

Heteroskedastic Spatial Models with Applications in Computer Experiments

Ke Wang, Wenying Huang, F. Jay Breidt

Colorado State University

Richard A. Davis

Columbia University

July 16, 2008

Abstract

We consider modeling a deterministic computer response as a realization from a stochastic heteroskedastic process (SHP), which incorporates a spatially-correlated volatility process into the traditional spatially-correlated Gaussian process (GP) model. Unconditionally, the SHP is a stationary non-Gaussian process, with stationary GP as a special case. Conditional on a latent process, the SHP is a non-stationary GP. The sample paths of this process offer more modeling flexibility than those produced by a traditional GP, and can better reflect prediction uncertainty. GP prediction error variances depend only on the locations of inputs, while SHP can reflect local inhomogeneities in a response surface through prediction error variances that depend on both input locations and output responses. We use maximum likelihood for inference, which is complicated by the high dimensionality of the latent process. Accordingly, we develop an importance sampling method for likelihood computation and use a low-rank kriging approximation to reconstruct the latent process. Responses at unobserved locations can be predicted using empirical best predictors or by empirical best linear unbiased predictors. Prediction error variances are also obtained. In examples with simulated and real computer experiment data, the SHP model is superior to traditional GP models.

Keywords: Adaptive sampling; Gaussian process; Stochastic heteroskedastic process (SHP); Latent process; Spatial covariance function; Uncertainty analysis

1 Introduction

Running complex computer simulations has been an influential and vital tool in many scientific areas for exploring complicated physical phenomena. Some good reviews of modeling, design and analysis of computer experiments can be found in Santer et al. (2003), Chen et al. (2003) and Fang et al. (2006). We briefly review some key ideas here. In a typical computer experiment, a high-dimensional vector $\mathbf{x} \in \mathbb{R}^d$ is used as input to a computer code, yielding an output $y(\mathbf{x})$. The output y is deterministic; i.e., running the code with the same inputs \mathbf{x} necessarily gives the same outputs. Because most codes are expensive to execute, one of the major goals of computer experiments is to seek an approximation model (metamodel) that gives outputs close to those produced by the true code but is faster to run. A statistical approach to the problem is to model the response $y(\mathbf{x})$ as a realization from a stochastic process and to construct a predictor appropriate for that process. For example, a stationary Gaussian process (GP) leads to kriging, or empirical best linear unbiased prediction (BLUP), which is a popular metamodeling technique in computer experiments (Sacks et al. (1989)). A closely-related approach is Bayesian prediction of the deterministic function under a GP model, which has also been studied extensively during the past several years (Currin et al. (1991)).

The stationary GP is popular in computer experiments because it is straightforward to fit and can produce prediction intervals. The stationarity assumption can be a severe restriction, however, especially for functions whose smoothness varies considerably over the input space. In computer experiments, some outputs do have inhomogeneous features. For example, the subsonic flow is quite different from supersonic flow in the application of computational fluid dynamics (Gramacy et al. (2004)). Using a stationary GP will oversmooth in some regions while undersmoothing in others.

To overcome this limitation of stationary GP, both regression and nonstationary GP techniques have been developed. Multivariate adaptive regression splines (MARS) have been used in the metamodeling of computer experiments by Jin et al. (2000) and Simpson et al. (2001a). The number of knots and locations for the splines are adaptively determined from the data to account for inhomogeneity. Artificial neural networks (ANN) are another approach for flexible modeling of the output from computer experiments (Chen and Varadarajan (1997) and Simpson et al. (2001b)). The MARS and ANN approaches have implicit covariance functions, and both have large numbers of coefficients with no clear interpretation.

Nonstationary GP models have been widely studied in the fields of statistics and geostatistics

(Higdon et al. (1999), Fuentes and Smith (2001), Gelfand et al. (2003), Sampson and Guttorp (1992) and Nychka et al. (2002)). Most of these nonstationary models work well in low-dimensional ($\mathbb{R}^2, \mathbb{R}^3$) physical experiments. Xiong et al. (2007) incorporates a nonstationary covariance function in modeling high-dimensional computer experiments. The nonstationary covariance function is formulated through a non-linear map with sparse parameterization, but the total number of model parameters may still be large in complicated cases. Gramacy et al. (2004) developed a tree-based Gaussian method to model nonstationarity in a response surface, by fitting individual GP models within subregions. The computational cost is reduced by this method, but discontinuities across subregions cannot be avoided.

In order to capture inhomogeneous features in computer experiment data, we adapt the idea of heteroskedasticity modeling from time series stochastic volatility (SV) models; see Shephard (1996) for a review. In a typical SV model, $y_t = \epsilon_t \exp(h_t/2)$, where $\{\epsilon_t\}$ is independent and identically distributed (iid) $N(0,1)$, and $\{h_t\}$ is a correlated, latent process. Taylor (1986) assumes an AR(1) model for the latent process. The SV model is parsimonious and interpretable. The conditional variance of y_t given h_t is e^{h_t} , so that the correlated latent process h_t captures volatility clustering effects. In his study of heteroskedasticity for spatial lattice data, Yan (2007) adapts the SV idea by introducing a spatial stochastic volatility (SSV) component into the widely-used conditional autoregressive (CAR) model.

Palacios and Steele (2006) extended the SV idea to a spatial context in which time t is replaced by a continuous (non-lattice) multi-dimensional index \boldsymbol{x} , and where the uncorrelated noise $\{\epsilon_t\}$ is replaced by a random field Z . In a standard GP model for spatial data, the random field Z is used to capture the small-scale variation in the deterministic computer experiment outputs. Palacios and Steele (2006) use a spatial stochastic volatility component in a GP model to capture non-Gaussian features in geospatial data, particularly outliers. We extend the model parameterization used by Palacios and Steele (2006) and model a deterministic computer response as a realization from a stochastic heteroskedastic process (SHP). Unconditionally, the SHP is a stationary non-Gaussian process, with stationary GP as a special case. Conditional on a latent process, the SHP is a non-stationary GP. The sample paths of this process offer more modeling flexibility than those produced by a traditional GP, and can better reflect prediction uncertainty. GP prediction error variances depend only on the locations of inputs, while SHP can reflect local inhomogeneities in a response surface through prediction error variances that depend on both input locations and output responses. We use maximum likelihood for inference, which is complicated by the high dimensionality of the latent process. Accordingly, we develop an importance sampling method for likelihood

computation and use low-rank kriging approximation to reconstruct the latent process. Responses at unobserved locations can be predicted using empirical best predictors (EBP) or by empirical best linear unbiased predictors (EBLUP). For more details on the properties of SHP and the implementation of the estimation methods, refer to Huang et al. (2008). This paper is organized as follows. In Section 2, we review GP models and introduce the SHP model and its properties. In Section 3, we outline the importance sampling method for likelihood calculations with the SHP model, and introduce a low-rank kriging approximation for the latent process. We also discuss the empirical BLUP and empirical BP for prediction. In Section 4, we illustrate the effectiveness of the SHP model and the proposed estimation method through a 2-d mathematical test function and through 4-d and 7-d computer experiment examples. The SHP is shown to be a flexible metamodeling approach in computer experiments that improves prediction accuracy and provides better quantification of prediction uncertainty than standard alternatives. Specifically, SHP-guided sampling of new inputs is more efficient at reducing prediction uncertainty than alternatives. Conclusions and further discussion are given in Section 5.

2 Spatial models

2.1 Gaussian process model

The standard GP model for computer simulation experiments (Sacks et al. (1989)) has the form

$$y(\mathbf{x}) = \mathbf{g}(\mathbf{x})^T \boldsymbol{\beta} + \sigma Z(\mathbf{x}), \quad \sigma > 0, \quad (1)$$

where $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \dots, g_p(\mathbf{x})]^T$ are known regression functions, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a vector of unknown regression coefficients, and $Z(\mathbf{x})$ is a stationary Gaussian stochastic process with mean zero, variance one and correlation function ρ_z .

Together with σ^2 , the correlation function ρ_z characterizes the distributional properties of the spatially correlated error process $Z(\mathbf{x})$. A major issue in modeling Z is the selection of correlation function. One popular choice of the correlation function used in computer experiments is the separable Gaussian correlation function, which has the form $\rho_z(\mathbf{x}, \mathbf{x}') = \exp(-\sum_{k=1}^d \phi_k |x_k - x'_k|^2)$, where ϕ_k is the unknown correlation parameter for the k^{th} covariate and x_k, x'_k are the k^{th} components of sample points \mathbf{x}, \mathbf{x}' . In this case, the generated sample path is infinitely differentiable in mean square (Santer et al. (2003)).

