Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

AN ADDITIVE GLOBAL AND LOCAL GAUSSIAN PROCESS MODEL FOR LARGE DATA SETS

Qun Meng

Department of Industrial andSystems Engineering
National University of Singapore
1 Engineering Drive 2 117576, SINGAPORE

Szu Hui Ng

Department of Industrial andSystems Engineering
National University of Singapore
1 Engineering Drive 2 117576, SINGAPORE

ABSTRACT

Many computer models of large complex systems are time consuming to experiment on. Even when surrogate models are developed to approximate the computer models, estimating an appropriate surrogate model can still be computationally challenging. In this article, we propose an Additive Global and Local Gaussian Process (AGLGP) model as a flexible surrogate for stochastic computer models. This model attempts to capture the overall global spatial trend and the local trends of the responses separately. The proposed additive structure reduces the computational complexity in model fitting, and allows for more efficient predictions with large data sets. We show that this metamodel form is effective in modelling various complicated stochastic model forms.

1 INTRODUCTION

In practice, simulation models are widely used to provide an effective and efficient way to evaluate the behavior of real systems. However with the stochastic and complex nature of most real systems, the simulation models can be time consuming to execute. An alternative is to create a statistical model of the simulation model, which is known as metamodel.

A metamodel is a simplification of a simulation model. Simpson et al. (2001) reviewed the metamodel application in engineering. Li et al. (2010) also provided a comprehensive comparison of metamodeling approaches that can also be well applied in simulation optimization. Among all these metamodels, the Gaussian Process model, also known as the kriging model (Cressie 1993), has been increasingly popular in recent years due to its adaptability and efficiency to model the computer outputs. More importantly, it can provide a unique statistical view of the prediction error, which makes it more useful in simulation optimization. Beyond the deterministic computers experiments, it has also been widely used in the stochastic simulation through stochastic kriging model (Ankenman et al. 2010) or the modified nugget effect model (Yin et al. 2011).

However, estimating the Gaussian Process model is a computational challenge when the data sets are large. Given the data size of N, estimating the model parameters with traditional methods like the maximum likelihood estimation and estimating the model predictors involve the inversion of a $N \times N$ covariance matrix, which typically requires $O(N^3)$ operations and $O(N^2)$ memory. This becomes computational intractable for a large N. As such, a desktop computer is unable to handle data sizes larger than several hundreds.

Many approximation approaches have been proposed to solve the computational problem with large data sets. Existing approximation methods may be divided into three categories: global approximation, localized regression and combination of global approximation and localized regression. The global approximation methods include *rank reduction* and *sparse approximation* (Quiñonero Candela and Rasmussen 2005, Banerjee et al. 2008). However these global approximation methods typically capture only the long lengthscale global trend of the spatial processes, leaving much of the local dependencies unexplained.

978-1-4673-9743-8/15/$31.00 ©2015 IEEE

The second category is localized regression, where model predictions are estimated based on local neighborhoods. Local Kriging fits different Kriging models in different subregions independently. Local Kriging is known for its adaptability to model nonstationary process and its efficiency in computation. However it suffers from the discontinuities at the boundaries due to its localized independent model estimation. Park et al. (2011) proposed an approach that smooths the discontinuities by adding equality constraints at the boundaries of neighboring subregions, but the additional computational time is required to estimate the boundary values. Another local approximation approach is to apply covariance tapering, which assumes that the distant pairs of observations are uncorrelated (Furrer, Genton, and Nychka 2006). Sparse matrix algorithm can then be applied to realize the computational efficiency. However, such approaches are unable to effectively capture the long lengthscale dependencies, missing often the larger global trend. Recent works by Gramacy and Apley (2014) and Gramacy and Haaland (2015) propose splitting the input domain into different segments where the parameters are estimated separately, enabling parallelization in the model estimation. However accomplishing the massive parallelization may still require large amounts of computation.

The last category combines the global approximation and the localized regression to overcome the disadvantages of the individual methods. A full scale approximation (FSA) of the covariance functions proposed by Sang and Huang (2012) approximates the covariance function through a combination of a reduced rank approximation and a tapered residual approximation. This attempts to capture both the long lengthscale dependence and shorter lengthscale dependence. The partially independent conditional approximation (PIC) approaches (Snelson and Ghahramani 2007) also successfully combines a reduced rank approximation and a locally independent residual approximation.

In this paper, we leverage on the benefits of combined approach and develop an Additive Global and Local Gaussian Process (AGLGP) model that facilitates computations of large data sets. It incorporates a global GP model and a piecewise local GP model into an additive GP model with a composite covariance structure. The local model allows for different correlation structures in different subregions, which is more flexible and better able to capture the nonstationary process with more numerical stability. The basic idea behind the AGLGP is to build a global model with a small set of inducing points to capture the global trend and build a local model to capture the residual process from the global model. Our central contribution is to develop a model that is computationally efficient and can capture the nonstationarity with this additive structure. In addition to the model prediction, the AGLGP can be potentially applied in optimization.

The rest of the paper is organized as follows. Section 2 introduces the background. Section 3 presents the new AGLGP model, its predictive distribution, discusses the parameter estimation based on the derived covariance structure and illustrates the mechanisms to fit the model. In Section 4 a numerical study is conducted to demonstrate the application of the new model and a comparison is made with other approximation models. Finally in Section 5, conclusions and future areas of development are discussed.

2 BASICS AND NOTATIONS

2.1 Kriging

The Kriging model is popularly used for approximating various highly flexible and nonlinear functional forms. For deterministic computer experiments, suppose we run the simulation at n inputs $\mathbf{x} = (x_1, x_2, ..., x_n)$ and obtain corresponding simulation outputs $\mathbf{y} = (y(x_1), y(x_2), ..., y(x_n))$. We assume that simulation outputs can be modeled as a Gaussian Process with mean $\mu(x)$ and covariance function $R(\cdot)$, *i.e.*, $y(x) = f(x)$, $f(x) \sim GP(\mu(x), \sigma^2 R(\cdot))$. Given the parameters, the kriging predictor and the mean square error (MSE) at an input x^* are given as

$$\hat{y}(x^*) = \mu(x^*) + \mathbf{r}'\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \tag{1}$$

$$\hat{s}^2(x^*) = \sigma^2 \left[1 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{r} + \frac{(1 - \mathbf{1}'\mathbf{R}^{-1}\mathbf{r})^2}{\mathbf{1}'\mathbf{R}^{-1}\mathbf{1}} \right] \tag{2}$$

where $\mathbf{r} = (R(d_{*,1}), R(d_{*,2}), ..., R(d_{*,n}))$ is the covariance between the point to be estimated and n observed points, and $d_{*,i}$ is the Euclidean distance between point x^* and x_i; $\mathbf{R}$ is $n \times n$ covariance matrix across all design points. The covariance function $R(\cdot)$ can have different forms and usually depends on a correlation parameter θ. A popular choice is the exponential family of correlation functions, which is commonly applied in computer experiments for its smooth characteristics, $R(d_{i,j}) = R[f(x_i), f(x_j)] = \sigma^2 exp[-\sum_{h=1}^{k} \theta_h |x_{ih} - x_{jh}|^{p_h}]$, where σ^2 is the variance of $f(x)$ of all x, θ_h control how fast the correlation decayes with the distance in the hth dimension. Stein (1991) discusses various families of correlation functions and their properties, including isotropic and anisotropic functions. When the mean function takes a constant form $\mu(x) = \mu$, it is known as the ordinary kriging model. This form is shown to be adequate in many applications (Ankenman et al. 2010, Sacks et al. 1989, Jones et al. 1998).

