

Composite Gaussian Process Models for Emulating Expensive Functions

Shan Ba and V. Roshan Joseph

School of Industrial and Systems Engineering

Georgia Institute of Technology

Atlanta, GA 30332-0205, USA

sba3@isye.gatech.edu and roshan@isye.gatech.edu

Abstract

A new type of non-stationary Gaussian process model is developed for approximating computationally expensive functions. The new model is a composite of two Gaussian processes, where the first one captures the smooth global trend and the second one models local details. The new predictor also incorporates a flexible variance model, which makes it more capable of approximating surfaces with varying volatility. Compared to the commonly used stationary Gaussian process model, the new predictor is numerically more stable and can more accurately approximate complex surfaces when the experimental design is sparse. In addition, the new model can also improve the prediction intervals by quantifying the change of local variability associated with the response. Advantages of the new predictor are demonstrated using several examples.

KEY WORDS: Computer experiments, Functional approximation, Kriging, Nugget, Non-stationary Gaussian process.

1 Introduction

The modern era witnesses the prosperity of computer experiments, which play a critical role in many fields of technological development where the traditional physical experiments are infeasible or unaffordable to conduct. By developing sophisticated computer simulators, people are able to evaluate, optimize and test complex engineering systems even before building expensive prototypes. The computer simulations are usually deterministic (no random error), yield highly nonlinear response surfaces, and are very time-consuming to run. To facilitate the analysis and optimization of the underlying system, surrogate models (or emulators) are often fitted to approximate the unknown simulated surface based on a finite number of evaluations (Sacks, Welch, Mitchell and Wynn 1989). Santner, Williams and Notz (2003) and Fang, Li and Sudjianto (2006) provide detailed reviews on the related topics.

In computer experiments, the stationary Gaussian process (GP) model is popularly used for approximating computationally expensive simulations. Its framework is built on modeling the computer outputs $Y(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^p$ as a realization of a stationary GP with constant mean μ and covariance function $\sigma^2 \text{cov}(Y(\mathbf{x} + \mathbf{h}), Y(\mathbf{x})) = \sigma^2 R(\mathbf{h})$, where the correlation $R(\mathbf{h})$ is a positive semidefinite function with $R(\mathbf{0}) = 1$ and $R(-\mathbf{h}) = R(\mathbf{h})$. When the above assumptions are satisfied, the corresponding predictor can be shown to be a *best linear unbiased predictor* (BLUP), in the sense that it minimizes the mean squared prediction error. Nevertheless, many studies in the literature have pointed out that the artificial assumption of second-order stationarity for the GP model are more for theoretical convenience rather than for representing reality, and they can be easily challenged in practice. If these assumptions deviate from the truth, the predictor is no longer optimal, and sometimes can even be problematic (see the discussions, for example, in Joseph 2006, Xiong et al. 2007, Gramacy and Lee 2011).

When the constant mean assumption for the GP model is violated, a frequently observed consequence is that the predictor tends to revert to the global mean, especially at locations far from design points. Consider a simple example from Xiong et al. (2007). Suppose the true function is $y(x) = \sin(30(x - 0.9)^4) \cos(2(x - 0.9)) + (x - 0.9)/2$ and we choose 17

unequally spaced points from $[0,1]$ to evaluate the function. The function and design points are illustrated in Figure 1. Obviously, the mean of this function in region $x \in [0, 0.4]$ is much smaller than the mean in region $x \in [0.4, 1]$. When the data are fitted with a stationary GP model with a Gaussian correlation function, a constant mean for the whole region is estimated as -0.147 by maximizing the likelihood function (Santner et al. 2003, page 66), and the corresponding predictor along with this mean value are shown in Figure 1. Clearly, the fit in region $x \in [0.4, 1]$ is not good, since the prediction is pulled down to the global mean.

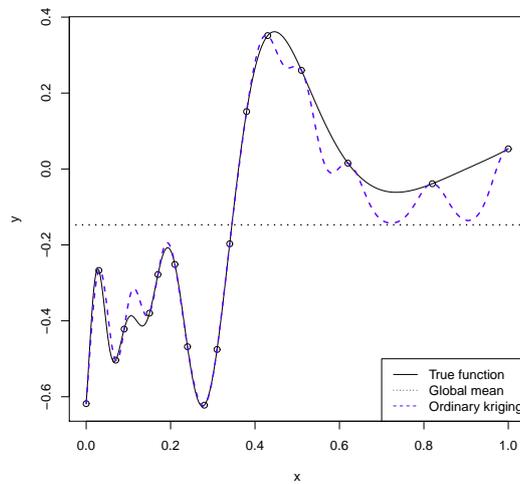


Figure 1: Plot of function $y(x) = \sin(30(x - 0.9)^4) \cos(2(x - 0.9)) + (x - 0.9)/2$, the global mean and the ordinary kriging predictor.

Just as a non-constant global trend can be quite common in engineering systems, the variability of simulated outputs can also change dramatically throughout the design region. Still consider the simple case in Figure 1 for example: the roughness of the one-dimensional function in region $x \in [0, 0.4]$ is much larger than in region $x \in [0.4, 1]$. For the GP model assuming a constant variance for the whole input region, the variance estimate for region $x \in [0.4, 1]$ tends to be inflated by averaging with that of the other part, which further contributes to the erratic prediction in this region. It is expected that as we increase the simulation sample size, the above problem can be mitigated. However, since most typical

applications of computer experiments involve high dimensional inputs, the data points always tend to be sparse in the design region and it is almost impossible to avoid such kind of gaps in practice.

In this article, we propose a more accurate modeling approach by incorporating a flexible global trend and a variance model into the GP model. The proposed predictor has an intuitive structure and can be efficiently estimated in a single stage. Not only can the new predictor mitigate the problems discussed above, it also enjoys several additional advantages such as better numerical stability, robustness to sparse design and improved prediction intervals.

The article is organized as follows. Section 2 introduces the notation and existing work. Section 3 presents the new predictor and shows its interesting connections with some existing methods. In Section 4 we discuss how to estimate the unknown parameters by maximum likelihood. Several properties of the new predictor are studied in Section 5, and in Section 6 we use several examples to demonstrate the advantages of the new method. Some final concluding remarks are given in Section 7.

2 Notation and Existing Work

In the computer experiments literature, the GP model is also often referred to as the *kriging* model (Currin et al. 1991), and these two terms are used interchangeably in this article. Suppose we have run the simulations under n different input settings $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^p$. Denote the corresponding computer outputs as $\mathbf{y} = (y_1, \dots, y_n)^\top$. A stationary GP model, called *ordinary kriging*, can be formally stated as

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x}), \tag{1}$$

where $Z(\mathbf{x}) \sim GP(0, \sigma^2 R(\cdot))$. The ordinary kriging predictor at an input location \mathbf{x} is given by

$$\hat{y}(\mathbf{x}) = \hat{\mu} + \mathbf{r}^\top(\mathbf{x})\mathbf{R}^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}), \tag{2}$$

where $\mathbf{r}(\mathbf{x}) = (R(\mathbf{x}-\mathbf{x}_1), \dots, R(\mathbf{x}-\mathbf{x}_n))^\top$, \mathbf{R} is an $n \times n$ correlation matrix with the (ij) th element $R(\mathbf{x}_i-\mathbf{x}_j)$, $\mathbf{1}$ is a n -dimensional vector with all elements 1, and $\hat{\mu} = (\mathbf{1}^\top \mathbf{R}^{-1} \mathbf{1})^{-1} (\mathbf{1}^\top \mathbf{R}^{-1} \mathbf{y})$.

To remedy the the predictor’s reversion to mean problem as discussed in the previous section, a common strategy is to relax the constant mean μ in ordinary kriging with a *global trend* $\mu(\mathbf{x})$, and modify the model in (1) as

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}). \quad (3)$$

If the global trend is comprised of some prescribed polynomial models $\mu(\mathbf{x}) = \mathbf{f}^\top(\mathbf{x})\boldsymbol{\beta}$, where $\mathbf{f}(\mathbf{x}) = (1, f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))^\top$ are known functions and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)^\top$ are unknown parameters, the model in (3) is called *universal kriging*. Define a $n \times (m+1)$ matrix $\mathbf{F} = (\mathbf{f}(\mathbf{x}_1), \dots, \mathbf{f}(\mathbf{x}_n))^\top$, and the corresponding optimal predictor under model (3) can be derived as

$$\hat{y}(\mathbf{x}) = \mathbf{f}^\top(\mathbf{x})\hat{\boldsymbol{\beta}} + \mathbf{r}^\top(\mathbf{x})\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}), \quad (4)$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{F}^\top \mathbf{R}^{-1} \mathbf{F})^{-1} (\mathbf{F}^\top \mathbf{R}^{-1} \mathbf{y})$. If $\mu(\mathbf{x}) = \mathbf{f}^\top(\mathbf{x})\boldsymbol{\beta}$ is close to the true global trend, then clearly this approach can give much better prediction than that of (2). However, in practice the correct functional form $\mathbf{f}(\mathbf{x}) = (1, f_1(\mathbf{x}), \dots, f_m(\mathbf{x}))^\top$ is rarely known, and a wrongly specified trend in universal kriging can make the prediction even worse. For this reason, Welch et al. (1992) suggested using ordinary kriging instead of universal kriging. Another practical approach, called *blind kriging*, is to relax the assumption that the $f_i(\mathbf{x})$ ’s are known and select them from a candidate set of functions using a variable selection technique (Joseph, Hung and Sudjianto 2008). Although this strategy usually leads to better fit, performing the variable selection while interacting with the second stage GP model is a non-trivial task. Considerable computational efforts are needed to properly divide up the total variation between the polynomial trend and the GP model. In addition, in some cases, polynomial models may not be adequate to fit the complex global trend well.