Given observed data, $(\mathbf{x}_1, y(\mathbf{x}_1)), \dots, (\mathbf{x}_n, y(\mathbf{x}_n))$, we want to predict $y(\cdot)$ at an untried input \mathbf{x}_0 . Let $\mathbf{y} = [y(\mathbf{x}_1), \dots, y(\mathbf{x}_n)]^T$, $\mathbf{r}_z = [\rho_z(\mathbf{x}_0, \mathbf{x}_1), \dots, \rho_z(\mathbf{x}_0, \mathbf{x}_n)]^T$, $R_z = [\rho_z(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^n$, $G =$

$[\mathbf{g}(\mathbf{x}_1), \dots, \mathbf{g}(\mathbf{x}_n)]^T$ and $\hat{\boldsymbol{\beta}} = (G^T R_z^{-1} G)^{-1} G^T R_z^{-1} \mathbf{y}$, the generalized least squares estimator. Then the best linear unbiased predictor (BLUP) for $y(\mathbf{x}_0)$ is

$$\hat{y}(\mathbf{x}_0) = \mathbf{g}(\mathbf{x}_0)^T \hat{\boldsymbol{\beta}} + \mathbf{r}_z^T R_z^{-1} (\mathbf{y} - G^T \hat{\boldsymbol{\beta}}), \quad (2)$$

with mean square prediction error

$$\text{var}(\hat{y}(\mathbf{x}_0)) = \sigma^2 \left\{ 1 - \mathbf{r}_z^T R_z^{-1} \mathbf{r}_z + (\mathbf{g}(\mathbf{x}_0) - G^T R_z^{-1} \mathbf{r}_z)^T (G^T R_z^{-1} G)^{-1} (\mathbf{g}(\mathbf{x}_0) - G^T R_z^{-1} \mathbf{r}_z) \right\}. \quad (3)$$

The BLUP interpolates all the observed data points, which is an advantage of GP in modeling deterministic computer code outputs. The empirical best linear unbiased predictor (EBLUP) for $y(\mathbf{x}_0)$ is then

$$\hat{y}(\mathbf{x}_0) = \mathbf{g}(\mathbf{x}_0)^T \hat{\boldsymbol{\beta}} + \hat{\mathbf{r}}_z^T \hat{R}_z^{-1} (\mathbf{y} - G^T \hat{\boldsymbol{\beta}}), \quad (4)$$

where $\hat{\mathbf{r}}_z$ and \hat{R}_z are obtained by plugging in estimates, such as those obtained by maximum likelihood, for the unknown parameters ϕ_1, \dots, ϕ_d .

2.2 Stochastic heteroskedastic process model

One approach to modeling nonstationarity is through scaling (Banerjee et al. (2003)). If $Z(\mathbf{x})$ is a stationary process with mean zero, variance one, and correlation function ρ_z and $\sigma(\mathbf{x})$ is a pre-specified deterministic function, then $W(\mathbf{x}) = \sigma(\mathbf{x})Z(\mathbf{x})$ is a nonstationary, heteroskedastic process. Alternatively, if $\sigma(\mathbf{x})$ is a random process, then $W(\mathbf{x})$ is conditionally heteroskedastic, but has a probabilistic structure that is characterized by a small number of distributional parameters instead of by an infinite-dimensional function. This is the motivation for the definition of the stochastic heteroskedastic process, or SHP, given as follows:

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{g}(\mathbf{x})^T \boldsymbol{\beta} + W(\mathbf{x}), \\ W(\mathbf{x}) &= \sigma \exp\left(\frac{\tau \alpha(\mathbf{x})}{2}\right) Z(\mathbf{x}), \quad \sigma > 0, \quad \tau > 0, \end{aligned} \quad (5)$$

where $\alpha(\mathbf{x})$ and $Z(\mathbf{x})$ are two independent stationary Gaussian processes with mean zero, variance one and correlation functions ρ_α and ρ_z respectively. In the remainder of this paper, we take ρ_α and ρ_z to be isotropic correlation functions with range parameters ϕ_α^{-1} and ϕ_z^{-1} respectively. The overall trend is $\mathbf{g}(\mathbf{x})^T \boldsymbol{\beta}$ and $W(\mathbf{x})$ represents local variation. The latent process $\alpha(\mathbf{x})$ is used to model local inhomogeneities, in the sense that large, positive values of $\alpha(\cdot)$ in a neighborhood of \mathbf{x} allow for large changes of the response $y(\cdot)$ in that neighborhood. Palacios and Steele (2006) consider model (5) for geospatial modeling, with $\rho_\alpha \equiv \rho_z$.

The y process has mean $\mathbf{g}(\mathbf{x})^T \boldsymbol{\beta}$, variance $\sigma^2 \exp(\tau^2/2)$ and kurtosis $3e^{\tau^2}$. Since the kurtosis is greater than 3, the y process has tails heavier than those of a normal distribution. Using the independence of the α and Z processes, it is easy to show that W is isotropic with unconditional correlation given by

$$\rho_Y(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{4}\tau^2 + \frac{1}{4}\tau^2 \rho_\alpha(\|\mathbf{x} - \mathbf{x}'\|)\right) \rho_z(\|\mathbf{x} - \mathbf{x}'\|). \quad (6)$$

Figure 1 shows realizations from GP and SHP models, where ϕ_α and ϕ_z are Matérn correlation functions with smoothness parameter $\nu = 2.5$. The figure shows the versatility of the SHP model. The SHP can produce realizations similar to those from GP for certain model parameter values, e.g., small values of ϕ_α and τ^2 , as shown in panel (c). Furthermore, the SHP model can also produce inhomogeneous realizations for large values of ϕ_α or τ^2 (panel (d)). While the unconditional correlation functions for GP and SHP in panels (a) and (b) are nearly identical, the realizations are remarkably different. The realization from the GP has a relatively uniform degree of smoothness over the whole input domain. In contrast, the realization from the SHP model has some local inhomogeneities. The local variation in the response surface becomes more obvious with increasing ϕ_α and τ^2 . The smoothness of the SHP realization varies over different parts of the region, which is due largely to the inhomogeneity of the conditional covariance, given the latent process $\boldsymbol{\alpha}$:

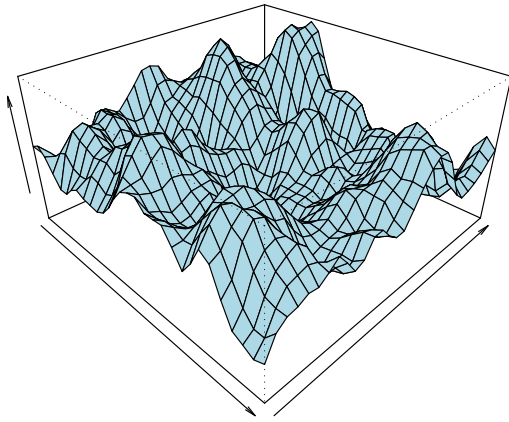
$$\gamma(\mathbf{x}, \mathbf{x}' | \boldsymbol{\alpha}) = \sigma^2 \exp\left(\frac{\tau\alpha(\mathbf{x})}{2}\right) \rho_z(\|\mathbf{x} - \mathbf{x}'\|) \exp\left(\frac{\tau\alpha(\mathbf{x}')}{2}\right). \quad (7)$$

Equations (6) and (7) indicate that W process is conditionally heteroskedastic and unconditionally stationary.