2.2 Stochastic Kriging

Stochastic Kriging extends the kriging model applied for the deterministic computer experiments to stochastic computer experiments. In the stochastic case, we assume that the simulation output $y(x)$ are realizations of a random process that can be described by the model

$$y(x) = f(x) + \varepsilon(x) \tag{3}$$

where $f(x)$ is the mean of the process, and $\varepsilon(x)$ is the random noise of the process. Here we assume $\varepsilon(x) \sim N(0, \sigma_\varepsilon^2(x))$, and are independently distributed across simulations. Furthermore, $\varepsilon(x)$ is assumed independent of $f(x)$. The error variances $\sigma_\varepsilon^2(x)$ may depend on x. The MSE-optimal predictor can be shown to be

$$\hat{y}(x^*) = \mu + \mathbf{r}'(\mathbf{R} + \mathbf{\Sigma}_\varepsilon)^{-1}(\mathbf{y} - \mathbf{1}'\mu) \tag{4}$$

where $\mathbf{\Sigma}_\varepsilon = diag(\sigma_\varepsilon^2(x_1), ...\sigma_\varepsilon^2(x_n))$. The optimal MSE is then given by:

$$\hat{s}^2(x^*) = E(f(x^*) - \hat{y}(x^*))^2 = R(d_{*,*}) - \mathbf{r}'(\mathbf{R} + \mathbf{\Sigma}_\varepsilon)^{-1}\mathbf{r} \tag{5}$$

This modeling form is known as stochastic kriging (Ankenman et al. 2010), and a similar form known as the Modified Nugget Effect Kriging can be found in Yin et al. (2011).

2.3 Composite Gaussian Process Model

The composite Gaussian Process model (CGP) (Ba and Joseph 2012) proposes a modeling approach by incorporating a flexible global trend into the GP model. This is useful when the second-order stationarity assumption for the traditional kriging model is violated. It replaces the polynomial mean model $\mu(x)$ in the universal kriging model by another GP model and introduces a local model for the local adjustment

$$y(x) = z_{global}(x) + z_{local}(x), \quad z_{global}(x) \sim GP(\mu, \tau^2 g(\cdot)), \quad z_{local}(x) \sim GP(0, \sigma^2(x) l(\cdot)) \tag{6}$$

The first GP $z_{global}(x)$ has a mean μ, variance τ^2 and correlation structure $g(\cdot)$. The second GP $z_{global}(x)$ has a mean 0, variance $\sigma^2(x)$ and a correlation structure $l(\cdot)$. Both $g(\cdot)$ and $l(\cdot)$ are specified to be Gaussian correlation functions. Overall, the model is equivalent to assuming the response $y(x) \sim GP(\mu, \tau^2 g(\cdot) + \sigma^2(x) l(\cdot))$. Suppose $\sigma^2(x)$ can be further expressed as $\sigma^2(x) = \sigma^2 v(x)$, then the best linear unbiased prediction can be shown to be

$$\hat{y}(x^*) = \mu + (\mathbf{g}(x^*) + \lambda v^{1/2}(x^*)\mathbf{\Sigma}^{1/2}\mathbf{l}(x^*))'(\mathbf{G} + \lambda \mathbf{\Sigma}^{\mathbf{1/2}}\mathbf{L}\mathbf{\Sigma}^{\mathbf{1/2}})^{\mathbf{-1}}(\mathbf{y} - \mathbf{1}'\mu) \tag{7}$$

where $\mathbf{g}(x^*) = (g(x^*, x_1), ..., g(x^*, x_n))$, $\mathbf{l}(x) = (l(x^*, x_1), ..., l(x^*, x_n))$. $\mathbf{G}, \mathbf{L}$ are $n \times n$ correlation matrix across all design points with specified correlation function of $g(\cdot)$ and $l(\cdot)$. $\lambda = \sigma^2/\tau^2$ and $\mathbf{\Sigma} = diag(v(x_1), ..., v(x_n))$ is denoted to represent the local variances at each design point.

3 AN ADDITIVE GLOBAL AND LOCAL GAUSSIAN PROCESS MODEL

To estimate the Gaussian Process model with large data sets, approximations are typically required. Considering the disadvantages of the discontinuities in localized kriging model, a combination of a global model and a local model is proposed here. Specifically, the proposed AGLGP model incorporates a global model to capture the overall global trend and a local model to capture the residual process in the local neighborhoods.

Consider a simple example function of $y(x) = \cos(100(x-0.2))e^{2x} + 7\sin(10x)$ in Figure 1. The function exhibits both a long lengthscale trend and a short lengthscale trend. To capture these, the global model is developed to capture the long lengthscale global trend $y_{global} = 7\sin(10x)$ and the local model is to capture the residual process (of shorter lengthscale), mainly $y_{local} = \cos(100(x-0.2))e^{2x}$. To address the computational constraints of the data sets, we propose to capture the global trend by a smaller number of well placed inducing points, and capture the residual process with a piecewise local model.

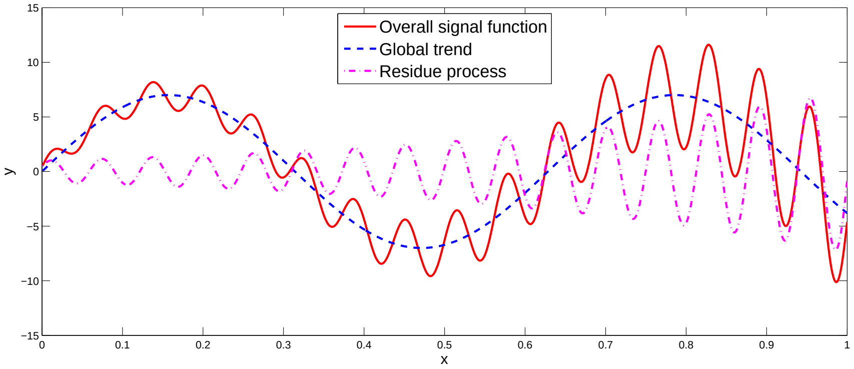

Figure 1: Decomposition of the signal function

The purpose of the inducing points is to sufficiently summarize the observed data points and smooth out the local fluctuations to highlight the global trend. To determine the inducing points, clusters are generated in such a way that the points within a cluster are close in terms of their distance x and response y.

The residuals from the global trend are then modeled by another piecewise GP. To fit the local model, the overall region is partitioned into local regions with a specific correlation structure. With appropriately selected regions, the residual process is expected to be more homogeneous within a local region, and the independent assumption is made across regions. Such *local kriging* models have been applied to model nonstationary process, but can be highly discontinuous across region as previously described. Our overall model combines both the global and local model that aims to capture both the global trend and local residuals, and the additive combination reduces the discontinuities of local kriging models. In the following subsections, we will describe the model development for the deterministic computer experiments (section 3.1) and the stochastic computer experiments (section 3.2).

