}{\mathbf{R}'_{\boldsymbol{\theta},\mathbf{k}}}^{-1}\left(\mathbf{y}-\mathbf{F}\hat{\boldsymbol{\beta}}\right), \tag{14}$$

With this redefinition of the correlation matrix $\mathbf{R}'$ it is possible to define the BLUP of the process independently of the variance σ^2 as

$$\hat{y}\left(\mathbf{x}\right) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\,\hat{\boldsymbol{\beta}} + (1-k)\mathbf{r}^{\mathrm{T}}(\mathbf{x})\,{\mathbf{R}'}^{-1}\left(\mathbf{y}-\mathbf{F}\hat{\boldsymbol{\beta}}\right). \tag{15}$$

The model parameter values that maximize the log-likelihood equation can be determined by substituting the optimal values of the trend function coefficients and the variance into the log-likelihood equation and maximizing it relative to the remaining covariance parameters. This resulting equation is typically called the profile log-likelihood equation. Even with this reduction in the number of parameters that must be determined via optimization, the process can still be quite computationally expensive. The resulting log-likelihood equation can have large, nearly flat regions, narrow ridges of near-optimal values, and multiple local maxima[8]. Global optimzation methods such as Simulated Annealing and Simplex methods have been used to determine the best parameter values but these tend to be slow, potentially requiring larger numbers of evaluations of the log-likelihood equation. The evaluation of the log-likelihood equation requires the calculation of the interpoint correlation matrix, $\mathbf{R}$, and its inverse.

A modified Newton-Raphson method was previously developed that incorporates the analytical evaluation of the Gradient and Hessian of the log-likelihood function to significantly reduce the number of correlation matrix evaluations and inversions that are requried during the use of finite-differencing the estimate the gradients and Hessians[9]. This previous work did not include the nugget parameter in its derivations.

This work does not use MLE to estimate the optimal nugget parameter. It explores the quality of the resulting metamodels given different nugget parameters. This is part of a development path to determining the best method to first determine if a nugget parameter is beneficial to creating a better metamodel and second what the value of that parameter should be if it is included in the model. As a result, the previous development of a closed form gradient and Hessian need only be slightly modified to include the k parameter. The new correlation matrix, $\mathbf{R}'$ from Eq. 12, can be substituted directly for the correlation matrix $\mathbf{R}$ used in the equations presented previously[9].

III. Discussion of Formulation

The form of the kriging model suggested here to use as a metamodel, a model of a model, is the *universal* kriging model with the added *nugget* parameter. The more typically used universal kriging model with observations of computer analyses is the interpolating type that assume s there is no nugget effect, i.e. $k = 0$. By assuming that the kriging model has no trend component other than a constant to remove the bias in the residuals, a requirement for a BLUP (unbiased), the ordinary kriging model results. The ordinary kriging model is often used in preference to the universal kriging model because of its flexibility in reproducing complex surfaces with a minimal number of observations and the relative simplicity of its form. In some instances it may be the best model form to reproduce the given observations.

The real desire is to develop a robust metamodel approach that can adequately deal with the situation of a small number of observations, a large numebr of observations, and the large number of input dimensions. Kriging is typically limited to 10 input dimensions and 100 observations. Beyond this, the model can become unstable and computationally expensive to estimate the required model parameters. The given form of the universal kriging model with the nugget effect provides the user with the opportunity to have a much more robust modeling system. Through the use of the nugget parameter, the universal kriging model can be smoothly adjusted from a fully parametric trend model in which there is no quantified spatial correlation ($k = 1$) to the full interpolating model in which the there exists a combination of important trend parameters and spatial correlation parameters on all of the input dimensions ($k = 0$). The spatial correlation portion of the model can be taken to its extreme with the ordinary kriging model where there is only a constant trend parameter and all of the observed variations are due to spatial correlations.

By not requiring the generated response surface to interpolate all of the observations, it is possible to create a kriging model that does not include all of the input dimensions. This can greatly simplify the model building process as well as reduce the number of observations required to make a *useful* model. The difficulty in this approach is to determine the dimensions that should be included in the model. A second difficulty lies in the fact that by decreasing the importance of the spatial correlation model, i.e. larger values of k, it can become more difficult to estimate the spatial correlation model parameters. It is for this reason that using just MLE to estimate all of the model parameters, including the nugget parameter, is not performed in this work. This provides the opportunity to understand some of the relationships that exist between the nugget parameter and the other parameters in the model.