Generalizing the GP model for non-stationary variance is an even more challenging task. None of the above remedies for the non-stationary mean can in any sense alleviate the constant variance restriction, and most studies in the literature focus on deriving complex non-stationary covariance functions such as by spatial deformations or kernel convolution

approaches (for example, see Sampson and Guttorp 1992, Higdon, Swall, and Kern 1999, Schmidt and O’Hagan 2003, Paciorek and Schervish 2006, Anderes and Stein 2008). However, those structures may easily get overparameterized in high dimensions and become computationally intractable to fit. In addition, many of them also require multiple observations, which is not applicable to the single set of outputs from computer experiments. Some other work includes Xiong et al. (2007), which adopts a non-linear mapping approach based on a parameterized density function to incorporate the non-stationary covariance structure. Gramacy and Lee (2008) utilize the Bayesian treed structure to implement a non-stationary GP model. However, by dividing the design space into subregions, the treed GP model may lose efficiency since the prediction is only based on local information, and its response can also be discontinuous across subregions. In the next section, we propose to solve the non-stationarity problem via a different approach. We show that the flexible mean and variance models can be incorporated into GP by using the *composite Gaussian process* (CGP) models.

3 Composite Gaussian Process Models

For clarity, in this section we develop the new method in two steps. First, a predictor that intrinsically incorporates a flexible mean model is presented, and then we further augment it with a variance model to simultaneously handle the change of variability in the response.

3.1 Improving the mean model

The universal kriging (or blind kriging) in (3) contains a polynomial mean model $\mu(\mathbf{x})$ as the global trend and a kriging model $Z(\mathbf{x})$ for local adjustments. To avoid the awkward variable selections in $\mu(\mathbf{x})$ and also make the mean model more flexible, we propose to use another GP to model the $\mu(\mathbf{x})$ as in the following form

$$\begin{aligned}
 Y(\mathbf{x}) &= Z_{global}(\mathbf{x}) + Z_{local}(\mathbf{x}), & (5) \\
 Z_{global}(\mathbf{x}) &\sim GP(\mu, \tau^2 g(\cdot)), \\
 Z_{local}(\mathbf{x}) &\sim GP(0, \sigma^2 l(\cdot)).
 \end{aligned}$$

Here the two GPs $Z_{global}(\mathbf{x})$ and $Z_{local}(\mathbf{x})$ are stationary and independent of each other. The first GP with variance τ^2 and correlation structure $g(\cdot)$ is required to be *smoother* to capture the global trend while the second GP with variance σ^2 and correlation $l(\cdot)$ is for local adjustments. Just as the universal kriging generalizes the ordinary kriging by adding a polynomial mean model $\mu(\mathbf{x})$, the new model in (5) can be viewed as a further extension which adopts a more sophisticated GP for global trend modeling. It is interesting to note that, the *linear model of regionalization* in geostatistics (Wackernagel 2003, Chapter 14) also employs a similar structure to model regionalized phenomena in geological data, but its final model form and estimation strategies are quite different from our approach.

Under the new assumptions in (5), the optimal predictor is easy to derive. Since the sum of two independent GPs is still a GP, we can equivalently express (5) as $Y(\mathbf{x}) \sim GP(\mu, \tau^2 g(\cdot) + \sigma^2 l(\cdot))$. Similar to ordinary kriging, the best linear unbiased predictor under the assumptions in (5) can be written as

$$\hat{y}(\mathbf{x}) = \hat{\mu} + (\mathbf{g}(\mathbf{x}) + \lambda \mathbf{l}(\mathbf{x}))^\top (\mathbf{G} + \lambda \mathbf{L})^{-1} (\mathbf{y} - \hat{\mu} \mathbf{1}) \quad (6)$$

where $\lambda = \sigma^2/\tau^2$ ($\lambda \in [0, 1]$) is the ratio of variances, $\mathbf{g}(\mathbf{x}) = (g(\mathbf{x} - \mathbf{x}_1), \dots, g(\mathbf{x} - \mathbf{x}_n))^\top$, $\mathbf{l}(\mathbf{x}) = (l(\mathbf{x} - \mathbf{x}_1), \dots, l(\mathbf{x} - \mathbf{x}_n))^\top$, \mathbf{G} and \mathbf{L} are two $n \times n$ correlation matrices with the (ij) th element $g(\mathbf{x}_i - \mathbf{x}_j)$ and $l(\mathbf{x}_i - \mathbf{x}_j)$ respectively, and $\hat{\mu} = (\mathbf{1}^\top (\mathbf{G} + \lambda \mathbf{L})^{-1} \mathbf{1})^{-1} \mathbf{1}^\top (\mathbf{G} + \lambda \mathbf{L})^{-1} \mathbf{y}$. Here the variance ratio λ is restricted to $[0, 1]$ because we expect the global trend to capture most of the variation in the response surface than the local process.

Although many possible correlation structures are available for $g(\cdot)$ and $l(\cdot)$, throughout this paper we follow the standard choice in computer experiments and specify them using the *Gaussian correlation functions*:

$$g(\mathbf{h}|\boldsymbol{\theta}) = \exp\left(-\sum_{j=1}^p \theta_j h_j^2\right), \quad l(\mathbf{h}|\boldsymbol{\alpha}) = \exp\left(-\sum_{j=1}^p \alpha_j h_j^2\right), \quad (7)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are unknown correlation parameters satisfying $\mathbf{0} \leq \boldsymbol{\theta} \leq \boldsymbol{\alpha}^l$ and $\boldsymbol{\alpha}^l \leq \boldsymbol{\alpha}$. The bounds $\boldsymbol{\alpha}^l$ are usually set to be moderately large, which ensures that the component $Z_{global}(\mathbf{x})$ is indeed smoother than $Z_{local}(\mathbf{x})$ in the fitted model.

The new predictor in (6) is still an interpolator, since $\hat{y}(\mathbf{x}_i) = \hat{\mu} + \mathbf{e}_i^\top(\mathbf{y} - \hat{\mu}\mathbf{1}) = y_i$ for $i = 1, \dots, n$, where \mathbf{e}_i is a unit vector with a 1 at its i th position. It can also be seen that when $\lambda = 0$ (i.e. $\sigma^2 = 0$), the new model reduces to ordinary kriging. When $\lambda \in (0, 1]$, the predictor in (6) can be written out as the sum of a global predictor and a local predictor

$$\hat{y}(\mathbf{x}) = \hat{y}_{global}(\mathbf{x}) + \hat{y}_{local}(\mathbf{x}), \quad (8)$$

$$\hat{y}_{global}(\mathbf{x}) = \hat{\mu} + \mathbf{g}^\top(\mathbf{x})(\mathbf{G} + \lambda\mathbf{L})^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}), \quad (9)$$

$$\hat{y}_{local}(\mathbf{x}) = \lambda\mathbf{l}^\top(\mathbf{x})(\mathbf{G} + \lambda\mathbf{L})^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}). \quad (10)$$

It is important to note that, since the lower bounds for $\boldsymbol{\alpha}$ in (7) are usually set to be moderately large, the off-diagonal elements in \mathbf{L} are closer to zero. Particularly, we can obtain $\mathbf{L} \rightarrow \mathbf{I}$ when $\boldsymbol{\alpha}$ take very large values. This immediately suggests two interesting properties for the CGP model. First, its global trend predictor $\hat{y}_{global}(\mathbf{x})$ in (9) resembles a kriging predictor with nugget effect as $\mathbf{L} \rightarrow \mathbf{I}$. When $\lambda > 0$, this nugget predictor is smooth but non-interpolating, and is commonly used in spatial statistics for modeling observational data with noise (Cressie 1991). Secondly, since $\mathbf{L} \approx \mathbf{I}$, the λ in $(\mathbf{G} + \lambda\mathbf{L})$ is mainly added to the diagonal elements. This makes $(\mathbf{G} + \lambda\mathbf{L})$ resistant to become ill-conditioned and the computation of $(\mathbf{G} + \lambda\mathbf{L})^{-1}$ in CGP can be numerically very stable. These two properties are elaborated in detail in Section 5.

3.2 Improving both the mean and variance models

To further relax the constant variance restriction, we introduce a variance model $\sigma^2(\mathbf{x})$ into (5) as follows

$$Y(\mathbf{x}) = Z_{global}(\mathbf{x}) + \sigma(\mathbf{x})Z_{local}(\mathbf{x}), \quad (11)$$

$$Z_{global}(\mathbf{x}) \sim GP(\mu, \tau^2 g(\cdot)),$$

$$Z_{local}(\mathbf{x}) \sim GP(0, l(\cdot)).$$

The $Z_{global}(\mathbf{x})$ above remains the same as in (5), since the global trend is smooth and can reasonably be assumed to be stationary. After subtracting $Z_{global}(\mathbf{x})$ from the response, the

second process is augmented with a variance model to quantify the change of local variability such that $\sigma(\mathbf{x})Z_{local}(\mathbf{x}) \sim GP(0, \sigma^2(\mathbf{x})l(\cdot))$. Overall, the model form in (11) is equivalent to assuming that the response $Y(\mathbf{x}) \sim GP(\mu, \tau^2g(\cdot) + \sigma^2(\mathbf{x})l(\cdot))$.

Without loss of generality, suppose the variance model can be expressed as $\sigma^2(\mathbf{x}) = \sigma^2v(\mathbf{x})$, where σ^2 is an unknown variance constant and $v(\mathbf{x})$ is the standardized volatility function which fluctuates around the unit value. In the following discussion, we first assume that $v(\mathbf{x})$ is known, and denote $\Sigma = \text{diag}\{v(\mathbf{x}_1), \dots, v(\mathbf{x}_n)\}$ to represent the standardized local variances at each of the design points $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. An efficient strategy for obtaining the $v(\mathbf{x})$ function is presented at the end of this section.