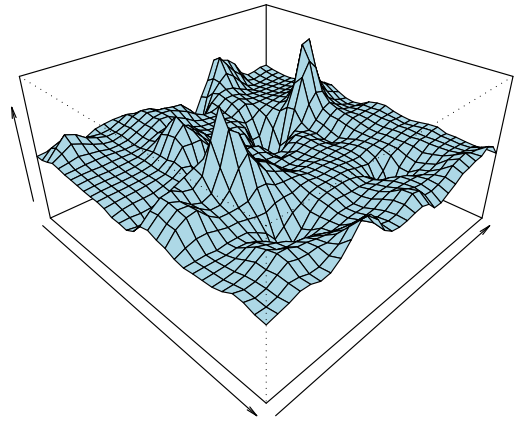
3 Estimation and prediction

3.1 Likelihood calculation

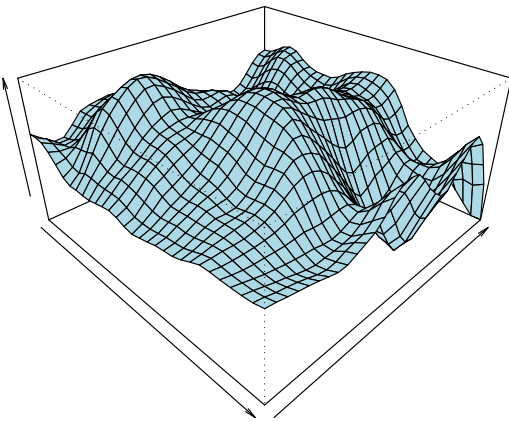
Due to the presence of the latent process α in the SHP model, there are no closed-form expressions for the likelihood function, and we consider simulation-based alternatives for its computation. We focus on importance sampling methods, which have proven successful in related time series models (Danielsson and Richard (1993), Durbin and Koopmans (1997), Davis and Rodriguez-Yam (2005)). Let $\mathbf{y} := (y_1, \dots, y_n)$ denote the vector of observations and $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_n)$ the vector of latent process values at input values $\mathbf{x}_1, \dots, \mathbf{x}_n$. Let $\boldsymbol{\psi} := (\boldsymbol{\theta}, \phi_\alpha)$ denote the model parameters where



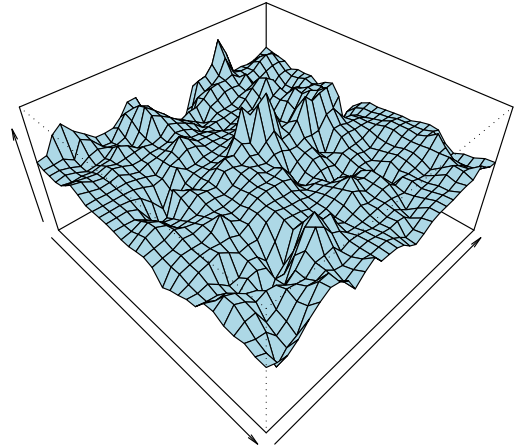
(a)



(b)



(c)



(d)

Figure 1: Simulated realizations from four processes. Panel (a): GP with $\phi = 6.4$. Panel (b): SHP with $\phi_\alpha = 8.5$ and $\phi_z = 4.4$. Panel (c): SHP with $\phi_\alpha = 1$ and $\phi_z = 3.3$. Panel (d): SHP with $\phi_\alpha = 12$ and $\phi_z = 11$. All correlation functions are Matérn with $\nu = 2.5$. All α processes use one common noise sequence, and all Z processes use another noise sequence.

$\boldsymbol{\theta} := (\sigma^2, \tau^2, \phi_z, \boldsymbol{\beta})$. The joint density of $(\mathbf{y}, \boldsymbol{\alpha})$ for the SHP model is given by

$$\begin{aligned} p(\mathbf{y}, \boldsymbol{\alpha} | \boldsymbol{\psi}) &= p(\mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha} | \phi_\alpha) \\ &= p(\mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\theta}) |R_\alpha|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{\alpha}^T R_\alpha^{-1} \boldsymbol{\alpha}\right) (2\pi)^{-\frac{n}{2}}, \end{aligned} \quad (8)$$

where $R_\alpha = [\rho_\alpha(\|\mathbf{x}_i - \mathbf{x}_j\|)]_{i,j=1}^n$, which only depends on ϕ_α .

It follows that the likelihood of the observed data is given by the n -fold integral

$$L(\boldsymbol{\psi}; \mathbf{y}) = \int p(\mathbf{y}, \boldsymbol{\alpha} | \boldsymbol{\psi}) d\boldsymbol{\alpha} = \int p(\mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha} | \phi_\alpha) d\boldsymbol{\alpha}. \quad (9)$$

In this paper, we use importance sampling as implemented in Davis and Rodriguez-Yam (2005) (see also Durbin and Koopmans (1997)) to approximate the likelihood function given in (9). We first obtain a density $p_a(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi})$ as an approximation to $p(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi})$ and then use $p_a(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi})$ as an importance density.

The construction of the importance density involves a Taylor series expansion of $\log p(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi})$ in a neighborhood of the posterior mode of $p(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi})$. Refer to Huang et al. (2008) for details of constructing the importance density. Let $\boldsymbol{\alpha}^*$ be the mode of $p(\mathbf{y}, \boldsymbol{\alpha} | \boldsymbol{\psi})$. Then the importance density is

$$p_a(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi}) = \text{N}(\boldsymbol{\alpha}^*, (K^* + R_\alpha^{-1})^{-1}) \quad (10)$$

and

$$\begin{aligned} K^* &= \frac{\tau^2}{4\sigma^2} (B + \text{diag}\{\mathbf{c}\}), \\ B &= \text{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\} \text{diag}\{\mathbf{y} - \mathbf{g}^T \boldsymbol{\beta}\} R_z^{-1} \text{diag}\{\mathbf{y} - \mathbf{g}^T \boldsymbol{\beta}\} \text{diag}\left\{\exp\left(-\frac{\tau\boldsymbol{\alpha}}{2}\right)\right\}, \\ \mathbf{c} &= \left[\exp\left(-\frac{\tau\boldsymbol{\alpha}}{2}\right)\right]^T \text{diag}\{\mathbf{y} - \mathbf{g}^T \boldsymbol{\beta}\} R_z^{-1} \text{diag}\{\mathbf{y} - \mathbf{g}^T \boldsymbol{\beta}\} \left[\exp\left(-\frac{\tau\boldsymbol{\alpha}}{2}\right)\right]. \end{aligned} \quad (11)$$

Using $p_a(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi})$ as an importance density function to implement Monte Carlo integration, the integral in (9) can be rewritten as

$$L(\boldsymbol{\psi}; \mathbf{y}) = \int \frac{p(\mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha} | \phi_\alpha)}{p_a(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi})} p_a(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi}) d\boldsymbol{\alpha} = \text{E}_a \left[\frac{p(\mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\theta}) p(\boldsymbol{\alpha} | \phi_\alpha)}{p_a(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi})} \right]. \quad (12)$$

If $\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(N)}$ are drawn from $p_a(\boldsymbol{\alpha} | \mathbf{y}, \boldsymbol{\psi})$, then (12) can be approximated by

$$L(\boldsymbol{\psi}; \mathbf{y}) \approx \frac{1}{N} \sum_{i=1}^N \left[\frac{p(\mathbf{y} | \boldsymbol{\alpha}^{(i)}, \boldsymbol{\theta}) p(\boldsymbol{\alpha}^{(i)} | \phi_\alpha)}{p_a(\boldsymbol{\alpha}^{(i)} | \mathbf{y}, \boldsymbol{\psi})} \right]. \quad (13)$$

To compute this approximation, we first simulate one set of common random draws $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(N)}$ to be used in every evaluation of the likelihood. Then, for a given value of $\boldsymbol{\psi}$, we generate $\boldsymbol{\alpha}^{(i)}$ from

$\mathbf{u}^{(i)}$ for $i = 1, \dots, N$ and use these to evaluate (13). It is standard practice to use common random numbers in computing a likelihood with an importance sampler, since this smooths the estimated likelihood surface, reducing Monte Carlo error in the approximation and facilitating convergence of the numerical optimizer.

In order to calculate the importance density (10), we need to find the posterior mode $\boldsymbol{\alpha}^*$ by maximizing (8). Since the dimensionality of $\boldsymbol{\alpha}$ is the same as the number of observations, it can be difficult to find $\boldsymbol{\alpha}^*$ especially for a large data set. To reduce the dimensionality in this optimization procedure, we adopt the *low-rank kriging* method (for details, refer to Ruppert et al. (2003)) to approximate the latent vector $\boldsymbol{\alpha}$. Let k_1, \dots, k_J be a set of knot locations. The knots are chosen from the observed locations, and evenly spread over the input domain. The simulation study shows that 10 – 12 knots work effectively in this low-rank approximation. Then $\boldsymbol{\alpha}$ is approximated by

$$\boldsymbol{\alpha} = B\boldsymbol{\omega}, \quad \boldsymbol{\omega} \sim N(\mathbf{0}, \Omega^{-1}), \quad (14)$$

where $B \equiv [\rho_\alpha(\|\mathbf{x}_i - \mathbf{k}_j\|)]_{1 \leq i \leq n, 1 \leq j \leq J}$ and $\Omega \equiv [\rho_\alpha(\|\mathbf{k}_j - \mathbf{k}'_{j'}\|)]_{1 \leq j, j' \leq J}$. We substitute (14) into (8) and maximize the likelihood with respect to $\boldsymbol{\omega}$ to get $\hat{\boldsymbol{\omega}}$. Then $\boldsymbol{\alpha}^*$ is approximated by $B\hat{\boldsymbol{\omega}}$.