IV. Demonstration of Model Form

A. Procedure for Comparisons

This section details the procedure used to investigate the effectiveness of the nugget parameter in creating a better quality surrogate model. The exploration of this resulting tradespace is achieved by varying the nugget parameter over the range of $[0, 0.6]$ out of the possible range of $[0, 1]$. Values greater than 0.6 were deemed to not be important and as a result are not included in the analysis. For each given value of the nugget parameter, the optimal kriging model was created. The creation of the optimal kriging model is a two loop process [14]. The outer loop selects a potential form for the trend function and the inner loop uses MLE to select the best model parameters given that form. The criterion used to determine the best model is the correct AIC metric. This metric is a parsimonious information-based criterion [15]. Starting with a second-order with interaction terms trend model, the trend parameter terms are removed from the model one at a time based upon the probability that they are part of the model. This is a function of the absolute value of the parameter estimate divided by the standard deviation of that estimate.

One of the most important qualities of a metamodel is its accuracy at estimating the original model. This can be best estimated by evaluating the original model at a large number of locations within its domain and comparing those results to that estimated by the metamodel. For the four example functions, 1000 uniformly random locations are evaluated and a non-dimensional model quality metric, ${R^2}_{actual}$, is calculated. This metric is defined as

$${R^2}_{actual} = 1 - \frac{\text{error variance}}{\text{total actual variance}}. \tag{16}$$

The difficulty with evaluating this metric of metamodel quality is the computational costs of its calculation renders it infeasible in most cases.

There are other metrics for metamodel quality that do not require the evaluation of a large number of validation points. The familiar coefficient of determination, R^2, does not apply for interpolating kriging models since it fits all of the observed data perfectly. As a result, the PRedicted Error Sum of Squares (PRESS) is used. With PRESS, each observation is left out of the model in turn and the resulting error at the left out observations is used to calculate the statistic. This is often called *leave-one-out* cross-validation. The form of the kriging model permits the very fast evaluation of this statistic. The PRESS provides a sum of the leave-one-out squared error. This sum can be used with the variance of the observations to calculate the ${R^2}_{prediction}$ metric [15]. The definition of the ${R^2}_{prediction}$ metric is very similar to that in Eq. 16. This ${R^2}_{prediction}$ metric is compared to ${R^2}_{actual}$ in the results of this section for different values of the nugget parameter for the four models.

Two additional model properties are also presented to aid in the analysis of the results. The first is the resulting log-likelihood calculated for each optimal model given the nugget parameter. The second is the product of the two spatial correlation parameters. The log-likelihood is important since the optimal model parameters are selected by maximizing this metric. As a result, it may be expected that the best quality model should occur at the maximum value of the log-likelihood function as a function of the nugget parameter. This is investigated in the results.

The rational for including the second model property lies in the form of the spatial correlation model. One of the main impacts of including the nugget parameter is the modification of the spatial correlation matrix. Larger values of the nugget parameter reduce the values of the off-diagonal terms and as a result, the influence of the spatial correlation model. In order to offset these changes, the values of the spatial correlations are expected to increase. Since the product correlation rule is used to combine the spatial correlations of multiple dimensions, a product of the optimal spatial correlation range parameters is also explored. The spatial correlation parameter values are directly related to the effective spatial correlation range in each dimension. Their product represents the area of correlation in two dimensions and can represent the overall spatial correlation in higher dimensions.

B. Example Functions

Four two-dimensional example functions are used to demonstrate the use of a non-interpolating kriging model form to approximate the original models. The first 2-D example function is highly non-linear in one dimension and is linear in the second dimension. This was introduced by Osio and Amon [10] and used by Jin et al. [11] as a test problem for sequential experimentation. The first function is defined over the domain, $\mathbf{x} \in [0,1]^2$, and is given by

$$y(\mathbf{x}) = \cos(6(x_1 - 0.5)) + 3.1(|x_1 - 0.7|) + 2(x_1 - 0.5) + \sin\left(\frac{1}{|x_1 - 0.5| + 0.31}\right) + 0.5x_2. \tag{17}$$