The model assumptions in (11) suggest that $y(\mathbf{x})$ and $\mathbf{y} = (y_1, \dots, y_n)^\top$ have the multivariate normal distribution

$$\begin{pmatrix} y(\mathbf{x}) \\ \mathbf{y} \end{pmatrix} \sim N_{1+n} \left[\begin{pmatrix} \mu \\ \mu \mathbf{1} \end{pmatrix}, \begin{pmatrix} \tau^2 + \sigma^2v(\mathbf{x}) & (\tau^2\mathbf{g}(\mathbf{x}) + \sigma^2v^{1/2}(\mathbf{x})\Sigma^{1/2}\mathbf{l}(\mathbf{x}))^\top \\ \tau^2\mathbf{g}(\mathbf{x}) + \sigma^2v^{1/2}(\mathbf{x})\Sigma^{1/2}\mathbf{l}(\mathbf{x}) & \tau^2\mathbf{G} + \sigma^2\Sigma^{1/2}\mathbf{L}\Sigma^{1/2} \end{pmatrix} \right]. \quad (12)$$

The best linear unbiased predictor under these assumptions can be derived as

$$\begin{aligned} \hat{y}(\mathbf{x}) &= \hat{\mu} + (\tau^2\mathbf{g}(\mathbf{x}) + \sigma^2v^{1/2}(\mathbf{x})\Sigma^{1/2}\mathbf{l}(\mathbf{x}))^\top (\tau^2\mathbf{G} + \sigma^2\Sigma^{1/2}\mathbf{L}\Sigma^{1/2})^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}) \\ &= \hat{\mu} + (\mathbf{g}(\mathbf{x}) + \lambda v^{1/2}(\mathbf{x})\Sigma^{1/2}\mathbf{l}(\mathbf{x}))^\top (\mathbf{G} + \lambda\Sigma^{1/2}\mathbf{L}\Sigma^{1/2})^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}), \end{aligned} \quad (13)$$

where $\lambda = \sigma^2/\tau^2$ ($\lambda \in [0, 1]$), $\hat{\mu} = (\mathbf{1}^\top(\mathbf{G} + \lambda\Sigma^{1/2}\mathbf{L}\Sigma^{1/2})^{-1}\mathbf{1})^{-1}\mathbf{1}^\top(\mathbf{G} + \lambda\Sigma^{1/2}\mathbf{L}\Sigma^{1/2})^{-1}\mathbf{y}$ and all the other notations remain the same as in (6). Note that after defining the ratio λ , the unknown σ^2 is no longer needed for prediction, because the predictor depends on the variance model $\sigma^2(\mathbf{x})$ only through λ and $v(\mathbf{x})$. The predictor includes (6) as a special case when the local volatility model $v(\mathbf{x})$ degenerates to a constant function. The predictor can also interpolate all the data points since $(\mathbf{g}(\mathbf{x}_i) + \lambda v^{1/2}(\mathbf{x}_i)\Sigma^{1/2}\mathbf{l}(\mathbf{x}_i))^\top (\mathbf{G} + \lambda\Sigma^{1/2}\mathbf{L}\Sigma^{1/2})^{-1} = \mathbf{e}_i^\top$ and $\hat{y}(\mathbf{x}_i) = \hat{\mu} + \mathbf{e}_i^\top(\mathbf{y} - \hat{\mu}\mathbf{1}) = y_i$ for $i = 1, \dots, n$. By decomposing the predictor (13) into

two parts

$$\hat{\mathbf{y}}(\mathbf{x}) = \hat{\mathbf{y}}_{global}(\mathbf{x}) + \hat{\mathbf{y}}_{local}(\mathbf{x}), \quad (14)$$

$$\hat{\mathbf{y}}_{global}(\mathbf{x}) = \hat{\mu} + \mathbf{g}^\top(\mathbf{x})(\mathbf{G} + \lambda \Sigma^{1/2} \mathbf{L} \Sigma^{1/2})^{-1}(\mathbf{y} - \hat{\mu} \mathbf{1}), \quad (15)$$

$$\hat{\mathbf{y}}_{local}(\mathbf{x}) = \lambda v^{1/2}(\mathbf{x}) \mathbf{l}^\top(\mathbf{x}) \Sigma^{1/2} (\mathbf{G} + \lambda \Sigma^{1/2} \mathbf{L} \Sigma^{1/2})^{-1}(\mathbf{y} - \hat{\mu} \mathbf{1}), \quad (16)$$

we can see that the global trend $\hat{\mathbf{y}}_{global}(\mathbf{x})$ in (15) reduces to a *stochastic kriging* predictor (Ankenman, Nelson and Staum 2010) when $\mathbf{L} \rightarrow \mathbf{I}$. Different from the nugget predictor in (9) where a universal term λ is used for adjusting the global trend throughout the whole region, the amount of shrinkage at each data point in (15) is proportional to the value of $\lambda v(\mathbf{x}_i)$. This *localized adjustment* scheme is advantageous in making the global trend smoother and more stable, since it is less affected by the data points with large variability.

The above predictor form is derived based on $Y(\mathbf{x}) \sim GP(\mu, \tau^2 g(\cdot) + \sigma^2(\mathbf{x})l(\cdot))$, which unifies the modeling assumptions (11) in a *single stage*. As a result, the new method can also be viewed as extending the kriging model with a non-stationary covariance structure $\tau^2 g(\cdot) + \sigma^2(\mathbf{x})l(\cdot)$. Different from this, another strategy to fulfill the new assumptions in (11) is to develop the global and local models *sequentially*: (i) Fit a global trend model as in (15) using the likelihood method. (ii) Obtain its residuals $\mathbf{s} = (\mathbf{y} - \hat{\mathbf{y}}_{global})$, where $\hat{\mathbf{y}}_{global} = (\hat{\mathbf{y}}_{global}(\mathbf{x}_1), \dots, \hat{\mathbf{y}}_{global}(\mathbf{x}_n))^\top$. If the estimated global trend interpolates all the data points ($\hat{\lambda} = 0$), we have $\mathbf{s} = \mathbf{0}$ and in this case the CGP just degenerates to a traditional single GP model. (iii) If $\mathbf{s} \neq \mathbf{0}$, standardize the residuals to achieve variance homogeneity $\mathbf{s}^* = \Sigma^{-1/2} \mathbf{s}$. (iv) Adjust the global trend by interpolating the standardized residuals via a simple kriging model $\hat{\mathbf{y}}_{adj}(\mathbf{x}) = \mathbf{l}^\top(\mathbf{x}) \mathbf{L}^{-1} \mathbf{s}^*$. In this way, we can form a sequential predictor as

$$\hat{\mathbf{y}}_{seq}(\mathbf{x}) = \hat{\mathbf{y}}_{global}(\mathbf{x}) + v^{1/2}(\mathbf{x}) \hat{\mathbf{y}}_{adj}(\mathbf{x}) = \hat{\mathbf{y}}_{global}(\mathbf{x}) + v^{1/2}(\mathbf{x}) \mathbf{l}^\top(\mathbf{x}) \mathbf{L}^{-1} \mathbf{s}^*. \quad (17)$$

It is of natural interest to ask whether this sequential predictor would make any difference from the single-stage predictor (13), and the following theorem establishes their connections.

Theorem 1. *Given the same parameter values, the single-stage predictor (13) and the sequential predictor (17) are equivalent.*

Proof of the theorem is left in the Appendix. Despite this equivalent model form, we want to emphasize that the single-stage fitting strategy is superior to the sequential one in parameter estimation. This is because all parameters in the single-stage predictor (13) can be optimized simultaneously, which takes into account the interactions between global and local models and automatically balances their effects. In contrast to this global optimization, the sequential fitting approach estimates the parameters in two separate steps, and each of them can at most achieve local optimality. Generally, the global trend is hard to identify correctly without considering the effects of the second stage model, and in many cases the performance of the final prediction can be quite sensitive to this “global-local tradeoff”. As a result, in this paper we only consider the single-stage modeling framework, and this is also a major advantage for the proposed method over other multi-step strategies such as blind kriging.

In the rest of this section, we present how to obtain the $v(\mathbf{x})$ function, which is required for the CGP predictor. As shown in (14), the CGP model can be decomposed into a global and a local component, and this structure provides us a convenient way to assess the change of local volatility. For a given global trend (15) (initially we can set $\Sigma = \mathbf{I}$), its squared residuals $\mathbf{s}^2 = (s_1^2, \dots, s_n^2)^\top$ are natural measures of the local volatility, which can be used as the bases to build the $v(\mathbf{x})$ function. Based on \mathbf{s}^2 , we propose an intuitive *Gaussian kernel regression model* for $v(\mathbf{x})$ as:

$$v(\mathbf{x}) = \frac{\mathbf{g}_b^\top(\mathbf{x})\mathbf{s}^2}{\mathbf{g}_b^\top(\mathbf{x})\mathbf{1}}, \quad (18)$$

where $\mathbf{g}_b(\mathbf{x}) = (g_b(\mathbf{x} - \mathbf{x}_1), \dots, g_b(\mathbf{x} - \mathbf{x}_n))^\top$ with $g_b(\mathbf{h}|\boldsymbol{\theta}, b) = \exp(-b \sum_{j=1}^p \theta_j h_j^2)$. Here $\boldsymbol{\theta}$ are the correlation parameters used in the global trend (15), $b \in [0, 1]$ is an extra bandwidth parameter such that $\mathbf{g}_b(\mathbf{x}) \rightarrow \mathbf{1}$ as $b \rightarrow 0$, and $\mathbf{g}_b(\mathbf{x}) = \mathbf{g}(\mathbf{x})$ if $b = 1$. Since $\mathbf{g}(\mathbf{x})$ is the correlation of the global trend, the underlying assumption behind (18) is that whenever two points in the global trend are strongly correlated, their variances also tend to be more related. The bandwidth parameter b adds additional flexibility in controlling the smoothness of the variance function: when equaling zero, it smoothes out $v(\mathbf{x})$ to a constant function even if the global trend is not flat.

From the $v(\mathbf{x})$ model in (18), we can evaluate $\hat{v}_i = v(\mathbf{x}_i)$ for $i = 1, \dots, n$ and update the matrix $\mathbf{\Sigma} = \text{diag}\{\hat{v}_1, \dots, \hat{v}_n\}$. Since $v(\mathbf{x})$ and $\mathbf{\Sigma}$ are the standardized local volatilities, we also need to re-scale them as

$$\mathbf{\Sigma} \leftarrow \mathbf{\Sigma} / \left(\frac{1}{n} \sum_{i=1}^n \hat{v}_i \right) \quad \text{and} \quad v(\mathbf{x}) \leftarrow v(\mathbf{x}) / \left(\frac{1}{n} \sum_{i=1}^n \hat{v}_i \right). \quad (19)$$

This standardization makes the diagonal elements of $\mathbf{\Sigma}$ having unit mean, which is essential for keeping the ratio of σ^2 to τ^2 consistent in the global trend. By plugging the updated (and standardized) $\mathbf{\Sigma}$ back into (15), we can repeat the above process for a few more times. Usually three or four iterations are sufficient to stabilize the volatility estimates. This iterative estimation for variance is similar in spirit to the *iteratively reweighted least squares* method in classical regression.