3.2 Estimation for σ^2

In simulations not reported here, we find that the likelihood tends to be flat for a wide range of large σ^2 values. Unlike the GP case, the parameter σ^2 cannot be profiled out of the SHP likelihood in (9), and it is numerically difficult to estimate σ^2 as part of the overall optimization. We propose an alternative method starting with the sample variance:

$$s^2 = \frac{1}{2n(n-1)} \sum_j \sum_k (y(\mathbf{x}_j) - y(\mathbf{x}_k))^2,$$

which has expectation

$$E(s^2) = \frac{2\sigma^2 \exp(\tau^2/2)}{2n(n-1)} \sum_j \sum_k (1 - \rho_y(\|\mathbf{x}_j - \mathbf{x}_k\|))$$

under the SHP model. Rearranging this expression, we have

$$\sigma^2 = \frac{n(n-1) \exp(-\tau^2/2) E(s^2)}{n^2 - \sum_j \sum_k \rho_y(\|\mathbf{x}_j - \mathbf{x}_k\|)}. \quad (15)$$

Rather than maximize the likelihood jointly with respect to all parameters, we fix σ^2 at the sample variance and maximize (13) only with respect to $(\tau^2, \phi_\alpha, \phi_z, \boldsymbol{\beta})$ to get estimates $(\hat{\tau}^2, \hat{\phi}_\alpha, \hat{\phi}_z, \hat{\boldsymbol{\beta}})$.

Then, by substituting $E(s^2)$ with s^2 and plugging other parameter estimates into (15), an approximately unbiased estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{n(n-1) \exp(-\hat{\tau}^2/2) s^2}{n^2 - \sum_j \sum_k \hat{\rho}_y(\|\mathbf{x}_j - \mathbf{x}_k\|)}. \quad (16)$$

3.3 Estimation of function of volatility

If $\boldsymbol{\psi}$ were known, a function $f(\cdot)$ of the latent process α at observed locations can be estimated as the conditional expectation $E[f(\alpha)|\mathbf{y}, \boldsymbol{\psi}]$, given by

$$\begin{aligned} E[f(\alpha)|\mathbf{y}, \boldsymbol{\psi}] &= \int f(\alpha) p(\alpha|\mathbf{y}, \boldsymbol{\psi}) d\alpha \\ &= \frac{\int f(\alpha) p(\mathbf{y}|\alpha, \boldsymbol{\theta}) p(\alpha|\phi_\alpha) d\alpha}{\int p(\mathbf{y}|\alpha, \boldsymbol{\theta}) p(\alpha|\phi_\alpha) d\alpha} \\ &= \frac{E_a[f(\alpha) p(\mathbf{y}|\alpha, \boldsymbol{\theta}) p(\alpha|\phi_\alpha) / p_a(\alpha|\mathbf{y}, \boldsymbol{\psi})]}{E_a[p(\mathbf{y}|\alpha, \boldsymbol{\theta}) p(\alpha|\phi_\alpha) / p_a(\alpha|\mathbf{y}, \boldsymbol{\psi})]}. \end{aligned} \quad (17)$$

We are specifically interested in estimating α and $\exp(\tau\alpha/2)$. Once the estimates of parameters $\hat{\boldsymbol{\psi}}$ are obtained, we sample $\alpha^{(1)}, \dots, \alpha^{(N)}$ from $p_a(\alpha|\mathbf{y}, \hat{\boldsymbol{\psi}})$ and approximate the conditional expectation in equation (17) by Monte Carlo integration.

Consider an unobserved location \mathbf{x}_0 . Given ϕ_α , the joint distribution of $(\alpha(\mathbf{x}_0), \alpha)$ is multivariate normal. The mean of $p(\alpha(\mathbf{x}_0)|\alpha, \phi_\alpha)$ is $\mathbf{r}_\alpha^T(\mathbf{x}_0) R_\alpha^{-1} \alpha$, where $\mathbf{r}_\alpha(\mathbf{x}_0) = [\rho_\alpha(\|\mathbf{x}_0 - \mathbf{x}_1\|), \dots, \rho_\alpha(\|\mathbf{x}_0 - \mathbf{x}_n\|)]^T$ and R_α is the $n \times n$ correlation matrix for α . It is easy to show that $E[\alpha(\mathbf{x}_0)|\mathbf{y}, \boldsymbol{\psi}] = \mathbf{r}_\alpha^T(\mathbf{x}_0) R_\alpha^{-1} E[\alpha|\mathbf{y}, \boldsymbol{\psi}]$. Plugging in estimates of α and ϕ_α , an empirical best predictor (EBP) of $\alpha(\mathbf{x}_0)$ is given by

$$\hat{\alpha}(\mathbf{x}_0) = \hat{\mathbf{r}}_\alpha^T(\mathbf{x}_0) \hat{R}_\alpha^{-1} \hat{\boldsymbol{\alpha}}. \quad (18)$$

3.4 Prediction of y process

For multivariate normal, it is well known that if all parameters are known, then the BP (best predictor in terms of minimum mean square prediction error) is the same as the BLUP (best linear unbiased predictor). The SHP model is unconditionally non-Gaussian and conditionally heteroskedastic Gaussian given the latent process α . If all parameters and the latent process are known, then the BP and BLUP are not the same under the SHP model.

Best Predictor

Given the latent process α , the joint distribution of $y(\mathbf{x}_0)$ and the observation vector \mathbf{y} is heteroskedastic Gaussian. The conditional covariances of $(y(\mathbf{x}_0), \mathbf{y})$ are given by

$$\begin{aligned}\text{Cov}(y(\mathbf{x}_0), \mathbf{y} | \boldsymbol{\psi}, \alpha(\mathbf{x}_0), \boldsymbol{\alpha}) &= \sigma^2 \exp(\tau\alpha(\mathbf{x}_0)/2) \mathbf{r}_z(\mathbf{x}_0)^T \text{diag} \{ \exp(\tau\alpha/2) \}, \\ \text{Var}(\mathbf{y} | \boldsymbol{\psi}, \boldsymbol{\alpha}) &= \sigma^2 \text{diag} \{ \exp(\tau\alpha/2) \} R_z \text{diag} \{ \exp(\tau\alpha/2) \}, \\ \text{Var}(y(\mathbf{x}_0) | \boldsymbol{\psi}, \alpha(\mathbf{x}_0)) &= \sigma^2 \exp(\tau\alpha(\mathbf{x}_0)),\end{aligned}\tag{19}$$

where $\mathbf{r}_z(\mathbf{x}_0) = [\rho_z(\|\mathbf{x}_0 - \mathbf{x}_1\|), \dots, \rho_z(\|\mathbf{x}_0 - \mathbf{x}_n\|)]^T$. It follows that the mean and variance of the conditional distribution $p(y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\psi}, \alpha(\mathbf{x}_0), \boldsymbol{\alpha})$ are

$$\begin{aligned}E(y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\psi}, \boldsymbol{\alpha}, \alpha(\mathbf{x}_0)) &= g(\mathbf{x}_0)^T \boldsymbol{\beta} \\ &\quad + \exp\left(\frac{\tau\alpha(\mathbf{x}_0)}{2}\right) \mathbf{r}_z(\mathbf{x}_0)^T R_z^{-1} \text{diag} \left\{ \exp\left(-\frac{\tau\alpha}{2}\right) \right\} (\mathbf{y} - G^T \boldsymbol{\beta}),\end{aligned}\tag{20}$$

$$\text{Var}(y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\psi}, \alpha(\mathbf{x}_0)) = \sigma^2 \exp(\tau\alpha(\mathbf{x}_0)) (1 - \mathbf{r}_z(\mathbf{x}_0)^T R_z^{-1} \mathbf{r}_z(\mathbf{x}_0)).\tag{21}$$