The second 2-D example function is a "mystery" multi-modal function in two dimensions, $\mathbf{x} \in [0,5]^2$, that comes from Sasena [12]

$$y(\mathbf{x}) = 2 + 0.01\left(x_2 - x_1^2\right)^2 + (1 - x_1)^2 + 2(2 - x_2)^2 + 7\sin(0.5x_1)\sin(0.7x_1x_2) \tag{18}$$

The third 2-D example function is the Branin test function [11], $x_1 \in [-5, 10], x_2 \in [0, 15]$, which has less oscillation over its domain than 2D-MF; it is defined as

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

The last 2-D example function is the six-hump camelback function [12], $x_1 \in [-2, 2], x_2 \in [-1, 1]$, which is defined as

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

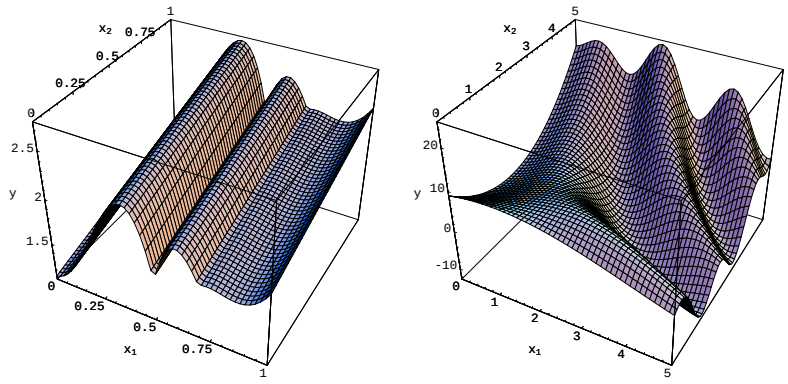
For all of these 2-D functions, the same set of 21 sample points is used to fit each kriging model. These points are generated using a Latin Hypercube [13] (see Figure 2). The domain of each example is mapped to $[0,1]^2$. A set of 1000 randomly sampled validation points is used to evaluate the quality of the created metamodels.

C. Results of Comparisons

The results of creating optimal kriging models for the four example functions while varying the nugget are parameter are presented. The details of each example function is presented first and the section will close with some general observations that can be made based on these for examples. The values of the nugget parameters were sampled at every 0.01 for the range of $[0, 0.2]$ and were sampled every 0.05 for the range $[0.2, 0.6]$. The actual and predicted R^2 are plotted first in the series of three plots for each example problem. This is followed by the resulting log-likelihood of the optimal model. The last plot in each case is the correlation "area" plot, i.e. a plot of the product of the two spatial correlation range parameters.

The Osio and Amon function is characterized as being a linear function in the x_2 dimension and being a periodic and abruptly changing function in the x_1 dimension. The nonlinearities in the orginal function provide a unique opportunity for the kriging model to demonstrate its capabilites. With the 21 point Design of Experiment (DoE), the kriging model is able to reproduce the original function very well.

The most striking result in this example (see Fig. 3) is the apparent discontinuity in all of the plots between the nugget values of 0.06 and 0.07. This can be attributed to a change in the optimal trend model form. For values of the nugget less than 0.06, the trend model form includes only a constant term and a linear term for x_2. For nugget values greater than 0.06 the trend model changes to additionally include both linear and quadratic terms for x_1. This change may be attributed to the increased dependence on the trend model part of the kriging model to represent the observed variability over the spatial correlation model.