Before concluding this section, we want to emphasize that the estimation of $v(\mathbf{x})$ does not need to be separately carried out before fitting the CGP model; instead, it can be seamlessly nested as an inner loop in estimating the whole model. The $v(\mathbf{x})$ function above is uniquely determined by the unknown parameters $\boldsymbol{\theta}$ and b . Since its correlation parameter $\boldsymbol{\theta}$ are always paired and synchronized with that of the global trend, inclusion of this volatility function $v(\mathbf{x})$ only adds one more parameter b to the whole model.

4 Estimation

In this section, we derive maximum-likelihood estimators (MLEs) for the unknown parameters in the CGP model. As suggested at the end of previous section, given each set of $(\lambda, \mu, \tau^2, \boldsymbol{\theta}, \boldsymbol{\alpha}, b)$ values, $v(\mathbf{x})$ and $\mathbf{\Sigma} = \text{diag}\{\hat{v}_1, \dots, \hat{v}_n\}$ values can be uniquely determined by nesting a small inner loop in the likelihood function.

Based on the multivariate normal assumptions in Section 3.2, the log-likelihood function (up to an additive constant) can be written as

$$\begin{aligned} l(\mu, \tau^2, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\alpha}, b) \\ = -\frac{1}{2} \log(\det(\tau^2 \mathbf{G} + \sigma^2 \mathbf{\Sigma}^{1/2} \mathbf{L} \mathbf{\Sigma}^{1/2})) - \frac{1}{2} (\mathbf{y} - \mu \mathbf{1})^\top (\tau^2 \mathbf{G} + \sigma^2 \mathbf{\Sigma}^{1/2} \mathbf{L} \mathbf{\Sigma}^{1/2})^{-1} (\mathbf{y} - \mu \mathbf{1}). \end{aligned}$$

Due to the invariant property of MLE under transformations, we can re-parameterize $\lambda = \sigma^2/\tau^2$ in the log-likelihood as

$$l(\lambda, \mu, \tau^2, \boldsymbol{\theta}, \boldsymbol{\alpha}, b) \tag{20}$$

$$= -\frac{1}{2}[n \log(\tau^2) + \log(\det(\mathbf{G} + \lambda \boldsymbol{\Sigma}^{1/2} \mathbf{L} \boldsymbol{\Sigma}^{1/2})) + (\mathbf{y} - \mu \mathbf{1})^\top (\mathbf{G} + \lambda \boldsymbol{\Sigma}^{1/2} \mathbf{L} \boldsymbol{\Sigma}^{1/2})^{-1} (\mathbf{y} - \mu \mathbf{1}) / \tau^2].$$

Since $\boldsymbol{\Sigma} = \text{diag}\{\hat{v}_1, \dots, \hat{v}_n\}$ can be known through the procedures presented in the last section, the MLEs for μ and τ^2 can be easily derived from (20) as

$$\hat{\mu}(\lambda, \boldsymbol{\theta}, \boldsymbol{\alpha}, b) = (\mathbf{1}^\top (\mathbf{G} + \lambda \boldsymbol{\Sigma}^{1/2} \mathbf{L} \boldsymbol{\Sigma}^{1/2})^{-1} \mathbf{1})^{-1} (\mathbf{1}^\top (\mathbf{G} + \lambda \boldsymbol{\Sigma}^{1/2} \mathbf{L} \boldsymbol{\Sigma}^{1/2})^{-1} \mathbf{y}), \tag{21}$$

$$\hat{\tau}^2(\lambda, \boldsymbol{\theta}, \boldsymbol{\alpha}, b) = \frac{1}{n} (\mathbf{y} - \hat{\mu} \mathbf{1})^\top (\mathbf{G} + \lambda \boldsymbol{\Sigma}^{1/2} \mathbf{L} \boldsymbol{\Sigma}^{1/2})^{-1} (\mathbf{y} - \hat{\mu} \mathbf{1}). \tag{22}$$

After substituting these values into (20), we can obtain the MLEs for $(\lambda, \boldsymbol{\theta}, \boldsymbol{\alpha}, b)$ by minimizing the following (negative) log profile likelihood

$$\phi(\lambda, \boldsymbol{\theta}, \boldsymbol{\alpha}, b) = n \log(\hat{\tau}^2(\lambda, \boldsymbol{\theta}, \boldsymbol{\alpha}, b)) + \log(\det(\mathbf{G} + \lambda \boldsymbol{\Sigma}^{1/2} \mathbf{L} \boldsymbol{\Sigma}^{1/2})), \tag{23}$$

where $\lambda \in [0, 1], b \in [0, 1], \theta_j \in [0, \alpha^l]$ and $\alpha_j \in [\alpha^l, \infty]$ for $j = 1, \dots, p$.

For p input variables, the above likelihood function contains $2p + 2$ unknown parameters. Compared to the stationary GP model whose likelihood contains only p unknown parameters, the CGP model becomes more difficult to estimate when the input dimension p gets large. To mitigate this disadvantage, we can further assume

$$\alpha_j = \theta_j + \kappa, \quad j = 1, \dots, p, \tag{24}$$

for the correlation parameters, which reduces the p unknown parameters $(\alpha_1, \dots, \alpha_p)$ into a single κ . By substituting (24) into (23), the CGP only involves $p + 3$ unknown parameters $(\lambda, \boldsymbol{\theta}, \kappa, b)$, whose MLEs can be obtained by minimizing

$$\phi(\lambda, \boldsymbol{\theta}, \kappa, b) = n \log(\hat{\tau}^2(\lambda, \boldsymbol{\theta}, \kappa, b)) + \log(\det(\mathbf{G} + \lambda \boldsymbol{\Sigma}^{1/2} \mathbf{L} \boldsymbol{\Sigma}^{1/2})), \tag{25}$$

subject to the constraints $\lambda \in [0, 1], b \in [0, 1], \kappa \in [\alpha^l, \infty]$ and $\theta_j \in [0, \alpha^l]$ for $j = 1, \dots, p$.

We now provide a general guideline for choosing the bound α^l . The idea is to specify the value of α^l based on the space-filling properties of the design points. Suppose the design

$D = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ has been standardized into the unit region of $[0, 1]^p$, and then define the following harmonic-type average inter-point distance d_{avg} to measure its space-filling properties (Ba and Joseph 2011)

$$d_{avg} = \left(\frac{2}{n(n-1)} \sum_{1 \leq i < k \leq n} \frac{1}{d(\mathbf{x}_i, \mathbf{x}_k)^2} \right)^{-\frac{1}{2}},$$

where $d(\mathbf{x}_i, \mathbf{x}_k) = \sqrt{(\sum_{j=1}^p (x_{ij} - x_{kj})^2)}$. When we assume $\theta_j = \theta$ and $\alpha_j = \alpha$ ($j = 1, \dots, p$) in the Gaussian correlation functions (7), correlations between points with distance d_{avg} are $g(\theta) = \exp(-\theta d_{avg}^2)$ and $l(\alpha) = \exp(-\alpha d_{avg}^2)$ for the global and local processes respectively. Because $\exp(-\alpha d_{avg}^2) \leq \exp(-\alpha^l d_{avg}^2) \leq \exp(-\theta d_{avg}^2)$, our recommendation for choosing α^l is to set $\exp(-\alpha^l d_{avg}^2) = 0.01$, which leads to

$$\alpha^l = \frac{\log 100}{d_{avg}^2}. \quad (26)$$

This bound is used for estimation throughout the paper.

5 Properties

5.1 Improved prediction for sparse dataset

As discussed in Section 1, the ordinary kriging predictor tends to revert to the global mean in regions where data are not available. This erratic phenomenon will be even more pronounced if the design points are sparse and cannot cover the input region reasonably well. The new predictor, however, relaxes the constant mean restriction in ordinary kriging and introduces another GP for modeling the mean. This global trend (mean model) is non-interpolating but smooth, which makes it immune to the erratic reversion problem in the data sparse region. Consider again the simple test function in Figure 1, where the ordinary kriging predictor ($\hat{\theta} = 400$) appears to be erratic. When the proposed CGP model is fitted ($\hat{\lambda} = 0.07, \hat{\theta} = 143.6, \hat{\alpha} = 1892.1, \hat{b} = 1$), its global trend is shown as the dotted line in Figure 2. Although it incurs large errors around data points in region $x \in [0, 0.4]$, it behaves well in the sparse region $[0.4, 1]$ due to the smoothness property. The final CGP predictor

after incorporating the local trend is shown as the dashed line in Figure 2. It can be seen that this predictor eliminates all the non-interpolating errors at design points. At locations far from data points, it tends to revert to the smooth global trend instead of a global constant, which avoids the erratic problem as in Figure 1 and yields much improved prediction. This shows the advantage of using the CGP predictor when data points are sparse in some parts of the design region. In practice, the sparseness of data points is quite common when input dimensions are high or a non-space-filling design is used.

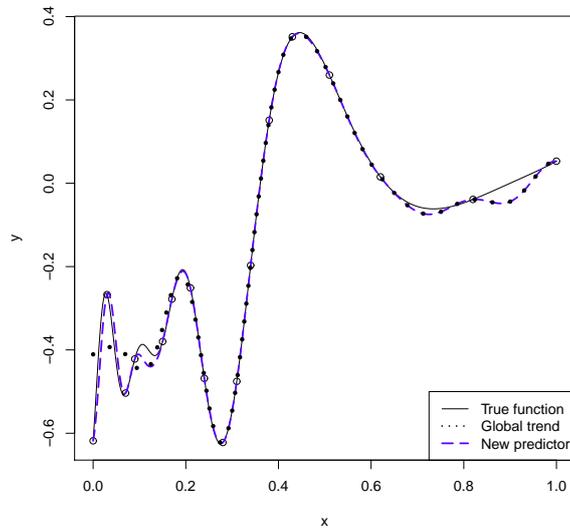


Figure 2: Plot of function $y(x) = \sin(30(x - 0.9)^4) \cos(2(x - 0.9)) + (x - 0.9)/2$, the global trend and the CGP predictor.