3.1 Deterministic Additive Global and Local Gaussian Process Model

For the deterministic computer experiments, the AGLGP models the response by an addition of a global and a local model.

$$\begin{gathered} y(x) = f_{global}(x) + f_{local}(x), \quad f_{global}(x) \sim GP(\beta_0, g(\cdot)), \quad f_{local}(x) \sim GP(0, l(\cdot)) \\ g(x_i - x_j) = \sigma^2 r_g(x_i - x_j, \boldsymbol{\theta}), \quad l(x_i - x_j) = \boldsymbol{\tau}^2 r_l(x_i - x_j, \boldsymbol{\alpha}) \end{gathered} \tag{8}$$

where $f_{global}(x)$ models the global trend that is captured through a small set of inducing points, and $f_{local}(x)$ models the residual process that is unexplained by $f_{global}(x)$. We assume $f_{global}(x)$ can be modeled by a deterministic GP model with a constant mean β_0 and variance of σ^2, while $f_{local}(x)$ can be modeled by another zero mean GP to capture the detrended residual process, which is the bias between the signal function and the global trend. $f_{global}(x)$ and $f_{local}(x)$ are assumed to be independent. This is similar to the assumption made in Composite Gaussian Process model (Ba and Joseph 2012). σ^2 and $\boldsymbol{\tau}^\mathbf{2}$ are the variance of the global and local model and $r_g(\cdot)$ and $r_l(\cdot)$ are the correlation structure of the individual processes. In this paper, we assume the *Gaussian correlation function*. Here $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ represent the sensitivity parameters of the correlation functions. The larger the parameters, the lower the correlation with respect to distance. As we expect the global model to capture the global trend while the local model to capture the residual details, it is reasonable to add constraints on the unknown correlation parameters $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ to satisfy $\mathbf{0} \leq \boldsymbol{\theta} \leq \boldsymbol{\alpha}$. As we assume independence across local regions for the local residuals, the covariance function $l(x_i - x_j)$ of the local model will be a piecewise function. Given k local regions, $r_1, ...r_k$, the local covariance structure can be expressed by $l(x_i - x_j) = \sum_{r=1}^{k} H(x_i, x_j) l_r(x_i - x_j)$

$$H(x_i, x_j) = \begin{cases} 1, x_i \in r_p, x_j \in r_q, p = q \\ 0, x_i \in r_p, x_j \in r_q, p \neq q \end{cases}$$

$l_r(\mathbf{x}_i - \mathbf{x}_j)$ is the specific covariance structure in local region $r = 1, ..., k$.

We define a set of inducing points $\mathbf{x}_g = (x_g^1, x_g^2, ..., x_g^m)$ in m dimensions, where $m \ll n$, with output of $\mathbf{y}_g = (y_g^1, y_g^2, ..., y_g^m)$. $\mathbf{y}_g$ is the realization of $f_{global}(x)$, which is a latent process. We first assume $\mathbf{y}_g$ is known, with given parameters $\beta_0, \boldsymbol{\theta}, \boldsymbol{\alpha}, \sigma^2, \boldsymbol{\tau}^\mathbf{2}$. The best linear unbiased global predictor can then be written as

$$\widehat{y}_{global}(x^*) = \beta_0 + \mathbf{g}' \mathbf{G_m^{-1}} (\mathbf{y}_g - \mathbf{1}' \beta_0) \tag{9}$$

where $\mathbf{g} = (g(x^* - x_g^1), ...g(x^* - x_g^m))$, $\mathbf{G_m}$ is $m \times m$ covariance matrix with $ijth$ element $g(x_g^i - x_g^j)$. The global predictor interpolates $\mathbf{y}_g$ since $\widehat{y}_{global}(\mathbf{x}_g^j) = \beta_0 + \mathbf{e_i'}(\mathbf{y}_g - \mathbf{1}'\beta_0) = y_g^i$, where $\mathbf{e_i'}$ is the i-th unit vector. With the fitted global model, we have $\widehat{\mathbf{y}}_{global} = (\widehat{y}_{global}(x_1), ..., \widehat{y}_{global}(x_n))$. The residuals are then obtained by $\mathbf{y_l} = \mathbf{y} - \widehat{\mathbf{y}}_{global}$. We assume that the residual process is correlated within a local region while independent across local regions, so different correlation functions are allowed in different local regions. This enables flexibility to capture nonstationarity in the process. The local predictor is given by

$$\widehat{y}_{local}(\mathbf{x}^*) = \mathbf{l}' \mathbf{L_n^{-1}} \mathbf{y_l} \tag{10}$$

where $\mathbf{l} = (l(x^* - x_1), ...l(x^* - x_n))$ and $\mathbf{L_n}$ is covariance matrix with (ij) element $l(x_i - x_j)$, $\mathbf{L_n}$ is a block diagonal matrix which can be expressed by $\mathbf{L_n} = diag(\mathbf{L_1}, ..., \mathbf{L_r}, ...\mathbf{L_k})$.

From (9) and (10), the overall AGLGP predictor can be expressed as

$$\widehat{y}(\mathbf{x}^*) = \widehat{y}_{global}(\mathbf{x}^*) + \widehat{y}_{local}(\mathbf{x}^*) = \beta_0 + \mathbf{g}' \mathbf{G_m^{-1}} (\mathbf{y}_g - \mathbf{1}' \beta_0) + \mathbf{l}' \mathbf{L_n^{-1}} \mathbf{y_l} \tag{11}$$

3.1.1 Deriving the Predictive Distribution

The derivation of the overall predictor (11) assumes $\mathbf{y_g}$ is known. However $\mathbf{y_g}$ is a latent process which can not be observed directly. Similarly $\mathbf{y_l}$ is also a latent process. In this section we derive the predictive distribution of any input x^* by integrating out the random variable $\mathbf{y_g}$ and $\mathbf{y_l}$. It is worthy to note that the integration does not complicate the covariance matrix. Detailed derivation of the covariance matrix will be discussed in section 3.1.2. We define the realization of f_{local} by $\mathbf{y_l} = \{y_l^i\}_{i=1}^n$ at input $\mathbf{x} = \{x_i\}_{i=1}^n$. Based on our assumptions for f_{global} and f_{local}, the distributions of $\mathbf{y_g}$ and $\mathbf{y_l}$ are given as

$$p(\mathbf{y_g} | \mathbf{x_g}) = N(\mathbf{y_g} | \boldsymbol{\beta_0}, \mathbf{G_m}), \quad p(\mathbf{y_l} | \mathbf{x}) = N(\mathbf{y_l} | \mathbf{0}, \mathbf{L_n}) \tag{12}$$

where $\mathbf{G_m}$ and $\mathbf{L_n}$ are functions of global inducing points $\mathbf{x_g}$ and the local regions. We discuss the selection of $\mathbf{x_g}$ and the division of local regions in detail later in Section 3.1.3. The main idea is to select these $\mathbf{x_g}$ such that they are sufficiently spread out to capture the changes in the global trend, and local regions divided such that points across regions are far apart (to approximate independence). From equation (8) and (12), the conditional likelihood of a single observed point x is given as