(a) Osio & Amon Function

(b) ”Mystery” Function

(c) Branin Function

(d) Six-Hump Camelback Function

Figure 1: Plots of the Four 2-D Test Functions

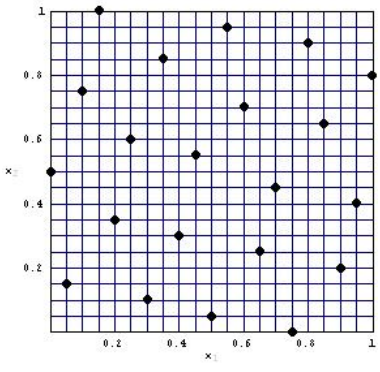
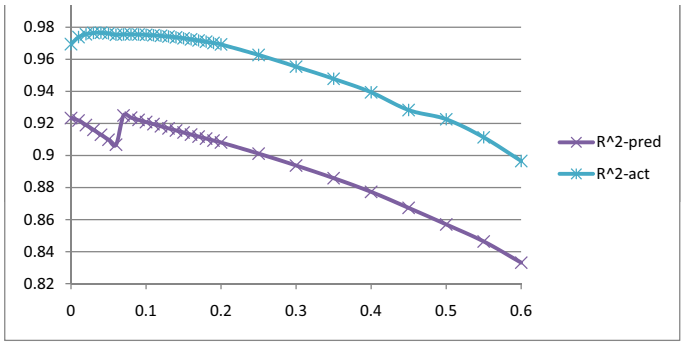
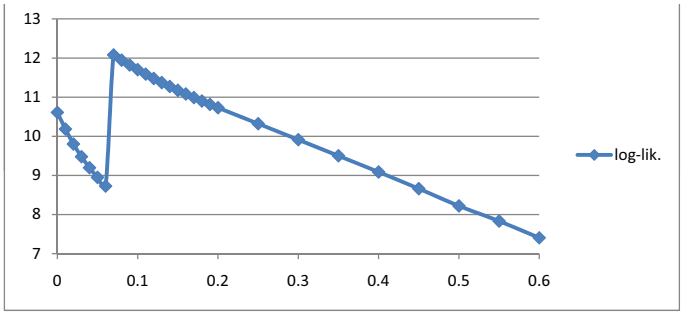


Figure 2: 2-D Latin Hypercube Sampled Points

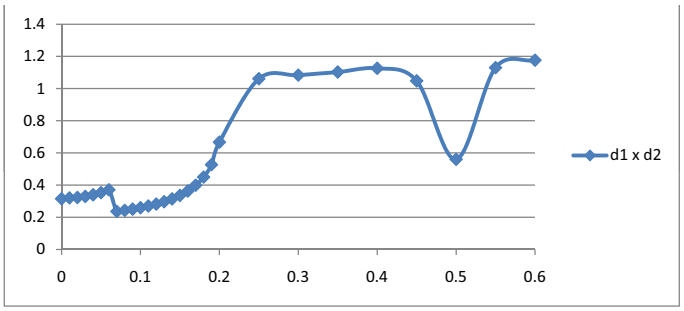
The best model, as defined by $R^2{}_{actual}$, occurs with a nugget value of 0.04, though there is very little difference for values between 0.02 and 0.13. It is interesting how the change in the trend model form does not significantly change the metamodel quality. The log-likelihood for each of the optimal models tends to decrease as the nugget parameter increases other than for the change in trend model form. The jump to a lower log-likelihood on the left of the plot is driven by the penalty applied in the corrected AIC metric for the additional model parameters for the model form selected by the larger values of the nugget parameter. The last thing to notice is the general increase in optimal correlation range product with increasing values of the nugget parameter. The variation in the correlation area plot seen at a nugget value of 0.5 may be due to the lack of importance in the correlation model for this relatively large nugget parameter value, i.e. it may just be the result of some "noise" and should not be seen as an important result.