5.2 Numerical stability

One well-documented problem with the GP model is the potential numerical instability when computing the inverse of its $n \times n$ correlation matrix \mathbf{R} . This correlation matrix can easily become ill-conditioned, for example, when sample size n is large, design points are close to each other, or the sample points get highly correlated while we search for the optimal correlation parameters (Ababou, Bagtzoglou and Wood 1994, Haaland and Qian

2012, Peng and Wu 2012). A near-singular correlation matrix in kriging will lead to serious numerical problems, which causes the resulting predictor unstable and unreliable.

To overcome this ill-conditioned problem, the popular approach is to add a non-zero nugget to the diagonal elements of the correlation matrix such that $\mathbf{R} \rightarrow (\mathbf{R} + \lambda \mathbf{I})$. Because including non-zero nugget has the inevitable drawback of making predictors over-smooth (non-interpolating), in this approach we need to reconcile the gains in numerical stability with the losses in interpolation property, and choose a trade-off value for the nugget (Ranjan, Haynes and Karsten 2012, Peng and Wu 2012).

As shown at the end of Section 3.1, the correlation matrix to invert in the proposed CGP model is $(\mathbf{G} + \lambda \mathbf{L})$. (Cases after including the variance matrix $\mathbf{\Sigma}$ remain similar.) Since the lower bounds for $\boldsymbol{\alpha}$ in (7) are moderately large and we have $\mathbf{L} \approx \mathbf{I}$, the λ in $(\mathbf{G} + \lambda \mathbf{L})$ automatically inflates the diagonal elements of the correlation matrix so that it is naturally resistant to become singular. In addition, different from the previous nugget case, the CGP model is always an interpolator and the λ value here can be freely estimated. In fact, whenever a traditional GP model has to include a non-zero nugget for numerical reasons, the CGP model can always improve it at least by removing its non-interpolating errors with a augmented $Z_{local}(\mathbf{x})$. This potential improvement is shown in next subsection.

5.3 Connection with the nugget predictor

To emulate deterministic outputs from computer experiments, Gramacy and Lee (2011) advocate always including a non-zero nugget in the kriging predictor for reasons even beyond computations. They argue that when model assumptions are violated or data points are sparse, the traditional GP predictor may lead to unpleasant results. Although adding a non-zero nugget to the predictor incurs extra errors around data points, it can be crucial for fitting a well-behaved (i.e. smooth) surface and avoiding erratic predictions in the unknown region. In a variety of situations, Gramacy and Lee (2011) show that overall this non-interpolating predictor can achieve better prediction accuracy.

Interestingly, when the local process in CGP has zero correlation ($\mathbf{L} = \mathbf{I}$), its global

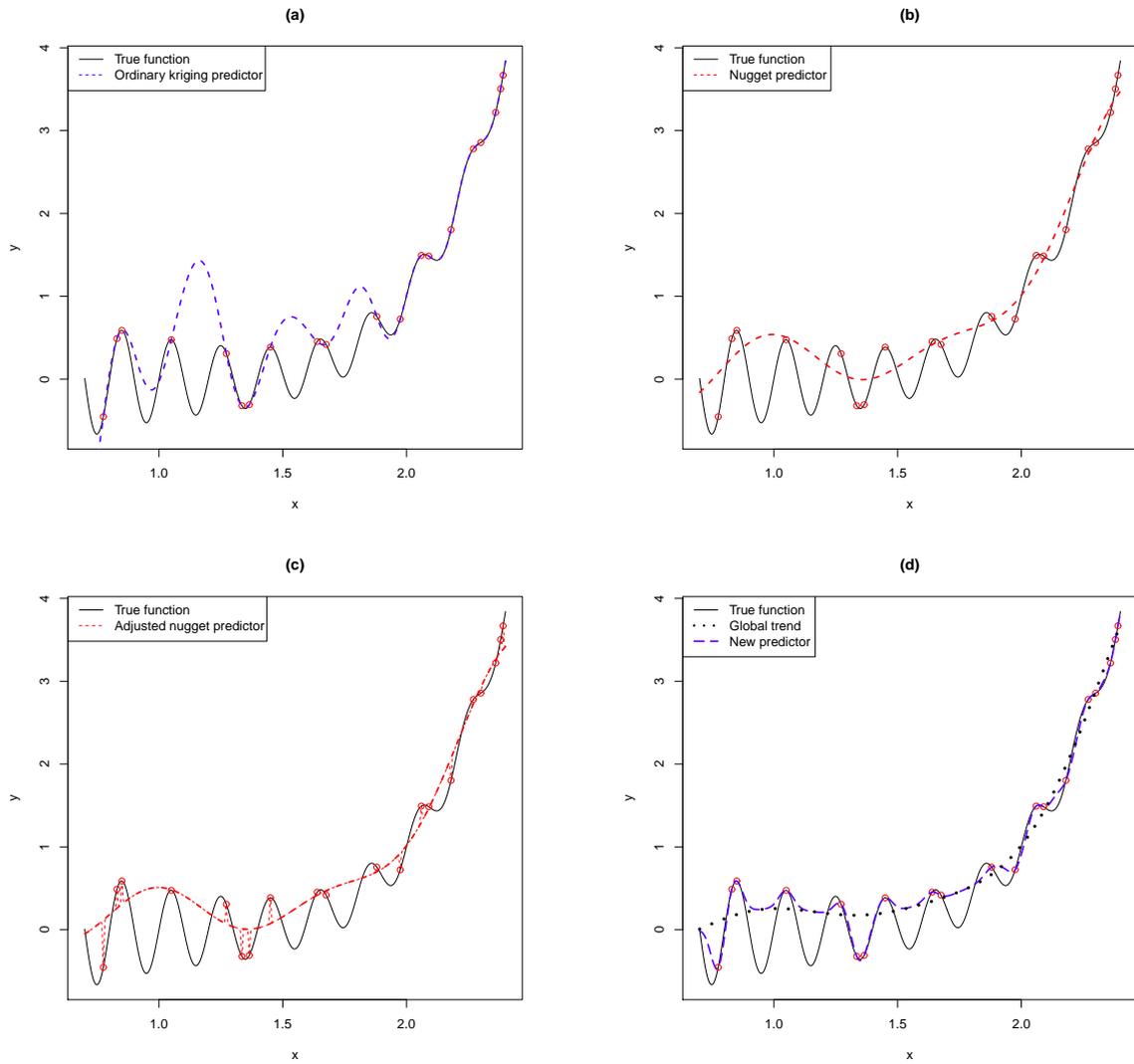


Figure 3: Plot of function $y(x) = \sin(10\pi x)/(2x) + (x - 1)^4$ with (a) the ordinary kriging predictor; (b) the kriging with nugget predictor; (c) the nugget predictor with adjustments around design points; (d) the optimized CGP predictor and its global trend.

trend just degenerates to a kriging predictor with nugget, and in this case the CGP predictor becomes $\hat{y}(\mathbf{x}) = \hat{y}_{nugget}(\mathbf{x}) + \hat{y}_{local}(\mathbf{x})$. In regions away from design points, since $\mathbf{l}(\mathbf{x}) = \mathbf{0}$ and $\hat{y}_{local}(\mathbf{x}) = 0$ for $\mathbf{x} \neq \mathbf{x}_i$ ($i = 1, 2, \dots, n$), the CGP model exactly matches the nugget predictor $\hat{y}_{nugget}(\mathbf{x})$. At the n design points, however, due to $\mathbf{l}(\mathbf{x}) = \mathbf{e}_i$ for $\mathbf{x} = \mathbf{x}_i$ ($i = 1, 2, \dots, n$), the $\hat{y}_{local}(\mathbf{x})$ still corrects the global trend and adjusts the CGP to interpolate all the data points. Just as the universal kriging generalizes the polynomial regression for interpolation, the CGP model can be similarly viewed as a generalization/improvement of the nugget predictor which eliminates errors at design points. When correlations in the local process of CGP are further estimated as positive, the above adjustments around data points tend to be continuous and smooth, which leads to a final CGP predictor inheriting the advantages from both the nugget predictor and the interpolating predictor.

Figure 3(a) demonstrates a simulated example from Gramacy and Lee (2011), where the test function $y(x) = \sin(10\pi x)/(2x) + (x-1)^4$ is evaluated at 20 unequally spaced locations to represent the sparseness of data points. Clearly, we can see that in this example the ordinary kriging predictor ($\hat{\theta} = 45.97$) makes predictions well outside the range of test function in many regions. The nugget predictor suggested by Gramacy and Lee (2011) is shown in Figure 3(b). Although non-interpolating, the nugget predictor overall gives smooth and reasonably good predictions, which reduces the *root mean squared prediction error* (RMSPE) from the previous 0.55 to 0.35. Here the $\text{RMSPE} = [\frac{1}{N} \sum_{i=1}^N \{\hat{y}(\mathbf{x}_i) - y(\mathbf{x}_i)\}^2]^{1/2}$ is computed based on $N = 5000$ randomly sampled data points from the design region. Now we further consider fitting the CGP model to this example. As shown in Figure 3(c), if we assume very small correlations in $Z_{local}(\mathbf{x})$, the new predictor remains almost the same as the nugget predictor within most regions; when it comes to around the design points, however, the predictor jumps to interpolate the data, which slightly reduces the RMSPE to 0.34. After we also fully estimate the correlations in $Z_{local}(\mathbf{x})$ and incorporate a variance model, Figure 3(d) gives the final CGP predictor ($\hat{\lambda} = 0.019, \hat{\theta} = 2.44, \hat{\alpha} = 578.09, \hat{b} = 1$), which is smooth and gives a RMSPE as low as 0.25.