$$p(y|x,\mathbf{x_g},\mathbf{y_g},\mathbf{x},\mathbf{y_l}) = N(y|\beta_0+\mathbf{g}'\mathbf{G_m^{-1}}(\mathbf{y_g}-\boldsymbol{\beta_0})+\mathbf{l}'\mathbf{L_n^{-1}}\mathbf{y_l},\lambda+\gamma) \tag{13}$$

where λ and γ represent the mean square prediction error for the global and local models respectively, with $\lambda = G_{nn}-\mathbf{g}'\mathbf{G_m^{-1}}\mathbf{g}, \gamma = L_{nn}-\mathbf{l}'\mathbf{L_n^{-1}}\mathbf{l}$, and G_{nn} and L_{nn} are the global and local variance at location x. The conditional distribution of $\mathbf{y_g}$ given the observed $(\mathbf{x},\mathbf{y})$ can be shown to be

$$p(\mathbf{y_g}|\mathbf{x_g},\mathbf{x},\mathbf{y}) = N(\mathbf{y_g}|\beta_0+\mathbf{G_m}\mathbf{Q_m^{-1}}\mathbf{G_{mn}}(\boldsymbol{\Lambda}+\mathbf{L_n})^{-\mathbf{1}}(\mathbf{y}-\beta_0),\mathbf{G_m}\mathbf{Q_m^{-1}}\mathbf{G_m}) \tag{14}$$

where $\boldsymbol{\Lambda} = diag(\boldsymbol{\lambda}),\boldsymbol{\lambda} = \mathbf{G_n}-\mathbf{G_{nm}}\mathbf{G_m^{-1}}\mathbf{G_{mn}}$ and $\mathbf{Q_m} = \mathbf{G_m}+\mathbf{G_{mn}}(\boldsymbol{\Lambda}+\mathbf{L_n})^{-1}\mathbf{G_{nm}}$, and the conditional distribution of $\mathbf{y_l}$ given $\mathbf{y_g}$ and $\mathbf{y}$ is given as

$$p(\mathbf{y_l}|\mathbf{x_g},\mathbf{y_g},\mathbf{x},\mathbf{y}) = N(\mathbf{y_l}|\boldsymbol{\Sigma}\boldsymbol{\Lambda}^{-1}(\mathbf{y}-\boldsymbol{\beta_0}-\mathbf{G_{nm}}\mathbf{G_m^{-1}}(\mathbf{y_g}-\beta_0)),\boldsymbol{\Sigma}) \tag{15}$$

where $\boldsymbol{\Sigma} = (\mathbf{L_n^{-1}}+\boldsymbol{\Lambda}^{-\mathbf{1}})^{-\mathbf{1}}$. Given a new input x^*, the unconditional predictive distribution can be obtained by integrating $\mathbf{y_g},\mathbf{y_l}$ from the likelihood function of (13)

$$p(y^*|x^*,\mathbf{x_g},\mathbf{x},\mathbf{y}) = \iint p(y^*|x^*,\mathbf{x_g},\mathbf{y_g},\mathbf{y_l},\mathbf{x},\mathbf{y})p(\mathbf{y_l}|\mathbf{x_g},\mathbf{y_g},\mathbf{x},\mathbf{y})p(\mathbf{y_g}|\mathbf{x_g},\mathbf{x},\mathbf{y})d\mathbf{y_g}d\mathbf{y_l} = N(y^*|\hat{\mu}^*,\hat{\sigma}^{*2}) \tag{16}$$

where

$$\begin{aligned}\hat{\mu}^* &= \beta_0+[\mathbf{g}'\mathbf{Q_m^{-1}}\mathbf{G_{mn}}+\mathbf{l}'(\boldsymbol{\Lambda}+\mathbf{L_n})^{-\mathbf{1}}(\boldsymbol{\Lambda}+\mathbf{L_n}-\mathbf{G_{nm}}\mathbf{Q_m^{-1}}\mathbf{G_{mn}})](\boldsymbol{\Lambda}+\mathbf{L_n})^{-\mathbf{1}}(\mathbf{y}-\beta_0)\\ \hat{\sigma}^{*2} &= G_{nn}-\mathbf{g}'(\mathbf{G_m^{-1}}-\mathbf{Q_m^{-1}})\mathbf{g}+\frac{(\mathbf{1}-\mathbf{1}'\mathbf{G_m^{-1}}\mathbf{g})^2}{\mathbf{1}'\mathbf{G_m^{-1}}\mathbf{1}}+L_{nn}-\mathbf{l}'[\boldsymbol{\Lambda}+\mathbf{L_n}]^{-\mathbf{1}}\mathbf{l}\end{aligned}$$

Given parameters $\boldsymbol{\theta},\boldsymbol{\alpha},\sigma^2,\boldsymbol{\tau^2}$, we can derive β_0 to maximize the likelihood function in the form $\hat{\beta}_0 = \frac{\mathbf{1}'\mathbf{R}^{-\mathbf{1}}\mathbf{y}}{\mathbf{1}'\mathbf{R}^{-\mathbf{1}}\mathbf{1}}$, where $\mathbf{R}$ is covariance matrix derived in the Section 3.1.2. Based on this, we have the following theorem,

Theorem 1 Given the parameter values $\boldsymbol{\theta},\boldsymbol{\alpha},\sigma^2,\boldsymbol{\tau^2}$, the predictive mean $\hat{\mu}^*$ is an unbiased predictor.

Proof of this theorem is provided in the Appendix A.

3.1.2 Parameter Estimation

The predictive distributions derived in equation (16) are given under the assumption that the parameters are known. However, in practice, these parameters need to be estimated through observations. We derive the estimator for the unknown parameters by maximizing likelihood function. First we derive the marginal likelihood of $\mathbf{y}$ by integrating out $\mathbf{y_g}$

$$p(\mathbf{y}|\mathbf{x},\mathbf{x_g}) = \int p(\mathbf{y}|\mathbf{x},\mathbf{x_g},\mathbf{y_g})p(\mathbf{y_g}|\mathbf{x_g})d\mathbf{y_g} = N(\beta_0,\mathbf{G}'_{\mathbf{mn}}\mathbf{G_m^{-1}}\mathbf{G_{mn}}+\boldsymbol{\Lambda}+\mathbf{L_n}) \tag{17}$$