We then compute

$$\begin{aligned}E(y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\psi}) &= E [E(y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\psi}, \boldsymbol{\alpha}, \alpha(\mathbf{x}_0)) | \mathbf{y}, \boldsymbol{\psi}] \\ &= g(\mathbf{x}_0)^T \boldsymbol{\beta} \\ &\quad + E \left[E \left[\exp\left(\frac{\tau\alpha(\mathbf{x}_0)}{2}\right) \middle| \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\psi} \right] \mathbf{r}_z(\mathbf{x}_0)^T R_z^{-1} \text{diag} \left\{ \exp\left(-\frac{\tau\alpha}{2}\right) \right\} \middle| \mathbf{y}, \boldsymbol{\psi} \right] (\mathbf{y} - G^T \boldsymbol{\beta}).\end{aligned}\tag{22}$$

Since the joint distribution of $(\alpha(\mathbf{x}_0), \boldsymbol{\alpha})$ is normal, we get

$$E \left[\exp\left(\frac{\tau\alpha(\mathbf{x}_0)}{2}\right) \middle| \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\psi} \right] = \exp\left(\frac{\tau\mu_0}{2} + \frac{\tau^2 v_0}{8}\right),\tag{23}$$

where $\mu_0 = \mathbf{r}_\alpha(\mathbf{x}_0)^T R_\alpha^{-1} \boldsymbol{\alpha}$ and $v_0 = 1 - \mathbf{r}_\alpha(\mathbf{x}_0)^T R_\alpha^{-1} \mathbf{r}_\alpha(\mathbf{x}_0)$ are the mean and variance of $p(\alpha(\mathbf{x}_0) | \boldsymbol{\alpha}, \phi_\alpha)$.

Plugging (23) into (22), we obtain the best predictor of $y(\mathbf{x}_0)$.

For $\text{Var}(y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\psi})$, we have that

$$\text{Var}(y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\psi}) = E [\text{Var}(y(\mathbf{x}_0) | \mathbf{y}, \alpha(\mathbf{x}_0), \boldsymbol{\psi}) | \mathbf{y}, \boldsymbol{\psi}] + \text{Var} (E [y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\alpha}, \alpha(\mathbf{x}_0), \boldsymbol{\psi}] | \mathbf{y}, \boldsymbol{\psi}),\tag{24}$$

where

$$E [\text{Var}(y(\mathbf{x}_0) | \mathbf{y}, \alpha(\mathbf{x}_0), \boldsymbol{\psi}) | \mathbf{y}, \boldsymbol{\psi}] = \sigma^2 E \left[\exp\left(\tau\mu_0 + \frac{\tau^2 v_0}{2}\right) (1 - \mathbf{r}_z(\mathbf{x}_0)^T R_z^{-1} \mathbf{r}_z(\mathbf{x}_0)) \middle| \mathbf{y}, \boldsymbol{\psi} \right]\tag{25}$$

and

$$\begin{aligned}
& \text{Var} \left(E[y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\alpha}, \alpha(\mathbf{x}_0), \boldsymbol{\psi}] \mid \mathbf{y}, \boldsymbol{\psi} \right) \\
&= E \left[\exp \left(\tau \mu_0 + \frac{\tau^2 v_0}{2} \right) r_z(\mathbf{x}_0)^T R_z^{-1} \text{diag} \{ \exp(-\tau \boldsymbol{\alpha} / 2) \} (\mathbf{y} - G^T \boldsymbol{\beta})(\mathbf{y} - G^T \boldsymbol{\beta})^T \right. \\
&\quad \left. \times \text{diag} \{ \exp(-\tau \boldsymbol{\alpha} / 2) \} R_z^{-1} r_z(\mathbf{x}_0) \mid \mathbf{y}, \boldsymbol{\psi} \right] - (E[y(\mathbf{x}_0) | \mathbf{y}, \boldsymbol{\psi}] - g(\mathbf{x}_0)^T \boldsymbol{\beta})^2. \tag{26}
\end{aligned}$$

Since (22) and (24) are functions of $\boldsymbol{\alpha}$, we can obtain the *empirical best predictor* (EBP) and its estimated prediction error variance by plugging in estimated model parameters and evaluating the posterior means of the $\boldsymbol{\alpha}$ functions by Monte Carlo approximation through (17).

Best Linear Unbiased Predictor

An alternative to the BP is the BLUP given unconditional distribution. The advantage of the BLUP is that it is easy to compute from equation (2). Of course, the BLUP would coincide with the best predictor if the joint distribution were Gaussian with unconditional covariance given by (6). For more details, refer to Huang et al. (2008).

4 Empirical Results

In this section, we use a two-dimensional test function and four-dimensional and seven-dimensional computer experiment examples to illustrate the prediction accuracy of the SHP model compared with the Gaussian process (GP) model. Out-of-sample root mean square prediction error (RMSE) is used as the criterion to assess model performance.

4.1 Two-dimensional test function

We consider $f(\mathbf{x}) = 10x_1 \exp(-x_1^2 - x_2^2)$, a rescaled version ($\times 10$) of the function used in Gramacy et al. (2004) to evaluate the treed GP fitting procedure. The function is evaluated on a uniform 21×21 grid of points on $[-2, 6] \times [-2, 6]$. The surface is plotted in Figure 2(a).

To compare the model accuracy of SHP with GP, a training data set of size 20 was chosen from the grid points. We used Latin hypercube sampling (LHS) to place 12 points in the quadrant $[-2, 2] \times [-2, 2]$ and 8 points in other areas. The sampling was implemented through the R package `tgpp` (Gramacy (2007)). After fitting the 20 points with a GP model and a SHP model (using isotropic Gaussian covariance functions for the GP and for α and Z in the SHP), we predicted the other 421 grid points and computed the RMSE. This process of sampling, fitting, and predicting

was repeated 100 times. Summary statistics for the 100 RMSEs for each method are given in the first row of Table 1. The SHP model with BP given by (20) has smaller RMSE in 75 of the 100 trials, and its average RMSE is about 19% smaller than that of GP. We give an example of fitted surfaces in Figure 2(b) and (c). The plots show that the SHP model is able to catch peaks better than the GP model. In this example, the latent process α does a good job of capturing the inhomogeneous features of the function.

The GP model is popular for metamodeling in computer experiments not only because it can fit complex functional forms, but also because it provides a measure of prediction uncertainty given by the prediction error variance. Adaptive sampling can then be conducted by selecting new inputs with probabilities proportional to their prediction error variances. The SHP model can also be used for adaptive sampling, using the empirical version of its prediction error variance (24). We compare the effectiveness of these prediction uncertainty measures in terms of out-of-sample prediction RMSE in adaptive sampling.

We fit each of the 100 initial tgp data sets of size 20 with the SHP model and quantify the prediction error variances. Another 20 points are adaptively selected from the grid with probabilities proportional to SHP model prediction error variances. This sampling without replacement is implemented by the R function `sample` (Team (2005)). Once we update the sample size to 40, we fit the 40 points with SHP and GP models, compute the RMSE at the remaining 401 grid points for each model, and select an additional 20 points with probabilities proportional to the updated SHP prediction error variances. We fit the 60 points with SHP and GP again and compute their RMSE at the remaining 381 grid points. These results are displayed in rows 3 and 5 of Table 1. The whole process was repeated using GP model prediction error variances in increasing the sample from 20 to 40 and then to 60. These results are displayed in rows 2 and 4 of Table 1. The table shows that regardless of the way the points are adaptively sampled, the SHP dominates the GP in every case by producing smaller average RMSEs. Further, the SHP does an excellent job of guiding the selection of new points in adaptive sampling. SHP adaptive sampling produces smaller average RMSEs regardless of whether predictions are computed from GP or SHP. Ranking the combinations of adaptive sampling strategy and prediction method by average RMSE (at sample size 40 or 60), the best is SHP sampling with SHP prediction, followed by SHP sampling with GP prediction, and GP sampling with SHP prediction. The worst combination is GP sampling with GP prediction.

The plots in Figure 3 and Figure 4 illustrate how GP and SHP, respectively, generate adaptive samples. Figure 3(a) is the image plot of true absolute errors $|\mathbf{y} - \hat{\mathbf{y}}|$ for the GP model based on

Table 1: Summary statistics for 100 replicates of out-of-sample RMSEs for GP and SHP(EBP), with different sample sizes and sampling strategies. The final column indicates the percentage of GP/SHP RMSE ratios that are greater than 1 (favoring SHP) out of 100 replicates.

sampling strategy	sample size	average GP RMSE	average SHP RMSE	percent > 1
20 tgp	20	0.513	0.418	75
20 tgp +20 GP	40	0.371	0.221	83
20 tgp +20 SHP	40	0.165	0.114	66
20 tgp +20 GP +20 GP	60	0.274	0.172	80
20 tgp +20 SHP +20 SHP	60	0.075	0.060	57

one set of 20 inputs, and Figure 3(b) is the image plot of the GP prediction error variances, which are fairly uniform away from initial sample locations. Accordingly, GP adaptive sampling selects 20 new inputs in a fairly uniform way across the previously unsampled part of the input space. This pattern is repeated in the second row of plots, Figures 3(c) and (d), as the sample is extended from 40 to 60 via GP adaptive sampling.