5.4 Improved prediction intervals

Apart from prediction, another frequently noted drawback of ordinary kriging is the poor coverage of its prediction intervals (Yamamoto 2000, Xiong et al. 2007, Gramacy and Lee 2011, Joseph and Kang 2011). By assuming a constant variance σ^2 throughout the whole input region, the $(1 - \alpha)$ prediction interval at location \mathbf{x} for ordinary kriging is given by

$$\hat{y}(\mathbf{x}) \pm z_{\alpha/2}\sigma\left\{1 - \mathbf{r}^\top(\mathbf{x})\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}) + \frac{(1 - \mathbf{r}^\top(\mathbf{x})\mathbf{R}^{-1}\mathbf{1})^2}{\mathbf{1}^\top\mathbf{R}^{-1}\mathbf{1}}\right\}^{1/2},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ critical value of the standard normal distribution. This prediction interval is often too restrictive and inadequate to cover some complex underlying surfaces since it fails to take into account the change of local variability in the design region. One typical example is demonstrated in Figure 4(a), where the test function fluctuates around zero with decreasing amplitude. The corresponding prediction intervals from ordinary kriging ($\hat{\theta} = 24.6$), however, yield the same variability pattern throughout the whole design region, which are obviously too narrow to cover the high volatility region in the left part, but also end up unnecessarily wide in the right part of the input region where the true function is almost flat. In this subsection, we introduce the prediction intervals for CGP models. By relaxing the constant variance restriction, these prediction intervals are self-adjusted according to the local variability, and can be expected to give much improved coverage.

In a Bayesian framework, the assumptions for a CGP model in (11) can be viewed as putting a prior distribution $y(\mathbf{x})|\mu \sim GP(\mu, \tau^2 g(\cdot) + \sigma^2(\mathbf{x})l(\cdot))$ on the function, which leads to the first-stage conditional distribution

$$\begin{pmatrix} y(\mathbf{x}) \\ \mathbf{y} \end{pmatrix} \Big| \mu \sim N_{1+n} \left[\begin{pmatrix} \mu \\ \mu \mathbf{1} \end{pmatrix}, \tau^2 \begin{pmatrix} 1 + \lambda v(\mathbf{x}) & \mathbf{q}^\top(\mathbf{x}) \\ \mathbf{q}(\mathbf{x}) & \mathbf{Q} \end{pmatrix} \right],$$

where $\lambda = \sigma^2/\tau^2$, $\mathbf{q}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) + \lambda v^{1/2}(\mathbf{x})\boldsymbol{\Sigma}^{1/2}\mathbf{l}(\mathbf{x})$, $\mathbf{Q} = \mathbf{G} + \lambda\boldsymbol{\Sigma}^{1/2}\mathbf{L}\boldsymbol{\Sigma}^{1/2}$ and all the other notations remain the same as in Section 3.2. Here for simplicity, the variance and correlation parameters are assumed to be known. If we further assume a second-stage noninformative prior for μ : $p(\mu) \sim 1$ and integrate it out, then the predictive distribution for $y(\mathbf{x})$ can be derived as

$$y(\mathbf{x})|\mathbf{y} \sim N_1(\mu_{0|n}(\mathbf{x}), v_{0|n}^2(\mathbf{x}))$$

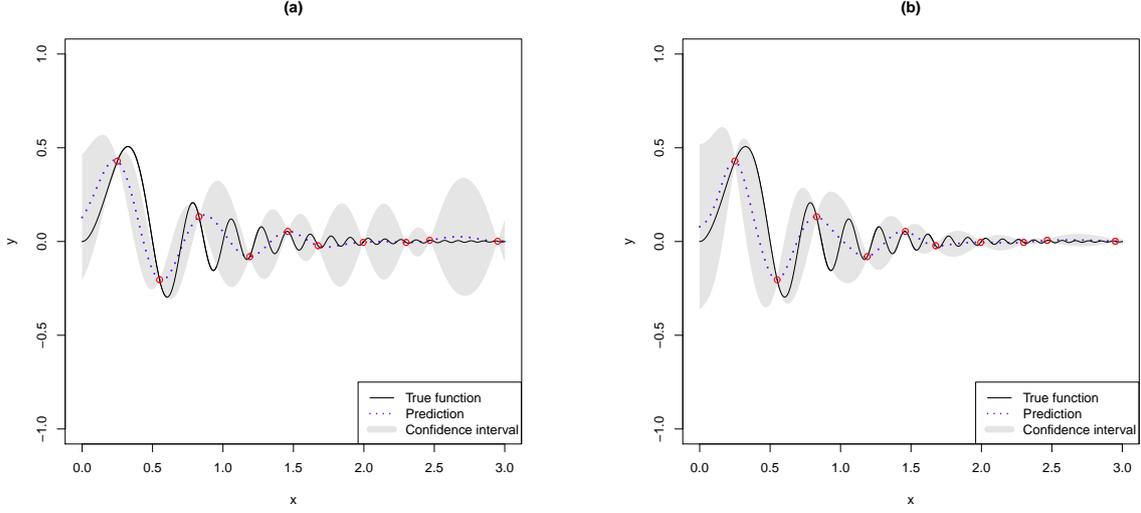


Figure 4: Plot of function $y(x) = \exp(-2x) \sin(4\pi x^2)$ and the prediction intervals from (a) ordinary kriging; (b) the CGP model.

where

$$\mu_{0|n}(\mathbf{x}) = \hat{\boldsymbol{\mu}} + \mathbf{q}^\top(\mathbf{x})\mathbf{Q}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}\mathbf{1}) \quad \text{for} \quad \hat{\boldsymbol{\mu}} = (\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{1})^{-1}(\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{y}),$$

and

$$v_{0|n}^2(\mathbf{x}) = \tau^2 \left\{ 1 + \lambda v(\mathbf{x}) - \mathbf{q}^\top(\mathbf{x})\mathbf{Q}^{-1}\mathbf{q}(\mathbf{x}) + \frac{(1 - \mathbf{q}^\top(\mathbf{x})\mathbf{Q}^{-1}\mathbf{1})^2}{\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{1}} \right\}. \quad (27)$$

The derivation for these results is tedious but standard, which follows similar development steps as in Santer et al. (2003, Chapter 4.3). It can be seen that our previously proposed predictor in (13) is nothing but the posterior mean of the function given the data. Now a (pointwise) prediction interval for this predictor can be constructed by

$$\hat{y}(\mathbf{x}) \pm z_{\alpha/2} v_{0|n}(\mathbf{x}), \quad (28)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ critical value of the standard normal distribution.

Note that, since $\mathbf{q}^\top(\mathbf{x}_i)\mathbf{Q}^{-1} = \mathbf{e}_i^\top$ and $\mathbf{e}_i^\top\mathbf{q}(\mathbf{x}_i) = 1 + \lambda v(\mathbf{x}_i)$, the above posterior variance $v_{0|n}^2(\mathbf{x})$ equals zero whenever $\mathbf{x} = \mathbf{x}_i$ for $i = 1, \dots, n$. Thus, as in ordinary kriging, the width of the prediction interval shrinks to zero at each data point, which is quite intuitive since

both models interpolate the responses at each observed location. On the other hand, however, different from ordinary kriging, the variance of predictive distribution in (27) depends on the local variability of the underlying surface, which intrinsically adjusts the widths of the prediction interval. Consider again the test function in Figure 4. It can be seen in Figure 4(b) that the prediction intervals from a CGP model ($\hat{\theta} = 2.1, \hat{\alpha} = 54.85, \hat{\lambda} = 1, \hat{b} = 1$) become much wider in the left region when the function fluctuates rapidly, but quickly narrow down as the underlying function becomes flat. Compared with the prediction intervals for ordinary kriging, the new intervals can more precisely demonstrate the change of prediction uncertainties throughout the input region: i.e. the predictive variances are much larger in the left part of region than in the right. One way to quantify such improvements is through computing the *interval score* for central prediction intervals (Gneiting and Raftery 2007) which is defined as $S_{\alpha}^{int}(l, u; x) = (u - l) + \frac{2}{\alpha}(l - x)\mathbb{1}\{x < l\} + \frac{2}{\alpha}(x - u)\mathbb{1}\{x > u\}$ for a $(1 - \alpha)\%$ central prediction interval $[l, u]$. This scoring rule (to be minimized) rewards narrow prediction intervals and also penalizes lack of coverage. For the prediction intervals in Figure 4, the average interval score (based on 3000 randomly sampled test points) for the ordinary kriging in (a) is 0.62 while for the CGP model in (b) is only 0.32, which shows almost 50% improvements.

5.5 Extensions to noisy data

In the previous sections, we model the deterministic outputs from a computer experiment by coupling two GPs. As an extension to this, sometimes it is also possible to use the sum of more than two GPs for gaining additional flexibility in the model and satisfying special needs. One important application of this extension is to modify the new predictor for modeling data with random errors.