The derived AGLGP model has an overall covariance structure $\mathbf{G_{nm}}\mathbf{G_m^{-1}}\mathbf{G_{mn}}+\boldsymbol{\Lambda}+\mathbf{L_n}$, where $\boldsymbol{\Lambda} = diag\{\mathbf{G}-\mathbf{G_{nm}}\mathbf{G_m^{-1}}\mathbf{G_{mn}}\}$ and $\mathbf{L_n}$ is a block diagonal matrix. When we let $m = n$ and $k = 1$, AGLGP model is equivalent to Composite Gaussian Process model (CGP) with a composite covariance matrix $\mathbf{G}+\mathbf{L}$. From this, we can see that the AGLGP model can also be viewed as a approximation of CGP, where $\mathbf{G_{nm}}\mathbf{G_m^{-1}}\mathbf{G_{mn}}+\boldsymbol{\Lambda}$ is an approximation for $\mathbf{G}$ and $\mathbf{L_n}$ is an approximation for $\mathbf{L}$. Our piecewise $\mathbf{L_n}$ approximation allows different covariance structure in different blocks and gives more flexibility for nonstationary features. Then the negative log-likelihood function dependent on $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$, σ^2 and $\boldsymbol{\tau^2}$ is given as

$$l(\boldsymbol{\theta},\boldsymbol{\alpha},\sigma^2,\boldsymbol{\tau^2}) = \frac{1}{2}\ln\det\mathbf{R}+\frac{1}{2}(\mathbf{y}-\widehat{\beta}_0)'\mathbf{R}^{-\mathbf{1}}(\mathbf{y}-\widehat{\beta}_0) \tag{18}$$

where $\mathbf{R} = \mathbf{G}'_{\mathbf{mn}}\mathbf{G}_{\mathbf{m}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{mn}} + \mathbf{\Lambda} + \mathbf{L}$. From the Woodbury formula (Higham 2002), we get

$$[\mathbf{G}_{\mathbf{nm}}\mathbf{G}_{\mathbf{m}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{mn}} + \mathbf{\Lambda} + \mathbf{L}_{\mathbf{n}}]^{-\mathbf{1}} = (\mathbf{I} - [\mathbf{\Lambda} + \mathbf{L}_{\mathbf{n}}]^{-\mathbf{1}}\mathbf{G}_{\mathbf{nm}}\mathbf{Q}_{\mathbf{m}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{mn}})[\mathbf{\Lambda} + \mathbf{L}_{\mathbf{n}}]^{-1}$$

where $\mathbf{L}_{\mathbf{n}}$ is a block diagonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix, so $\mathbf{\Lambda} + \mathbf{L}_{\mathbf{n}}$ can be inverted in blocks.

To minimize the negative log-likelihood function (18), optimization algorithms like the quasi-Newton methods can be applied. However, as the number of parameters to be optimized increases with the number of local regions, the approximations of the Hessian matrix can become a computational burden. Fortunately, based on our assumptions of independence across local regions, most of the terms in the Hessian matrix are zeros, *i.e.* $\frac{\partial^2 l}{\partial \alpha_{ih} \partial \alpha_{jk}} = 0, \forall h \neq k$, where α_{ih}, α_{jk} represent the local parameters of input dimensions i and j in local regions h and k. We further suggest to start the algorithm from multiple starting points to improve the convergence of these optimization methods. To address the issue of the increasing number of parameters to estimate when input dimensions increases, further assumption such as $\alpha_{ih} = \theta_i + \kappa_h$ can be made. A similar assumption is made in the estimation of the Composite Gaussian Process model (Ba and Joseph 2012).

Based on our independence assumption across the local regions, some discontinuities can be present in our estimated model. However, as our model is an additive global (smooth) model and local model, the discontinuities tend to be smaller than those from a localized model approach only. If continuity in the overall model is required, further continuity restrictions can be applied. For example, Park et al. (2011) smoothes the discontinuities in local models by adding extra constraints on the subregion boundaries when combining the local predictors. However substantial computational time is required to determine the values at the boundaries. Continuity can also be achieved by releasing the independence assumption over non-overlapping local regions. The original matrix can instead be partitioned into overlapping local regions and this helps to smooth out the local regions. However, this increases the computation required for the inversion of the new local matrix, and the larger the overlapping areas, the more complicated the matrix inversion will be. Another alternative is to constrain the local model predictors to be zero at the boundaries, and have only the global model dominate at the boundaries.

3.1.3 Selection of Inducing Points and Local Regions

To develop the AGLGP model, inducing points that summarize the observed data points and smooth out short-term fluctuations need to be selected for the global model. In addition, the whole input space needs to be divided into local regions for a piecewise local model.

In selecting inducing points for sparse matrix approximation methods, various selection criterion have been summarized in Quiñonero Candela and Rasmussen (2005). For the global model, it is desirable to place these artificial (inducing) points such that collectively, they are able to capture the global trend, not only in the location x, but also in the observation values y.

To divide the input space with a given a set of input designs and corresponding observations, we first classify the data through hyperplanes that represent the largest separation, or margin, between classes so that the independent assumptions across local regions can be reasonably made. Hence, points are grouped according to their Euclidean distance and the boundaries are drawn such that the largest separation or margin between two subregions is achieved. Figure 2 illustrates the desirable characteristics of the local regions decomposition and the selection of inducing points.

Here we describe in detail the algorithm applied to attain this. First we divide the whole input space R into k different local regions $r_1, ..., r_k$. Suppose n observed points $\mathbf{x} = (x_1, ..., x_n)$ are separated into k sets of data $\mathbf{x} = (\mathbf{x}_{s_1}, ..., \mathbf{x}_{s_k})$, where $\mathbf{x}_{s_i} \in r_i, \forall i = 1, ..., k$. Points are separated through k-means, which based on the *Euclidean* distance of input $\mathbf{x}$ to minimize the within-region sum of squares $\arg\min_{\mathbf{s}} \sum_{i=1}^{k} \sum_{x \in \mathbf{x}_{s_i}} \|x - \mu_i\|^2$. μ_i is the mean of points in $\mathbf{x}_{s_i}$. Then based on the data sets $\mathbf{x}_{s_i}$ in a local region r_i, we construct boundaries of local regions through Support Vector Machine (SVM), which chooses the best hyperplane that represents the largest separation or margin between two neighboring local regions. Hsu and Lin (2002) summarizes

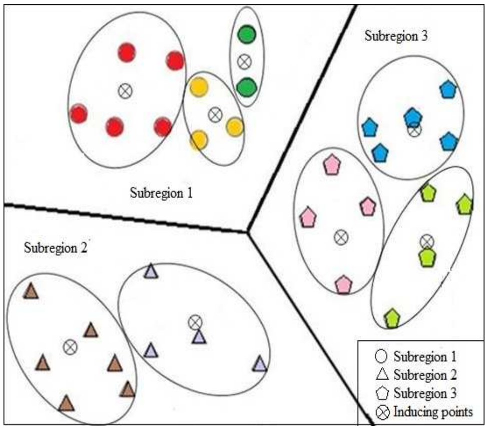

Figure 2: Domain Decomposition and inducing points generation

the methods for multiclass SVM. In this paper, we generate pairwise classifiers. Two hyperplanes that separate data set $\mathbf{x}_{s_i}$ and $\mathbf{x}_{s_j}, i \neq j$ with no points between them are described by a set of points $\mathbf{x}_{\mathbf{p_1}}$ and $\mathbf{x}_{\mathbf{p_2}}$ that satisfies $\mathbf{w} \cdot \mathbf{x}_{\mathbf{p_2}} - b = 1$, $\mathbf{w} \cdot \mathbf{x}_{\mathbf{p_1}} - b = -1$, where $\mathbf{w}$ is the normal vector to the hyperplane. Parameters are optimized by maximizing the distance between this two hyperplanes $\frac{2}{||\mathbf{w}||}$, which is equivalent to

$$\begin{aligned} min \quad & ||\mathbf{w}|| \\ s.t. \quad & \mathbf{w} \cdot x_i - b \geq 1, x_i \in \mathbf{x}_{s_i} \\ & \mathbf{w} \cdot x_j - b \leq -1, x_j \in \mathbf{x}_{s_j} \end{aligned}$$

where the constraints ensure that there is no data between two hyperplanes.