Similarly, Figure 4(a) is the image plot of true absolute errors $|\mathbf{y} - \hat{\mathbf{y}}|$ for the SHP model based on one set of 20 inputs, and Figure 4(b) is the image plot of the SHP prediction error variances. In contrast with Figure 3(b), these prediction error variances are far from uniform away from initial sample locations, and instead have hot spots of high uncertainty. Accordingly, SHP adaptive sampling selects 20 new inputs intensively in the hot spots. This pattern is repeated in the second row of plots, Figures 4(c) and (d), as the sample is extended from 40 to 60 via SHP adaptive sampling. In this example, with clear inhomogeneity in the surface, the SHP not only produces better predictors but also provides a much better adaptive sampling scheme.

4.2 Four-dimensional computer simulation experiment

To illustrate the potential of SHP in metamodeling, we consider an example of a computer simulation experiment given in Qian et al. (2006). The data consist of the outputs from computer simulations for a heat exchanger used in electronic cooling applications. The response y of interest is the steady heat transfer rate depending on four inputs: the mass flow rate of entry air \dot{m} , the temperature of entry air T_{in} , the temperature of heat source T_{wall} and solid material thermal

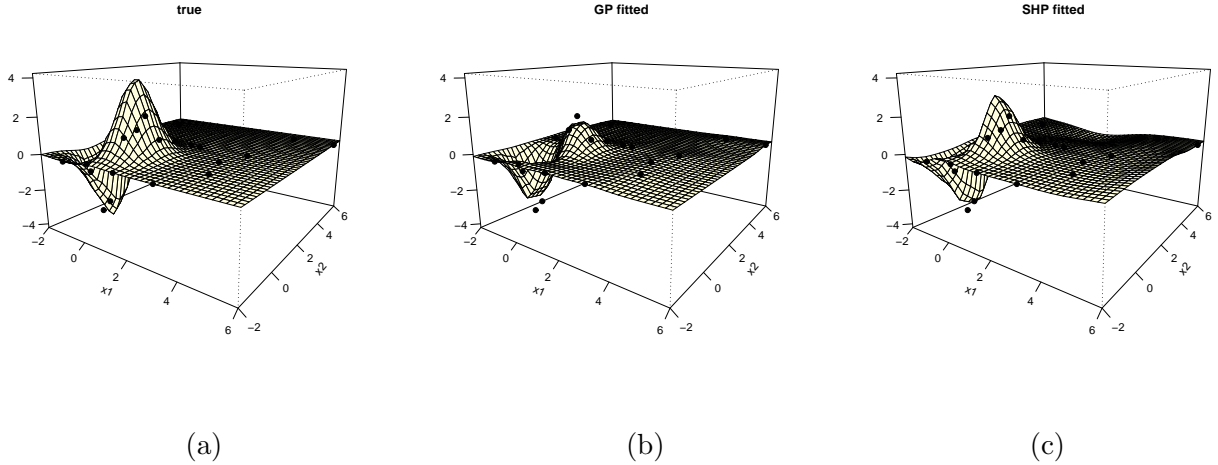


Figure 2: (a) True response surface. (b) GP-fitted surface based on 20 inputs. (c) SHP-fitted surface based on 20 inputs.

conductivity k . There are two types of simulations used in their study: an approximate simulation (AS) and a detailed simulation (DS). The training data consist of AS data for 64 input points and DS data for 22 out of these 64 input points. The testing data consist of 14 set-aside input points with both AS and DS data. Following their notation, \mathbf{y}_a represents the AS outputs and \mathbf{y}_d represents the DS outputs.

To explore the relationship between the design factors and heat transfer rate, Qian et al. (2006) proposed a two-step approach to build a surrogate model that can produce predictions close to the DS data. The first step uses the 64 AS training data to build a “base surrogate model” given in (27), and the second step adjusts the fitted model in step one with the 22 DS training data to create the “final surrogate model” given in (28):

$$y_a(\mathbf{x}) = \beta_{a0} + \sum_{h=1}^d \beta_{ah} x_h + \epsilon_a(\mathbf{x}), \quad (27)$$

$$y_d(\mathbf{x}) = \theta(\mathbf{x}) y_a(\mathbf{x}) + \delta(\mathbf{x}). \quad (28)$$

Here, $\epsilon_a(\mathbf{x})$ is a stationary GP with zero mean and separable Gaussian covariance function, $\delta(\mathbf{x})$ is a stationary GP with unknown constant mean and separable Gaussian covariance function, and $\theta(\mathbf{x}) = \theta_0 + \sum_{j=1}^d \theta_j x_j$. For more modeling and engineering details, see Qian et al. (2006).

We use the same 64 AS training data to build base surrogate models with anisotropic GP (reproducing Qian et al.’s separable model), isotropic GP, and SHP. We compare the three approaches using leave-one-out cross-validation. The cross validation score for the SHP model with EBP is

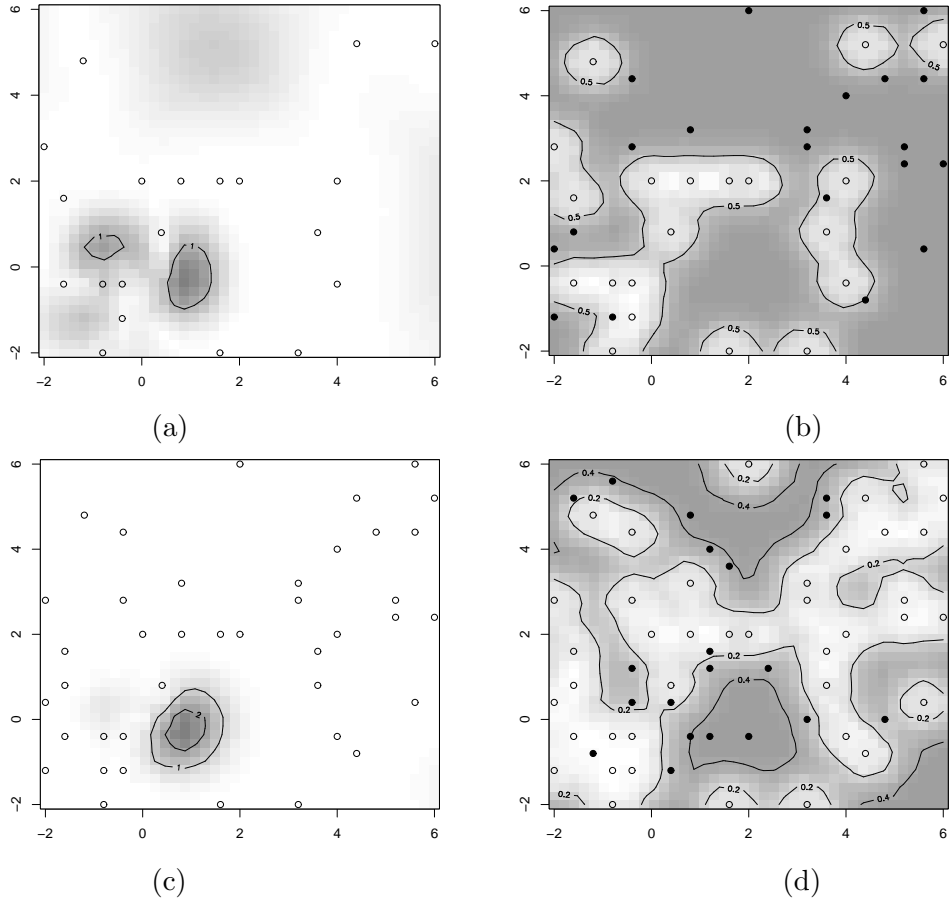


Figure 3: (a) Image plot of absolute error $|\mathbf{y} - \hat{\mathbf{y}}|$ from GP predictions using 20 initial sample points given by the open circles. (b) Image plot of GP prediction error variance based on the 20 initial locations (open circles). Solid dots are the 20 locations adaptively sampled using GP. (c) Image plot of absolute error from GP predictions based on 40 points in (b). (d) Image plot of GP prediction error variance based on the 40 input locations (open circles) in (c). Solid dots are the next 20 locations adaptively sampled using GP.