Based on the previous model form in Section 3.2, we can add a third GP (with zero correlation) to account for the white noise as follows

$$Y(\mathbf{x}) = Z_{global}(\mathbf{x}) + \sigma(\mathbf{x})Z_{local}(\mathbf{x}) + \varepsilon(\mathbf{x}),$$

where $Z_{global}(\mathbf{x})$, $Z_{local}(\mathbf{x})$ are the same stationary GPs as in (11), and the error term $\varepsilon(\mathbf{x})$ is

assumed to be $N(0, \sigma_\varepsilon^2(\mathbf{x}))$ distributed, uncorrelated at different input locations and also independent of the other two GPs. Suppose the error variances $\Sigma_\varepsilon = \text{diag}\{\sigma_\varepsilon^2(\mathbf{x}_1), \dots, \sigma_\varepsilon^2(\mathbf{x}_n)\}$ are given, then the best linear unbiased predictor can be easily updated by modifying (13) as follows

$$\begin{aligned}\hat{\mathbf{y}}(\mathbf{x}) &= \hat{\mu} + (\tau^2 \mathbf{g}(\mathbf{x}) + \sigma^2 v^{1/2}(\mathbf{x}) \Sigma^{1/2} \mathbf{l}(\mathbf{x}))^\top (\tau^2 \mathbf{G} + \sigma^2 \Sigma^{1/2} \mathbf{L} \Sigma^{1/2} + \Sigma_\varepsilon)^{-1} (\mathbf{y} - \hat{\mu} \mathbf{1}) \\ &= \hat{\mu} + (\mathbf{g}(\mathbf{x}) + \lambda v^{1/2}(\mathbf{x}) \Sigma^{1/2} \mathbf{l}(\mathbf{x}))^\top (\mathbf{G} + \lambda \Sigma^{1/2} \mathbf{L} \Sigma^{1/2} + \rho \Sigma_\varepsilon)^{-1} (\mathbf{y} - \hat{\mu} \mathbf{1}),\end{aligned}$$

where $\rho = 1/\tau^2$, $\hat{\mu} = (\mathbf{1}^\top (\mathbf{G} + \lambda \Sigma^{1/2} \mathbf{L} \Sigma^{1/2} + \rho \Sigma_\varepsilon)^{-1} \mathbf{1})^{-1} (\mathbf{1}^\top (\mathbf{G} + \lambda \Sigma^{1/2} \mathbf{L} \Sigma^{1/2} + \rho \Sigma_\varepsilon)^{-1} \mathbf{y})$ and all the other notations remain the same as in (13). This predictor for noisy data is no longer an interpolator, and its parameter estimation can be similarly carried out as in the previous sections, except for $(\mathbf{G} + \lambda \Sigma^{1/2} \mathbf{L} \Sigma^{1/2})$ replaced by $(\mathbf{G} + \lambda \Sigma^{1/2} \mathbf{L} \Sigma^{1/2} + \rho \Sigma_\varepsilon)$ in the models.

6 Examples

Example 1. For any non-stationary modeling approach, one commonly raised concern is that if the true surface is indeed a realization from a stationary Gaussian process, whether the “unnecessarily sophisticated” non-stationary modeling approach can perform as good as the “correct” stationary model. To test the performance of our proposed model in such cases, we simulate sample paths from various two-dimensional stationary Gaussian processes for 50 times, and fit both the CGP and the stationary GP models to each of them for comparison. A 24-run maximin distance Latin Hypercube Design (LHD) is used in these simulations, and for each time the true correlation parameters in GP are randomly generated from [1, 5]. In each iteration, once the design and correlation parameters are fixed, a 24×24 correlation matrix \mathbf{R} is uniquely determined. A sample path from the corresponding stationary GP can then be drawn by simulating a random sample vector from the multivariate normal distribution $N_n(\mu \mathbf{1}^n, \sigma^2 \mathbf{R}^n)$ with $n = 24$, $\mu = 0$, $\sigma^2 = 1$.

After drawing stationary sample paths as above for 50 times, we fit CGP models to each of them. Among the 50 fitted models, 42 out of them have $\hat{\lambda} = 0$, which shows that the

Table 1: RMSPE values for three predictors based on 5000 testing data.

Method	Maximin LHD	Adaptive Design
GP	0.188	0.266
CGP	0.144	0.159
TGP	0.312	0.465

CGP has perfectly degenerated to the stationary GP model. For the other eight CGP models, their $\hat{\lambda}$ values are also extremely small, with the largest one only as 0.003. Measured by the leave-one-out cross validation error, the prediction accuracy of CGP model and the stationary GP model are almost identical in these cases.

Example 2. In this example, we provide two test functions possessing non-stationary features: one in two dimensions and the other has 10-dimensional inputs. The first function is the two-dimensional $f(x_1, x_2) = \sin(1/(x_1x_2))$, $(x_1, x_2 \in [0.3, 1])$, whose surface fluctuates rapidly when x_1 or x_2 is small, but gradually becomes smooth as x_1 and x_2 increase toward one. The second test function (known as the Michalewicz’s function) is in 10 dimensions, which has the following form:

$$f(\mathbf{x}) = - \sum_{i=1}^{10} \sin(x_i) \left[\sin\left(\frac{ix_i^2}{\pi}\right) \right]^{2m}, 0 \leq x_i \leq \pi, i = 1, \dots, 10.$$

Typically, this function is used with $m = 10$, which leads to a high dimensional surface containing many local optima, and its volatility varies dramatically throughout the input region.

We use a 24-run maximin distance LHD and a 24-run adaptive design from Xiong et. al (2007) to evaluate the first test function. Both the GP and CGP models are fitted to these two designs, and their RMSPEs are compared based on additional 5000 randomly sampled testing data. From the results in Table 1, we can see that the CGP predictor improves the accuracy of the GP predictor by 23% and 40% for each design. In Table 1, we also fit the Bayesian treed Gaussian process (TGP) model (Gramacy and Lee 2008) to the two designs

for comparison. The RMSPEs of this non-stationary treed model are relatively large, which probably are due to its inefficient partitioning of the input region.

To further test the performance of CGP predictor based on different designs, we generate fifty 100-run random LHDs to evaluate the second test function, and fit the GP and CGP models to each of them. RMSPEs of the two predictors are plotted in Figure 5 for the 50 random designs. It can be seen that, compared to the GP model, the CGP predictor can always give better approximations to this complex surface based on any random LHD. The RMSPEs of the two predictors based on a 100-run maximin distance LHD are also marked in this plot.

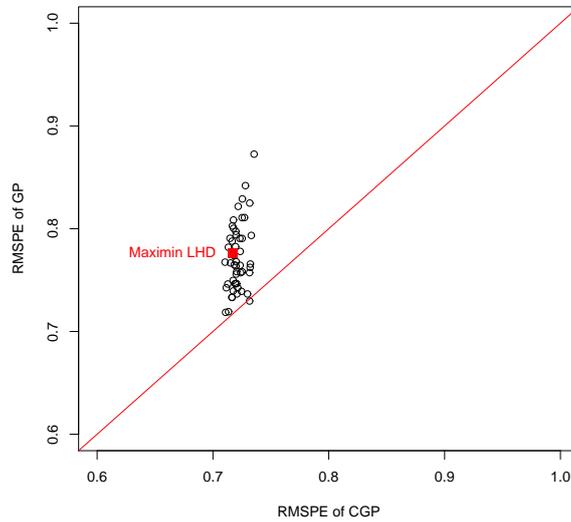


Figure 5: RMSPEs of GP and CGP models for the Michalewicz’s function. Points falling above the diagonal line indicating larger prediction errors for the GP model.

Example 3. Qian et al. (2006) described a computer simulation of a heat exchanger for electronic cooling applications. The device under study consists of linear cellular materials and is used for dissipating the heat generated by some sources such as a microprocessor. The response of interest is the total rate of steady state heat transfer of the device, which depends on the mass flow rate of entry air $\dot{m} \in (0.00055, 0.001)$, the temperature of entry air $T_{in} \in (270, 303.15)$, the solid material thermal conductivity $k \in (330, 400)$ and the

temperature of the heat source $T_{wall} \in (202.4, 360)$. The device is assumed to have fixed overall width (W), depth (D), and height (H) of 9, 25, and 17.4 millimeters, respectively. In Qian et al. (2006), the study involved two types of simulators: an expensive finite element simulator and a relatively cheaper finite difference simulator. Since the latter type of simulation was systematically conducted in the design space while the previous one only available at limited locations, here we only focus on using the finite difference simulation results to compare the prediction accuracy of several different models. Because the four input variables are in very different scales, all of them are standardized into the (0,1) region before analysis.

Qian et al. (2006) used a 64-run orthogonal array-based Latin Hypercube design for running the finite difference simulations with an extra 14-run test data set for assessing the predictions from surrogate model. If no prior information is available for the function and an ordinary kriging with Gaussian correlation function is directly fitted, the maximum likelihood estimates for its correlation parameters are (0.22, 4.37, 0.14, 7.24), which yield a RMSPE of 5.15. However, for this particular problem, the physical domain knowledge indicates that a linear component is very likely to exist between the response and factors. As a result, Qian et al. (2006) included the linear trend into the model and fitted a universal kriging to the data. Their results showed that the linear effects for T_{in} and T_{wall} are significant but for the other two variables are almost negligible. By including these two linear effects into the global trend, the RMSPE can be successfully reduced to only 2.588. Now we fit a CGP model to the data for comparison. Based on the maximum likelihood method in Section 4, we can estimate the unknown parameters as $\hat{\boldsymbol{\theta}}=(0.008,0.3,0.01,11.74)$, $\hat{\boldsymbol{\alpha}}=(11.81,12.17,11.94,23.48)$, $\hat{\lambda}=0.019$ and $\hat{b}=1$. The RMSPE for this new predictor is 2.24, which is much better than the ordinary kriging and even smaller than the previous improved result from universal kriging. Note that in the global trend of this new predictor, the two correlation parameters $\hat{\theta}_2$ and $\hat{\theta}_4$ (for T_{in} and T_{wall}) are remarkably larger than the others, which perfectly coincides with the two significant linear trends in universal kriging. This demonstrates the effectiveness of CGP model for capturing the global trend. In most common situations where no functional relationship in the global trend can be known in advance, the ability to automatically estimate the trend

and the variance is a great advantage for the new predictor over the other methods.

7 Conclusions

In this article, we present an intuitive approach for approximating complex surfaces that are not second-order stationary. The new predictor intrinsically incorporates a global trend and a flexible variance model, and all of its parameters can be estimated in a single stage. Compared with many existing methods, the new model enjoys several advantages such as numerical stability, improved prediction accuracy and flexible prediction intervals. R codes for fitting the CGP model can be obtained from the authors' website.

For modeling the non-stationarity in variance, one reviewer draws our attention to a related idea called *scaling* in the geostatistical literature (Banerjee, Charlin, and Gelfand 2003). The scaling approach is given in the form $Y(\mathbf{x}) = \sigma(\mathbf{x})Z(\mathbf{x})$, where $Z(\mathbf{x})$ denotes a stationary process and $\sigma^2(\mathbf{x})$ is a variance function that needs to be specified. By choosing $\sigma^2(\mathbf{x})$ as the exponent of another Gaussian process, Huang, Wang, Breidt and Davis (2011) proposed a *stochastic heteroscedastic process* (SHP) model $y(\mathbf{x}) = \mathbf{g}^\top(\mathbf{x})\boldsymbol{\beta} + \sigma \exp(\tau\alpha(\mathbf{x})/2)Z(\mathbf{x})$ for low-dimensional environmental applications, where $\alpha(\mathbf{x})$ is defined to be another stationary Gaussian process that is independent of $Z(\mathbf{x})$. Although this SHP model does not have a flexible global trend, its variance model is more sophisticated than our CGP model. This additional flexibility in variance, however, comes with the expenses of a very difficult and complicated estimation procedure. Since the likelihood function of the SHP model has no closed-form expression, simulation-based approximations have to be applied for the likelihood value during each step of its optimization. Obviously, this can be computationally very challenging (or even infeasible) when the dimension of unknown parameters is high, which limits its application in computer experiments.