With the defined local regions, the set of points $\mathbf{x}_{s_i}$ within a certain local region r_i are then further divided into clusters based on their observations $\mathbf{y}_{s_i}$: firstly we define the range of overall observations $\Delta y = y_{max} - y_{min}$ and the range of points within cluster Δ; next, contour lines are drawn with an interval of Δ, i.e. $L = \{(x,y)|y = c\}$ where $c \in \{y_{min}, y_{min} + \Delta, ..., y_{max} - \Delta, y_{max}\}$. Finally, the set of points $\mathbf{x}_{s_i}$ is further divided into ϕ clusters $\mathbf{x}_{s_i} = (\mathbf{c}_{i1}, ..., \mathbf{c}_{i\phi})$, where $\mathbf{c}_{ij} = \{x | y_{min} + \Delta \times (j-1) \leq y(x) \leq y_{min} + \Delta \times j\}, j = 1, ..., \phi$.

If $||x_h - x_k|| > ||x_l - x_h||, \forall x_h, x_k \in \mathbf{c}_{ij}, \exists x_l \notin \mathbf{c}_{ij}$, the cluster $\mathbf{c}_{ij}$ will be further divided into two subclusters between x_i and x_j. Finally we have m subclusters $\mathbf{c} = (c_1, ..., c_m)$, where the variability within a cluster is relatively small compared to the total variability of the whole domain. Then the points in a subcluster are aggregated to generate an inducing point. The summary of the algorithm is described in Table 1.

Table 1: Inducing points selection and input space decomposition

Step 1	Separate points $\mathbf{x}$ into k sets $\mathbf{s} = (s_1, ..., s_k)$ through k-means based on locations of $\mathbf{x}$
Step 2	Generate boundaries for the k set of data $\mathbf{s} = (s_1, ..., s_k)$ via SVM that gives the largest separation between data set $\mathbf{s}$ to get the local regions $\{r_i\}_{i=1,...,k}$
Step 3	Further divide points $\mathbf{s}_i$ in each local region into clusters $\mathbf{c}_i = (\mathbf{c}_{i1}, ..., \mathbf{c}_{i\phi})$ based on a set of contour lines $L = \{(x,y)|y = c\}$ where $c \in \{y_{min}, y_{min} + \Delta, ..., y_{max} - \Delta, y_{max}\}$
Step 4	If $||x_h - x_k|| > ||x_l - x_h||, \forall x_h, x_k \in \mathbf{c}_{ij}, \exists x_l \notin \mathbf{c}_{ij}$, the cluster $\mathbf{c}_{ij}$ will be further divided into subclusters. Finally points in each subcluster are aggregated into inducing points

With this algorithm, two key points need to be specified, specifically the number of inducing points and the number of regions. Here we recommend the number of inducing points selected to be under a hundred (for computational efficiency), and the number of local regions to be no more than ten (for the computational efficiency of the multiclass SVM classifier). This however has to be traded-off with the data size and the number of parameters to be estimated. In the numerical examples, we find that regions divided such that the number of parameters is limited within ten work reasonably well.

3.2 Stochastic AGLGP

For stochastic simulation, the mean of the simulation output is modeled by

$$y(x) = f(x) + \varepsilon(x) = f_{global}(x) + f_{local}(x) + \varepsilon(x) \tag{19}$$

where $f(x)$ describes the mean of the process and $\varepsilon(\mathbf{x})$ is the random noise. We assume $\varepsilon(\mathbf{x}) \sim N(0, \sigma_\varepsilon^2(\mathbf{x}))$ and the error variance may depends on x. In stochastic simulation when replicates at each observation points are observed, the sample mean of the replicates are typically used in model estimation. We denote the sample mean and sample variance as $y(\mathbf{x}_i) = \frac{\sum_{j=1}^{r} y_j(\mathbf{x}_i)}{r}, v(\mathbf{x}_i) = \frac{\sum_{j=1}^{r}(y_j(\mathbf{x}_i) - y(\mathbf{x}_i))^2}{r-1}$. Similar to the deterministic case, the output of the stochastic AGLGP is also additive, with $\mathbf{y} = \mathbf{y_g} + \mathbf{y_l}$. In this stochastic case, we still assume no noise in the latent process $\mathbf{y_g}$. This is a reasonable assumption as the process is unobservable. Then we have the distribution of $\mathbf{y_l}$ as $p(\mathbf{y_l}|\mathbf{x}) = N(\mathbf{y_l}|\mathbf{0}, \mathbf{L_n} + \Sigma_\varepsilon)$. The likelihood of a single point is then given by

$$p(y|x, \mathbf{x_g}, \mathbf{y_g}, \mathbf{x}, \mathbf{y_l}) = N(y|\beta_0 + \mathbf{g}'\mathbf{G_m^{-1}}(\mathbf{y_g} - \beta_0) + \mathbf{l}'(\mathbf{L_n} + \boldsymbol{\Sigma_\varepsilon})^{\mathbf{-1}}\mathbf{y_l}, \lambda + \gamma + \sigma_\varepsilon^2(x)) \tag{20}$$

The conditional distribution of $\mathbf{y_g}$ and $\mathbf{y_l}$ can be shown to be

$$p(\mathbf{y_g}|\mathbf{x_g}, \mathbf{x}, \mathbf{y}) = N(\mathbf{y_g}|\beta_0 + \mathbf{G_m Q_m^{-1} G_{mn} K^{-1}}(\mathbf{y} - \beta_0), \mathbf{G_m Q_m^{-1} G_m}) \tag{21}$$

$$p(\mathbf{y_l}|\mathbf{x_g}, \mathbf{y_g}, \mathbf{x}, \mathbf{y}) = N(\mathbf{y_l}|\mathbf{L_n K^{-1}}\{\mathbf{y} - \beta_0 - \mathbf{G_{nm} G_m^{-1}}(\mathbf{y_g} - \beta_0)\}, \mathbf{L_n} - \mathbf{L_n K^{-1} L_n} + \boldsymbol{\Sigma}_\varepsilon) \tag{22}$$

where $\mathbf{Q_m} = \mathbf{G_m} + \mathbf{G_{mn} K^{-1} G_{nm}}$ and $\mathbf{K} = \mathbf{L_n} + \boldsymbol{\Lambda} + \boldsymbol{\Sigma}_\varepsilon$. Given a new input x^*, with (21) and (22), the predictive distribution is obtained by integrating the likelihood function (20) to give

$$p(y^*|x^*, \mathbf{x_g}, \mathbf{x}, \mathbf{y}) = N(y^*|\hat{\mu}^*, \hat{\sigma}^{*2}) \tag{23}$$