(a) R^2 Plot



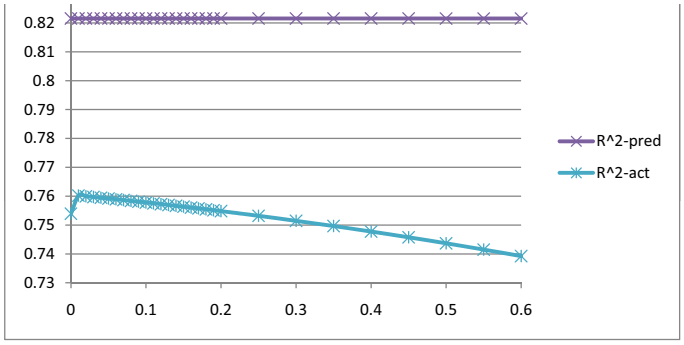
(b) Log-Likelihood Plot



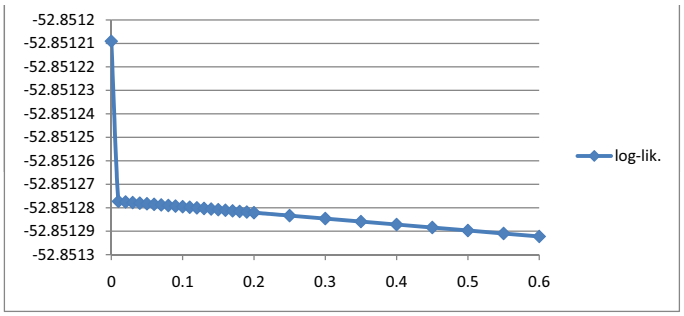
(c) Correlation Area Plot

Figure 3: Osio and Amon Function Results

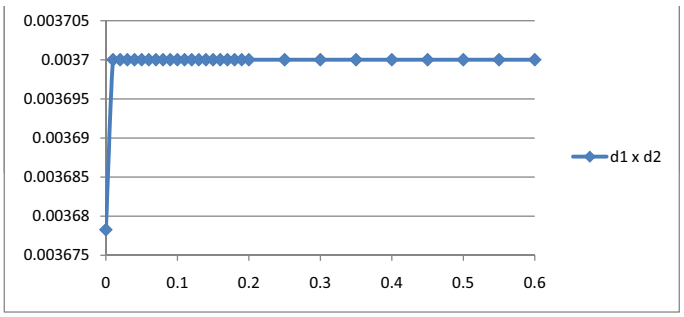
The Mystery function (see the plots in Fig. 4) is very difficult to fit with nearly any metamodel form, including the kriging model, given its cyclic pattern that is a function of x_1 and x_2 combined. By using only 21 observations to fit the model, it is difficult to adequately capture the response surface of the original function. For any of the created models the $R^2{}_{predicted}$ is nearly constant. The $R^2{}_{actual}$ makes a slight jump once the kriging model is no longer required to interpolate all of the observations. It get progressively worse as the nugget parameter is increased. The log-likelihood stays nearly constant throughout. The small descrease for larger nugget values is only noticeable because of the expanded scale of the plot. The same can also be said for the correlation area plot; it is nearly constant across the different values of the nugget parameter. Given the poor quality of any of the created models, few conclusions can be drawn from the results of this example function.