Recently, we also noticed an interesting work from Haaland and Qian (2011), which uses the sum of multiple GPs to emulate outputs from large scale computer experiments. However, the purposes of their work is different from ours. The aim of Haaland and Qian (2011) is mainly to control the numerical error in computing interpolators based on huge

amount of data. Their multiple GP models are fitted sequentially and each of them is only based on a subset of data points. On the contrary, our method is developed to improve the precision in modeling expensive simulation results that are not second-order stationary. Both our global and local GPs are fitted based on the entire data set and all parameters in our model are also estimated in a single stage.

For p input factors, the proposed CGP model involves $p + 3$ unknown parameters, which is computationally slightly more expensive to fit than the ordinary kriging. This is the price we need to pay for incorporating the extra flexibility in modeling the global trend and the change of variance. We want to note that although the number of parameters in ordinary kriging can also be extended from p to $2p$ by generalizing its Gaussian correlation function to the *power exponential correlation function* $r(\mathbf{h}|\boldsymbol{\theta}, \mathbf{w}) = \exp(-\sum_{j=1}^p \theta_j |h_j|^{w_j})$ or even a *Matern correlation function*, this extension alone cannot solve the problems discussed in this paper, since the resulting predictor still remains second-order stationary.

Appendix: Proof of Theorem 1

Since both the single-stage predictor (14) and the sequential predictor (17) contain the same global trend $\hat{y}_{global}(\mathbf{x})$ as in (15), we only need to prove $\hat{y}_{local}(\mathbf{x}) = v^{1/2}(\mathbf{x})\hat{y}_{adj}(\mathbf{x})$.

$$\begin{aligned}
v^{1/2}(\mathbf{x})\hat{y}_{adj}(\mathbf{x}) &= v^{1/2}(\mathbf{x})\mathbf{l}^\top(\mathbf{x})\mathbf{L}^{-1}\mathbf{s}^* \\
&= v^{1/2}(\mathbf{x})\mathbf{l}^\top(\mathbf{x})\mathbf{L}^{-1}\boldsymbol{\Sigma}^{-1/2}[\mathbf{y} - \hat{\mu}\mathbf{1} - \mathbf{G}(\mathbf{G} + \lambda\boldsymbol{\Sigma}^{1/2}\mathbf{L}\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1})] \\
&= v^{1/2}(\mathbf{x})\mathbf{l}^\top(\mathbf{x})\mathbf{L}^{-1}\boldsymbol{\Sigma}^{-1/2}[\mathbf{I} - \mathbf{G}(\mathbf{G} + \lambda\boldsymbol{\Sigma}^{1/2}\mathbf{L}\boldsymbol{\Sigma}^{1/2})^{-1}](\mathbf{y} - \hat{\mu}\mathbf{1}) \\
&= \lambda v^{1/2}(\mathbf{x})\mathbf{l}^\top(\mathbf{x})\boldsymbol{\Sigma}^{1/2}(\lambda\boldsymbol{\Sigma}^{1/2}\mathbf{L}\boldsymbol{\Sigma}^{1/2})^{-1}[\mathbf{I} - \mathbf{G}(\mathbf{G} + \lambda\boldsymbol{\Sigma}^{1/2}\mathbf{L}\boldsymbol{\Sigma}^{1/2})^{-1}](\mathbf{y} - \hat{\mu}\mathbf{1}) \\
&\stackrel{(*)}{=} \lambda v^{1/2}(\mathbf{x})\mathbf{l}^\top(\mathbf{x})\boldsymbol{\Sigma}^{1/2}(\mathbf{G} + \lambda\boldsymbol{\Sigma}^{1/2}\mathbf{L}\boldsymbol{\Sigma}^{1/2})^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}) \\
&= \hat{y}_{local}(\mathbf{x}),
\end{aligned}$$

where the equality $\stackrel{(*)}{=}$ holds because $(\lambda\boldsymbol{\Sigma}^{1/2}\mathbf{L}\boldsymbol{\Sigma}^{1/2})^{-1}[\mathbf{I} - \mathbf{G}(\mathbf{G} + \lambda\boldsymbol{\Sigma}^{1/2}\mathbf{L}\boldsymbol{\Sigma}^{1/2})^{-1}](\mathbf{G} + \lambda\boldsymbol{\Sigma}^{1/2}\mathbf{L}\boldsymbol{\Sigma}^{1/2}) = \mathbf{I}$.

References

- Ababou, R., Bagtzoglou, A. C., and Wood, E. F. (1994), “On the Condition Number of Covariance Matrices in Kriging, Estimation, and Simulation of Random Fields,” *Mathematical Geology*, 26, 99-133.
- Ankenman, B., Nelson, B. L., and Staum, J. (2010), “Stochastic Kriging for Simulation Metamodeling,” *Operations Research*, 58, 371-382.
- Anderes, E. B., and Stein, M. L. (2008), “Estimating Deformations of Isotropic Gaussian Random Fields on the Plane,” *Annals of Statistics*, 36, 719-741.
- Ba, S., and Joseph, V. R. (2011), “Multi-Layer Designs for Computer Experiments,” *Journal of the American Statistical Association*, 106, 1139-1149.
- Banerjee, S., Charlin, B. P., and Gelfand, A. E. (2003), *Hierarchical Modeling and Analysis for Spatial Data*, Florida: Chapman and Hall/CRC.
- Cressie, N. A. (1991), *Statistics for Spatial Data*, New York: Wiley.
- Currin, C., Mitchell, T. J., Morris, M. D. and Ylvisaker, D. (1991), “Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments,” *Journal of the American Statistical Association*, 86, 953-963.
- Fang, K. T., Li, R., and Sudjianto, A. (2006), *Design and Modeling for Computer Experiments*, London: Chapman and Hall.
- Gneiting, T., and Raftery, A. E. (2007), “Strictly Proper Scoring Rules, Prediction, and Estimation,” *Journal of the American Statistical Association*, 102, 359-378.
- Gramacy, R. B., and Lee, H. K. H. (2008), “Bayesian Treed Gaussian Process Models with an Application to Computer Modeling,” *Journal of the American Statistical Association*, 103, 1119-1130.

- Gramacy, R. B., and Lee, H. K. H. (2011), “Cases for the Nugget in Modeling Computer Experiments,” *Statistics and Computing*, to appear.
- Higdon, D. M., Swall, J., and Kern, J. (1999), “Non-Stationary Spatial Modeling,” *Bayesian Statistics 6, Proceedings of the Sixth Valencia International Meeting*, 761-768, Oxford University Press.
- Haaland, B., and Qian, P. Z. G. (2011), “Accurate Emulators for Large-Scale Computer Experiments,” *Annals of Statistics*, 39, 2974-3002.
- Huang, W., Wang, K., Breidt, F. J., and Davis, R. A. (2011), “A Class of Stochastic Volatility Models for Environmental Applications,” *Journal of Time Series Analysis*, 32, 364-377.
- Joseph, V. R. (2006), “Limit Kriging,” *Technometrics*, 48, 458-466.
- Joseph, V. R., Hung, Y., and Sudjianto, A. (2008), “Blind Kriging: A New Method for Developing Metamodels,” *ASME Journal of Mechanical Design*, 130, 031102 (8 pages).
- Joseph, V. R., and Kang, L. (2011), “Regression-Based Inverse Distance Weighting with Applications to Computer Experiments,” *Technometrics*, 53, 254-265.
- Morris, M. D., Mitchell, T. J., and Ylvisaker, D. (1993), “Bayesian Design and Analysis of Computer Experiments: Use of Derivatives in Surface Prediction,” *Technometrics*, 35, 243-255.
- Paciorek, C., and Schervish, M. (2006), “Spatial Modelling Using a New Class of Nonstationary Covariance Functions,” *Environmetrics*, 17, 483-506.
- Peng, C. Y., and Wu, C. F. J. (2012), “Regularized Kriging,” submitted.
- Qian, P. Z. G., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Wu, C. F. J. (2006), “Building Surrogate Models with Detailed and Approximate Simulations,” *ASME Journal of Mechanical Design*, 128, 668-677.

- Ranjan, P., Haynes, R., and Karsten, R. (2012), “Gaussian Process Models and Interpolators for Deterministic Computer Simulators,” *Technometrics*, to appear.
- Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), “Design and Analysis of Computer Experiments,” *Statistical Science*, 4, 409-423.
- Sampson, P. D., and Guttorp, P. (1992), “Nonparametric Estimation of Nonstationary Spatial Covariance Structure,” *Journal of the American Statistical Association*, 87, 108-119.
- Santner, T. J., Williams, B. J., and Notz, W. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer.
- Schmidt, A. M., and O’Hagan, A. (2003), “Bayesian Inference for Nonstationary Spatial Covariance Structure via Spatial Deformations,” *Journal of the Royal Statistical Society, Series B*, 65, 745-758.
- Wackernagel, H. (2003), *Multivariate geostatistics* (3rd ed.), New York: Springer-Verlag.
- Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992), “Screening, Predicting, and Computer Experiments,” *Technometrics*, 34, 15-25.
- Xiong, Y., Chen, W., Apley, D. W., and Ding, X. (2007), “A Non-Stationary Covariance-Based Kriging Method for Metamodelling in Engineering Design,” *International Journal for Numerical Methods in Engineering*, 71, 733-756.
- Yamamoto, J. K. (2000), “An Alternative Measure of the Reliability of Ordinary Kriging Estimates,” *Mathematical Geology*, 32, 430-439.