where

$$\hat{\mu}^* = \beta_0 + [\mathbf{g}'\mathbf{Q_m^{-1} G_{mn}} + \mathbf{l}'(\mathbf{L_n} + \boldsymbol{\Sigma}_\varepsilon)^{-1}\mathbf{L_n K^{-1}}(\mathbf{K} - \mathbf{G_{nm} Q_m^{-1} G_{mn}})]\mathbf{K^{-1}}(\mathbf{y} - \beta_0)$$

$$\hat{\sigma}^{*2} = G_{nn} - \mathbf{g}'(\mathbf{G_m^{-1}} - \mathbf{Q_m^{-1}})\mathbf{g} + \frac{(\mathbf{1} - \mathbf{1}'\mathbf{G_m^{-1}}\mathbf{g})^{\mathbf{2}}}{\mathbf{1}'\mathbf{G_m^{-1}}\mathbf{1}} + L_{nn} - \mathbf{l}'[\boldsymbol{\Sigma}_\varepsilon + \mathbf{L_n}]^{\mathbf{-1}}\mathbf{L_n K^{-1} L_n}[\boldsymbol{\Sigma}_\varepsilon + \mathbf{L_n}]^{\mathbf{-1}}\mathbf{l} + \sigma_\varepsilon^2(x^*)$$

The predictor is still an unbiased predictor but is no longer an interpolator. The parameter estimation can be similarly estimated based on the MLE method discussed in Section 3.1.2 but with $\mathbf{L_n} + \boldsymbol{\Lambda}$ is replaced by $\mathbf{K}$.

In the heterogeneous case where the random error is assumed to be independent but not identical, variance information is not available to estimate $\sigma_\varepsilon^2(x^*)$ unless the location has been previously observed. Here we propose to model $\log(\sigma_\varepsilon^2)$ as a Gaussian Process. This natural log transformation has the nice properties of approximating normality, stabilizing variance, and ensuring inverse transformation back to the positive scale. The details can be found in Ng and Yin (2012).

4 NUMERICAL STUDY OF TEST FUNCTIONS

In this section, we study the performance of the model. We measure the prediction accuracy compared with some existing approximation approach. The approximation methods compared includes FSA (Sang and Huang 2012), localized GP model (i.e. local kriging), reduced rank approximation (Banerjee, Gelfand, Finley, and Sang 2008) and PIC (Snelson and Ghahramani 2007).

The implementation of all methods were conducted in MATLAB. As the implementations of the other methods are not available, we wrote our own codes for PIC, FSA, local and reduced rank. Throughout the numerical analysis, we used the Gaussian covariance function. All numerical studies were performed on a processor with quad-core 3.3 GHz CoreTM i5 CPU and 8 GB memory.

Two test functions are applied. The first is a one dimension function $y(x) = \cos(100(x-0.2))e^{2x} + 7\sin(10x)$ as shown in Figure 1. It has long lengthscale global trend and also local (shorter lengthscale) variation. Computer experiments are simulated with three noise level functions, namely $\sigma_e = 0$, $5.5 + 4.5\sin(10x)$, $15.5 + 14.5\sin(10x)$. 1000 sampled locations are taken and 20 replications are simulated at each sampled location. The average mean squared error (IMSE) between the predictor and signal function at $N = 1000$ unsampled points is used as the performance measure. Specifically in each macro replication k, we observe $MSE(k) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}(x_i) - y(x_i))^2$, and the average of the MSE based on $M = 1000$ macro replications is used, $IMSE = \frac{1}{M}\sum_{k=1}^{M} MSE(k)$. For FSA and PIC, we chose the same number of inducing inputs as AGLGP and this differs in each macro-replication. The tapering scale of FSA is 0.2. PIC shares the same decomposition method and same number of local regions five with AGLGP in this example.

Table 2: IMSE of approximation models with one-dimension test function

Model	AGLGP	Local	FSA	Reduced Rank	PIC
$\sigma_e = 0$	0.0212	0.0208	0.0875	0.0982	0.0832
$\sigma_e = 5.5 + 4.5\sin(10x)$	0.1542	0.1616	0.5365	0.6501	0.5292
$\sigma_e = 15.5 + 14.5\sin(10x)$	1.6235	1.6491	1.8476	1.7622	1.6536

Table 2 summarizes the results. Furthermore, pair-wise t-test results show that the mean square prediction errors for the AGLGP model and Local Kriging are statistically better than the other models (FSA, PIC, Reduced Rank) at $\alpha = 0.05$. The difference of the mean square prediction error between AGLGP model and Local Kriging is not statistically significant at $\alpha = 0.05$. However, the Local Kriging suffers larger discontinuities on the boundary. Figure 3 illustrates the predictors outputs of AGLGP model and LocalKriging model. From this figure, we can see obvious discontinuities of LocalKriging model at the boundaries of local regions. The AGLGP model displays a much smoother fitted response.

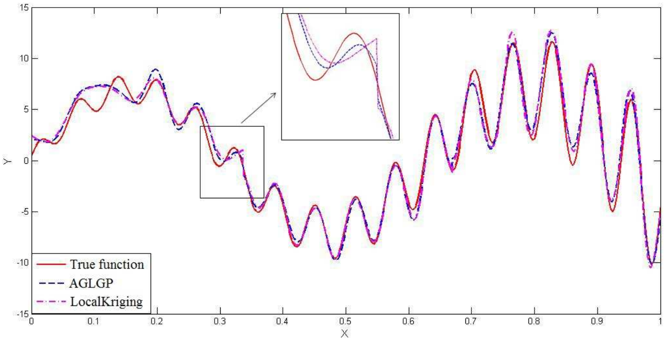

Figure 3: Fitted one-dimension test function via AGLGP and LocalKriging

The second test function studied is the Six-Hump Camel function $(4 - 2.1x_1^2 + x_1^4/3)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2$. We randomly selected 400 input locations from the region [-2, 2]×[-2, 2]. The noise function $\sigma_e = 6 + 5\sin(x_1)$ is used and 20 replicates are taken at each input location. 6561 separate points are then predicted and the mean square prediction error observed is 1.3897. The simulation results is shown in Figure 4.

The left graph in Figure 4 is the predicted surface while the right is the error between the prediction and the true value $e = \hat{y}(x^*) - y(x^*)$. Compared to the noise variance level, we can see that the error is small for most points unless the points located at the boundary of the whole region. Specifically, among

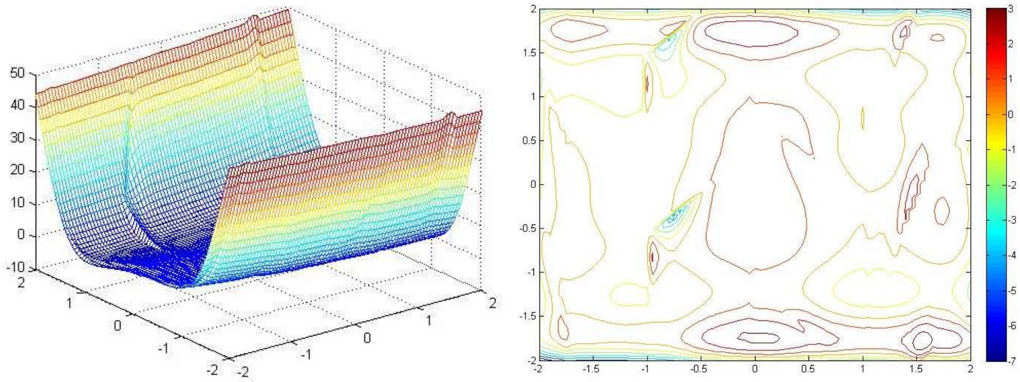

Figure 4: Simulation for Six-Hump Camel function

all the 6561 predicted points, 4431 points have errors $e \in [-1,1]$ while 6048 points have errors $e \in [-2,2]$. The result shows a good interpolation prediction with AGLGP model.