(a) R^2 Plot



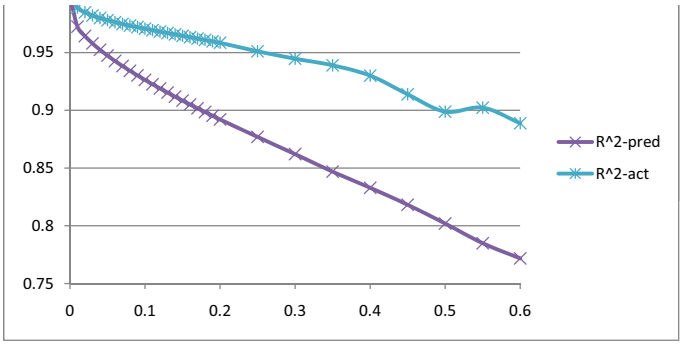
(b) Log-Likelihood Plot



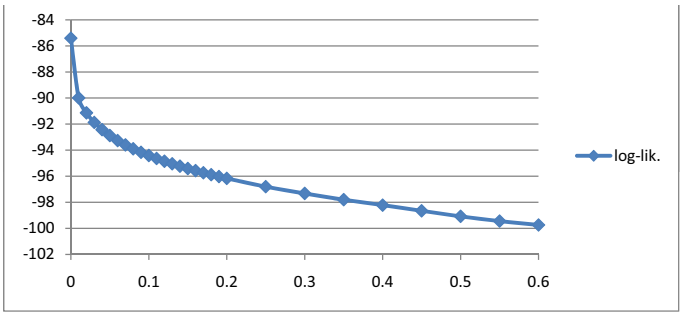
(c) Correlation Area Plot

Figure 4: Mystery Function Results

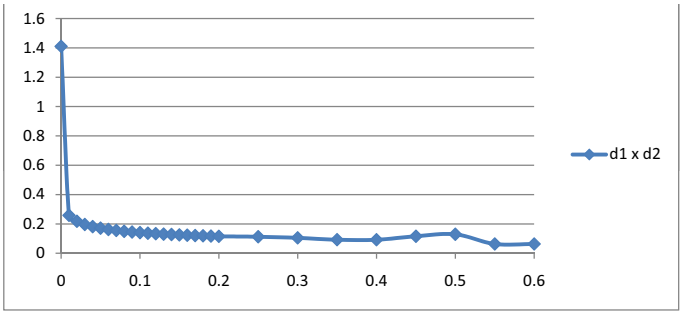
The Branin function example (see Fig. 5) behaves a bit differently than the rest of the examples. For this example, the best model results from the interpolating model, i.e. a nugget parameter value of 0.0. In this example, both the $R^2{}_{predicted}$ and $R^2{}_{actual}$ are nearly 1.0 for the interpolating model and fall off as the value of the nugget parameter increases. The same is true for the log-likelihood value; it decreases for increasing values of the nugget parameter. The last observation to make is the resulting decrease in the correlation area for increasing values of the nugget parameter. This seems to be different than the expected increase in the correlation area with increasing nugget parameter values.



(a) R^2 Plot



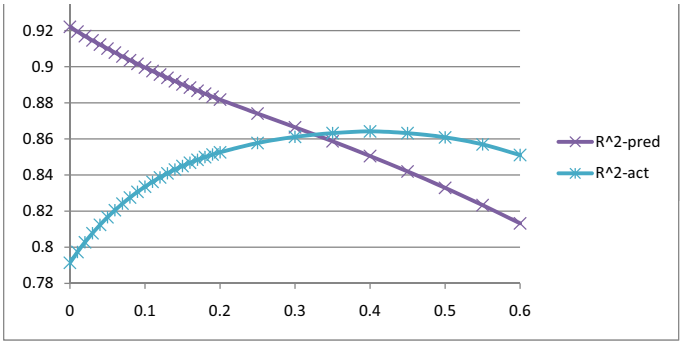
(b) Log-Likelihood Plot



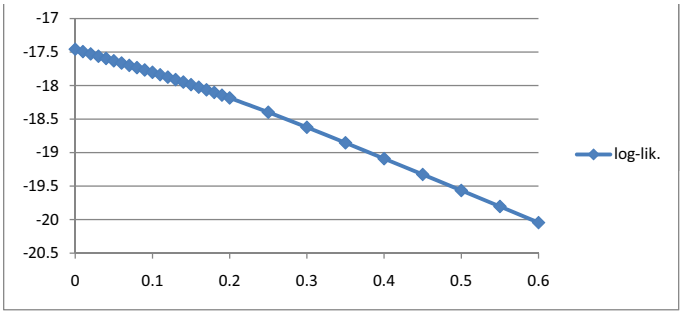
(c) Correlation Area Plot

Figure 5: Branin Function Results

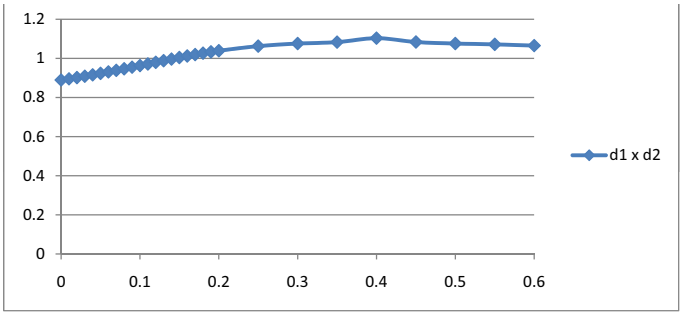
The results of the last example, the Six-Hump Camelback function, are shown in Fig. 6. The resulting response surface of this function is multi-modal with six local minima. The optimal kriging model created with 21 observations results in a marginal approximation to the original function. In this example, like the previous ones, the log-likelihood and the $R^2{}_{predicted}$ of the optimal kriging model decrease as the nugget parameter increases. The maximum value of the $R^2{}_{actual}$ occurs at a nugget parameter value of 0.4. The correlation area appears to have its maximum at a nugget parameter value of 0.4. This example is different than the previous ones in that the best kriging model results from a relatively large value of the nugget parameter.