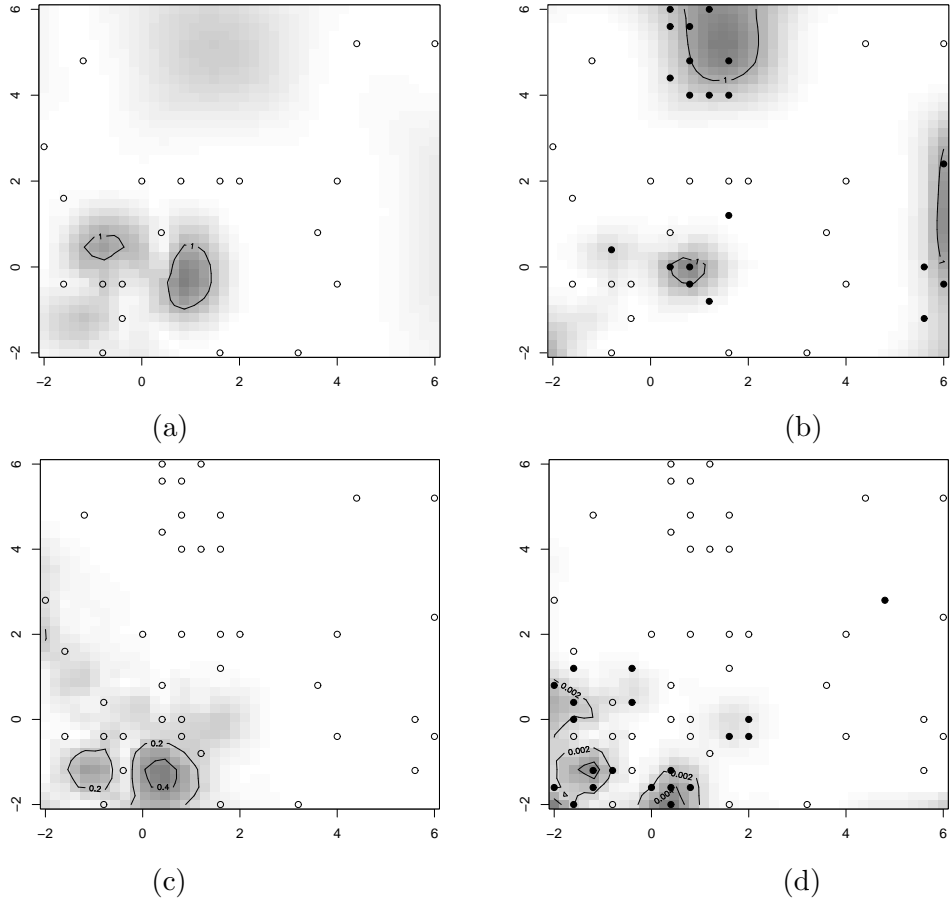


Figure 4: (a) Image plot of absolute error $|\mathbf{y} - \hat{\mathbf{y}}|$ from the SHP model using 20 initial sample points given by the open circles. (b) Image plot of SHP prediction error variance based on the 20 initial locations (open circles). Solid dots are the 20 locations adaptively sampled using SHP. (c) Image plot of absolute error from SHP model based on 40 points in (b). (d) Image plot of SHP prediction error variance based on the 40 input locations (open circles) in (c). Solid dots are the next 20 locations adaptively sampled using SHP.

0.311, versus 0.509 for the isotropic GP model and 0.642 for the separable GP model used by Qian et al. (2006). The poorer performance of the separable model in this case is due to the three additional parameters in the separable GP, resulting in increased variation in estimation and prediction with this small data set. It is worth noting that the cross validation score for SHP model with EBLUP is 0.427, which is also better than either GP model.

We now turn to the modeling of the DS data. Qian et al. (2006) use the 22 input points with both AS and DS training data to fit model (28), with δ as a GP with separable Gaussian covariance function. For \mathbf{x} in the set of 14 set-aside locations, they predict $\delta(\mathbf{x})$ using EBLUPs from this fitted model, predict $y_a(\mathbf{x})$ from the separable GP fitted to the 64 AS training data, and predict $y_m(\mathbf{x})$ by plugging their estimates and predictions into (28).

To facilitate comparisons, we use Qian et al.’s final surrogate model for DS predictions, including their estimate of $\theta(\cdot)$ and their predicted values of $\delta(\mathbf{x})$. The only difference in our DS predictions is that we plug in SHP-predicted values of $y_a(\mathbf{x})$ instead of GP-predicted values.

The proposed SHP model provides significant improvement in terms of prediction accuracy for both the 14 AS and the 14 DS set-aside values. For AS, the RMSE based on SHP is 2.073 versus 2.588 for Qian et al. For DS, the RMSE based on SHP-predicted inputs is 3.133, versus 3.795 for Qian et al.

Using the SHP model to fit AS data improves the prediction accuracy for DS data. This example illustrates the potential of SHP modeling in application of multi-level computer experiments and model validation. The SHP model can be used not only to model lower-level outputs but also to model the bias term among different level outputs or the bias between computer code output and physical data in model validation. We did not use SHP to model the bias term in this example because there are only 22 DS data in a 4-dimensional space. For a very small data set evenly distributed over the input domain, the GP model can do a comparable job to the SHP model.

4.3 SIR model: Seven-dimensional computer simulation experiment

An SIR (Susceptible-Infected-Recovered) model describes the time dynamics of a contagious disease through a system of ordinary differential equations, with one equation for susceptible individuals $S(t)$, one for infected $I(t)$, and one for recovered $R(t)$. We use one example from Estep and Neckels (2006) to investigate the behavior of an SIR model that allows for birth, death due to natural causes, death due to disease, and the possibility that the offspring of the resistant class may inherit

the resistance. This SIR model is described by

$$\begin{cases} \dot{S} = r_n(1 - \frac{N}{k})(S + I + (1 - p_R)R) - d_n S - r_I SI, \\ \dot{I} = r_I SI - (d_n + d_I)I - a_R I, \\ \dot{R} = p_R r_n(1 - \frac{N}{k})R - d_n R + a_R I, \\ S(0) = S_0, I(0) = I_0, R(0) = R_0, \end{cases}$$

where $\dot{S}, \dot{I}, \dot{R}$ denote time derivatives, $N = S(t) + I(t) + R(t)$ is the population size, and $\mathbf{x} = (a_R, r_n, k, p_R, d_n, r_I, d_I)^T$ are the input parameters recovery rate, natural growth rate, carrying capacity, probability of inheriting resistance, natural death rate, contraction rate, and death rate from disease. The domains for each input parameter are specified in Table 7.3.1 of Estep and Neckels (2006). Responses of interest for the SIR could be some functional of S, I, R . We considered three responses: the average number of infected individuals, the average number of susceptible individuals and the average number of resistant individuals over a time interval $[0, T]$:

$$q(\mathbf{x})_1 = \frac{1}{T} \int_0^T S(s, \mathbf{x}) ds, \quad q(\mathbf{x})_2 = \frac{1}{T} \int_0^T I(s, \mathbf{x}) ds, \quad q(\mathbf{x})_3 = \frac{1}{T} \int_0^T R(s, \mathbf{x}) ds.$$

We model the three quantities of interest individually but report only on results for q_2 with $T = 10$ as in Estep and Neckels (2006). Similar results are obtained for q_1 and q_3 and reported in Wang (2008). Note that these responses are not expensive to compute, so exact results can be obtained and compared to predictions based on models fitted to samples of inputs.

Using the generalized Green's function and a variational analysis, Estep and Neckels (2006) compute not only the quantity of interest but also the derivatives at sampled input points. This derivative information is used in Estep and Neckels (2006) to create what they refer to as the "higher-order parameter sampling" method, or HOPS, to approximate the quantity of interest at untried locations. We compare stochastic modeling using GP and SHP to the HOPS method.

We first used Latin hypercube sampling to select 70 data points at random from the input domain, which is standardized to $[0, 1]^7$. We then fit using HOPS, GP and SHP and predicted values of q_2 at 1000 points uniformly distributed in the input domain. These predictions were compared to the true values of the quantities for those 1000 points. The process of sampling, fitting, and predicting was repeated 100 times. The first two rows of Table 2 show the summary statistics for the 100 replicates of the RMSE ratios of HOPS/SHP and GP/SHP. SHP is often much better and never much worse than either of the competing methods.