5 CONCLUSION

This paper proposes an additive global and local Gaussian Process (AGLGP) model to approximate the response surface of computer models where large data sets are observed. The proposed AGLGP model incorporates a global model and a local model, which combines to be a good approximation of a Gaussian Process with a composite covariance structure. In addition to reducing the computational burden of large data sets, the additive structure of the model is flexible and is able to capture nonstationarity features of the computer models.The numerical study illustrate the performance of AGLGP with several other approximating GP models, and the results of AGLGP are promising.

To further understand the performance advantages of AGLGP, an extension of this work is to conduct an analytical comparison of the approximation components, including the covariance structure of the various approximation models. Another related extension of this work is its potential application in simulation optimization, where individual components of the additive model can be adopted to efficiently explore and exploit the design space.

A UNBIASEDNESS

For the AGLGP model, given the model parameters, the predictive mean $\hat{\mu}^*$ can be expressed by $\hat{\mu}^* = \hat{\beta}_0 + (\mathbf{g}'\mathbf{G_m^{-1}G_{mn}} + \mathbf{l}')\mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}'\hat{\beta}_0)$, where $\mathbf{R} = \mathbf{G}'_{\mathbf{mn}}\mathbf{G_m^{-1}G_{mn}} + \mathbf{\Lambda} + \mathbf{L_n}$ and $\hat{\beta}_0 = \frac{\mathbf{1'R^{-1}}}{\mathbf{1'R^{-1}1}}$. $\hat{\mu}^*$ is also a linear combination of $\mathbf{y}$, *i.e.* $\hat{\mu}^* = \sum_{i=1}^{n} \lambda_i y_i$,

$$\lambda_i = \left[\frac{\mathbf{1'R^{-1}}}{\mathbf{1'R^{-1}1}} + (\mathbf{g}'\mathbf{G_m^{-1}G_{mn}} + \mathbf{l}')\mathbf{R}^{-1}(1 - \mathbf{1}'\frac{\mathbf{1'R^{-1}}}{\mathbf{1'R^{-1}1}}) \right] e_i$$

where $e_i = [0,0,..., \underbrace{1}_{the\ ith\ element},...,0,0]$; $\sum_{i=1}^{n} \lambda_i = 1$. Hence, $E[\hat{\mu}^*] = E[\sum_{i=1}^{n} \lambda_i y_i] = \sum_{i=1}^{n} \lambda_i E[y_i] = E[y(x^*)]$.

REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2010. “Stochastic Kriging for Simulation Metamodeling”. *Operations Research* 58 (2): 371–382.

Ba, S., and V. R. Joseph. 2012. “Composite Gaussian process models for emulating expensive functions”. *The Annals of Applied Statistics* 6 (4): 1838–1860.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang. 2008. "Gaussian predictive process models for large spatial data sets.". *Journal of the Royal Statistical Society. Series B, Statistical methodology* 70 (4): 825–848.

Cressie, N. 1993. *Statistics for Spatial Data*. Wiley, New York.

Furrer, R., M. G. Genton, and D. Nychka. 2006. "Covariance Tapering for Interpolation of Large Spatial Datasets". *Journal of Computational and Graphical Statistics* 15 (3): 502–523.

Gramacy, R. B., and D. W. Apley. 2014. "Local Gaussian process approximation for large computer experiments". *Journal of Computational and Graphical Statistics*:1–28.

Gramacy, R. B., and B. Haaland. 2015. "Speeding up neighborhood search in local Gaussian process prediction". *Technometrics* (to appear).

Higham, N. J. 2002. *Accuracy and stability of numerical algorithms*. SIAM.

Hsu, C.-W., and C.-J. Lin. 2002. "A comparison of methods for multiclass support vector machines". *Neural Networks, IEEE Transactions on* 13 (2): 415–425.

Jones, D., M. Schonlau, and W. Welch. 1998. "Efficient global optimization of expensive black-box functions". *Journal of Global optimization*:455–492.

Li, Y., S. Ng, M. Xie, and T. Goh. 2010. "A systematic comparison of metamodeling techniques for simulation optimization in Decision Support Systems". *Applied Soft Computing* 10 (4): 1257–1273.

Ng, S. H., and J. Yin. 2012. "Bayesian Kriging Analysis and Design for Stochastic Simulations". *ACM Transactions on Modeling and Computer Simulation* 22 (3): 1–26.

Park, C., J. Huang, and Y. Ding. 2011. "Domain decomposition approach for fast Gaussian process regression of large spatial data sets". *The Journal of Machine Learning Research* 12:1697–1728.

Quiñonero Candela, J., and C. Rasmussen. 2005. "A unifying view of sparse approximate Gaussian process regression". *The Journal of Machine Learning Research* 6:1939–1959.

Sacks, J., W. Welch, T. Mitchell, and H. Wynn. 1989. "Design and analysis of computer experiments". *Statistical science* 4 (4): 409–435.

Sang, H., and J. Huang. 2012. "A full scale approximation of covariance functions for large spatial data sets". *Journal of the Royal Statistical Society* 74:111–132.

Simpson, T. W., J. Poplinski, P. N. Koch, and J. K. Allen. 2001. "Metamodels for computer-based engineering design: survey and recommendations". *Engineering with computers* 17 (2): 129–150.

Snelson, E., and Z. Ghahramani. 2007. "Local and global sparse Gaussian process approximations". *International Conference on Artificial Intelligence and Statistics*.

Stein, M. L. 1991. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Yin, J., S. Ng, and K. Ng. 2011. "Kriging metamodel with modified nugget-effect: The heteroscedastic variance case". *Computers & Industrial Engineering* 61 (3): 760–777.

AUTHOR BIOGRAPHIES

QUN MENG received the B.E. degree in industrial engineering and logistics management from Shanghai Jiao Tong University, Shanghai, China, in 2012. She is currently working towards the Ph.D. degree at the Department of Industrial and Systems Engineering, National University of Singapore. Her research interests are in the area of simulation optimization. Her email address is mengqun@u.nus.edu.

SZU HUI NG is an Associate Professor in the Department of Industrial and Systems Engineering at the National University of Singapore. She holds B.S., M.S., and Ph.D. degrees in Industrial and Operations Engineering from the University of Michigan. Her research interests include computer simulation modeling and analysis, design of experiments, and quality and reliability engineering. She is a member of IEEE and INFORMS and a senior member of IIE. Her email address is isensh@nus.edu.sg and her web page is http://www.ise.nus.edu.sg/staff/ngsh/index.html.