(a) R^2 Plot



(b) Log-Likelihood Plot



(c) Correlation Area Plot

Figure 6: Six Hump Camelback Function Results

V. Conclusions and Future Work

The addition of a nugget parameter to the kriging model that removes the requirement to interpolate all of the result is presented in this paper. This addition is originally from field of geostatistics and is not widely used in multidisciplinary design. In geostatistics, the nugget is typically estimated graphically through the use of a variogram. This work investigates the appropriateness of using of MLE to estimate the optimal nugget parameter. MLE is an objective and most common method used to estimate kriging model in the are of Design and Analysis of Computer Experiments (DACE).

It was found, by analyzing the results of creating kriging models for four 2-D functions, that the initial belief that Maximum Likelihood Estimation is the best method to etimate the nugget parameter value is not correct. Additionally, using the estimated model quality metric, $R^2{}_{predicted}$ does not appear to be a good method to estimate the best nugget parameter value since it tends to follow the same trends as the resulting log-likelihood of the optimal kriging models given the nugget parameter. The one nearly consistent observation is the best kriging model results when the correlation area was maximized. The one exception to this observation exists with the first example problem. The model with the maximum correlation area for the trend model form selected for the interpolating model is still the best model, but the absolute maximum correlation area is not the best model.

Future research should look into the ramifications of creating non-interpolating kriging models with larger numbers of observations and larger numbers of input dimensions. It is expected that the non-interpolating kriging models will reduce some of the computational difficulties often seen when inverting the resulting spatial correlation matrices. As a follow-on, the uncertainty of the spatial correlation parameter should also be estimated by evaluating the diagonal terms of the Hessian of the log-likelihood function, values that are already calculated during the MLE parameter estimation process. For those dimensions in which the uncertainty is greatest, an alternative model form, that doesn't include that specific input dimension, can be constructed.

VI. Acknowledgment

The author would like to thank Dr. Scott Hassan of ONR 333 for his generous support on Contract No. N00014-10-G-0259.

References

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

[2]Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P., 1989, "Design and Analysis of Computer Experiments", *Statistical Science,* **4**(4), pp. 409–435.

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

[4]Schonlau, M., Welch, W. J. and Jones, D. R., 1998, "Global Versus Local Search in Constrained Optimization of Computer Models", In *New Developments and Applications in Experimental Design*, N. Flournoy, W. F. Rosenberger, and W. K. Wong, eds., Institute of Mathematical Statistics, pp. 11–25.

[5]Sasena, M. J., Papalambros, P. and Goovaerts, P., 2002, "Exploration of Metamodeling Sampling Criteria for Constrained Global Optimization", *Engineering Optimization,* **34**(3), pp. 263–278.

[6]Huang, D., Allen, T. T., Notz, W. I. and N., Z., 2006, "Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models", *Journal of Global Optimization,* **34**(3), pp. 441–466.

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

[8]Mardia, K. and Marshall, R., 1984, "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression", *Biometrika,* **71**(1), pp. 135–146.

[9]Martin, J. D., 2009, "Computational Improvements to Estimating Kriging Metamodel Parameters", *ASME Journal of Mechanical Design,* **131**(8).

[10]Osio, I. G. and Amon, C. H., 1996, "An Engineering Design Methodology with Multistage Bayesian Surrogate and Optimal Sampling", *Research in Engineering Design,* **8**(4), pp. 189–206.

[11]Jin, R., Chen, W. and Sudjianto, A., 2002, "On Sequential Sampling for Global Metamodeling in Engineering Design", In *2002 ASME Design Engineering Technical Conference*, DETC2002/DAC-34092.

[12]Sasena, M. J., 2002, "Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations". *Ph.D. Dissertation*, University of Michigan, Ann Arbor, MI.

[13]McKay, M., Conover, W. and Beckman, R., 1979, "A Comparison of Three Methods for Selecting Values of Input in the Analysis of Output from a Computer Code", *Technometrics,* **21**(1), pp. 239–245.

[14]Martin, J. D. and Simpson, T. W., 2006, "A Methodology to Manage Uncertainty During System-Level Conceptual Design", *ASME Journal of Mechanical Design,* **128**(4), pp. 959–968.

[15]Martin, J. D. and Simpson, T. W., 2005, "On the Use of Kriging Models to Approximate Deterministic Computer Models", *AIAA Journal,* **43**(4), pp. 853–863.

[16]Joseph, V. R., Hung, Y. and Sudjianto, A., 2008, "Blind Kriging: A New Method for Developing Metamodels", *ASME Journal of Mechanical Design,* **130**(3).