It is also of interest to investigate the performance of predictors locally, for sub-domains in the parameter space. In the SIR setting, a potential sub-domain of interest might be a "good

Table 2: Summary statistics of 100 replicates of RMSE ratios for HOPS/SHP and GP/SHP for q_2 on the entire input domain and on two sub-domains.

		min	25 th	median	mean	75 th	max	percentage
global	HOPS/SHP	0.985	1.374	1.551	1.556	1.743	2.313	99
	GP/SHP	0.868	1.042	1.118	1.135	1.209	1.504	83
good	GP/SHP	0.375	0.937	1.116	1.122	1.303	1.736	67
bad	GP/SHP	0.521	1.045	1.348	1.392	1.569	4.116	79

population” with high recovery rate, high natural growth rate and high probability of inheriting resistance, but low natural death rate, low contraction rate and low death rate from disease. Another sub-domain of interest might be a “bad” population with low recovery rate, low natural growth rate and low probability of inheriting resistance, but high natural death rate, high contraction rate and high death rate from disease. Among the 1000 test data values, there are 11 in the good sub-domain and 9 in the bad sub-domain. For the 100 repeated training data sets, the ratios of RMSEs for GP/SHP for q_2 in the good and bad sub-domains are summarized in the last two rows of Table 2. In this example, the SHP predictions outperform GP not only globally but also locally.

An important problem in science and engineering is the determination of the effect of variation in input parameters on the uncertainty of output. This kind of uncertainty analysis is a major objective in Estep and Neckels (2006). In particular, Estep and Neckels (2006) use a 64-point HOPS to approximate the distribution of q_2 when the input parameters are independently distributed as uniform on their respective ranges. The exact cumulative distribution function (cdf) for the quantity of interest was approximated with a massive Monte-Carlo simulation of 30,000 points. We used LHS to select 64 points from the SIR model (input and output), fitted the 64 data points with HOPS, SHP and GP, and then predicted the responses q_2 at the 30,000 randomly selected inputs as \hat{q}_2 . Compared to the MC30000 distribution, the two-sample Kolmogorov-Smirnov (K-S) test statistics $\max_{u \in \mathbb{R}} |F_{q_2}(u) - F_{\hat{q}_2}(u)|$ for HOPS, GP, and SHP are 0.057, 0.053 and 0.044, respectively. The SHP model outperforms GP and HOPS, indicating the potential of stochastic modeling in the uncertainty analysis.

5 Conclusions

In this paper, we have investigated a new class of models for metamodeling in computer experiments. The deterministic computer response is modeled as a realization from a SHP, which has a rich class of sample paths and correlation functions. For certain parameterizations, the SHP produces Gaussian-like sample paths, but the sample paths of this process allow for local inhomogeneities, unlike those produced by a traditional GP model. Using test functions and computer experiment examples, we have shown that the SHP model is more capable of capturing the peaks and valleys of response surfaces, providing better prediction accuracy and quantification of prediction uncertainty than the GP model. Moreover, the SHP methodology provides information on local volatility in the input space, so that more data points can be placed around sensitive areas. This gives one possible approach to adaptive sampling in computer experiments.

Not surprisingly, the increased flexibility of SHP comes at the price of increased computational cost compared to GP. There is, however, a trade-off between computational cost and model accuracy. Even though the computational cost of SHP is higher than GP, this cost is still negligible compared with that of running expensive computer codes.

In this paper, we use the isotropic covariance function for α and Z processes in the SHP model. The separable covariance function is commonly used in GP models to allow different correlations for different design factors. By letting the covariance function of the Z process be separable, the SHP model can be easily extended to a separable SHP model. The SHP model is broadly applicable to spatial data (Palacios and Steele (2006), Huang et al. (2008)). Further, by adding a measurement error to (5), the SHP model can be extended in the context of non-parametric regression. These topics will be explored elsewhere.

6 Acknowledgment

This research was supported by NSF grant MSPA-CSE-0434354. The authors also wish to thank Don Estep for his help on the SIR example.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC.

- Chen, V., Tsui, K., Barton, R., and Allen, J. (2003). A review of design and modeling in computer experiments. *Handbook of Statistics*, 22:231–261.
- Chen, W. and Varadarajan, S. (1997). Integration of design of experiments and artificial neural network for achieving affordable concurrent design. *38th AIAA/ASME/ASCE/AHA/ASC Structures, Structural Dynamics, and Materials Conference and AIAA/ASME/AHS Adaptive Structures Forum*, 2:1316–1324.
- 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.
- Danielsson, J. and Richard, J. F. (1993). Accelerated Gaussian importance sampler with applications to dynamic latent variable models. *Journal of Applied Econometrics*, 8:153–173.
- Davis, R. A. and Rodriguez-Yam, G. (2005). Estimation for state-space models: an approximate likelihood approach. *Statistica Sinica*, 15:381–406.
- Durbin, J. and Koopmans, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84:669–684.
- Estep, D. and Neckels, D. (2006). Fast and reliable methods for determining the evolution of uncertain parameters in differential equations. *Journal of Computational Physics*, 213:530–556.
- Fang, K. T., Li, R., and Sudjianto, A. (2006). *Design and Modeling for Computer Experiments*. Boca Raton, FL : Chapman and Hall/CRC.
- Fuentes, M. and Smith, R. L. (2001). Modeling nonstationary processes as a convolution of local stationary processes. Technical report, North Carolina State University, Dept. of Statistics.
- Gelfand, A. E., Kim, H. J., Sirmans, C. F., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of American Statistical Association*, 98:387–396.
- Gramacy, R. B. (2007). tgp: An R package for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian process models. *Journal of Statistical Software*, 19(9).
- Gramacy, R. B., Lee, H. K. H., and Macready, W. G. (2004). Parameter space exploration with Gaussian process trees. *Proceedings of the 21st International Conference on Machine Learning*, pages 353–360.

- Higdon, D., Swall, J., and Kern, J. (1999). Non-stationary spatial modeling. In *Bayesian Statistics 6*, pages 761–768. Oxford: Oxford University Press. Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (Eds.).
- Huang, W., Wang, K., Breidt, F. J., and Davis, R. A. (2008). Spatial processes with heteroscedasticity.
- Jin, R., Chen, W., and Simpson, T. W. (2000). Comparative studies of metamodeling techniques under multiple modeling criteria. *8th AIAA/NASA/USAF/ISSMO Symposium on Multidisciplinary Analysis and Optimization*. AIAA, Long Beach, CA, AIAA-2000-4801.
- Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2:315–331.
- Palacios, M. B. and Steele, M. F. J. (2006). Non-Gaussian Bayesian geostatistical modeling. *Journal of the American Statistical Association*, 101:604–618.
- Qian, Z., Seepersad, C., Joseph, R., Allen, J., and Wu, C. F. J. (2006). Building surrogate models with detailed and approximate simulations. *AMSE Journal of Mechanical Design*, 128:668–677.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.
- Sacks, J. W., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiment. *Statistical Science*, 4:409–423.
- Sampson, P. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.*, pages 108–119.
- Santer, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer.
- Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields*, pages 1–67. Chapman and Hall, London. In: Cox, D. R., Hinkley, D. V. and Barndorff-Nielsen, O. E. (Eds.).
- Simpson, T. W., Lin, D. K. J., and Chen, W. (2001a). Sampling strategies for computer experiments: Design and analysis. *International Journal of Reliability and Applications*, 2(3):209–240.

- Simpson, T. W., Peplinski, J. D., Koch, P. N., and Allen, J. K. (2001b). Metamodels for computer-based engineering design: survey and recommendations. *The Journal of Engineering with Computers, Special Issue Honoring Professor Steven J. Fenves*, 17:129–150.
- Taylor, S. J. (1986). *Modelling Financial Time Series*. Chichester: John Wiley.
- Team, R. D. C. (2005). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0.
- Wang, K. (2008). *Spatial Models with Applications in Computer Experiments*. PhD thesis, Colorado State University, Fort Collins, CO.
- Xiong, Y., Chen, W., Apley, D., and X., D. (2007). A non-stationary covariance-based kriging method for metamodeling in engineering design. *International Journal for Numerical Methods in Engineer*, 71:733–756.
- Yan, J. (2007). Spatial stochastic volatility for lattice data. *Journal of Agricultural, Biological, and Environmental Statistics*, 12(1):25–40.