Statistica Sinica **28** (2018), 621-648
doi:https://doi.org/10.5705/ss.202015.0249

GAUSSIAN PROCESS MODELING WITH BOUNDARY INFORMATION

Matthias Hwai Yong Tan

City University of Hong Kong

Abstract: Gaussian process (GP) models are widely used to approximate time consuming deterministic computer codes, which are often models of physical systems based on partial differential equations (PDEs). Limiting or boundary behavior of the PDE solutions (e.g., behavior when an input tends to infinity) is often known based on physical considerations or mathematical analysis. However, widely used stationary GP priors do not take this information into account. It should be expected that if the GP prior is forced to reproduce the known limiting behavior, it will give better prediction accuracy and extrapolation capability. This paper shows how a GP prior that reproduce known boundary behavior of the computer model can be constructed. Real examples are given to demonstrate the improved prediction performance of the proposed approach.

Key words and phrases: Computer experiments, constrained Gaussian process emulator, extrapolation in finite element simulations.

1. Introduction

Due to the increase in computing power, there is widespread development and use of computer simulation models of physical systems. Many of the computer models are based on PDEs solved using finite difference or finite element methods. These deterministic simulators can be time consuming and are often approximated with cheap-to-compute surrogates constructed with data from computer experiments. GP models are a popular class of surrogates. In GP modeling, a stationary GP is often employed as a prior for the continuous functional relationship f between the real valued output $y \in \mathbb{R}$ and inputs $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}_+^d$, where $\mathbb{R}_+ = [0, \infty)$. This prior process is updated with experiment data, giving a posterior process that is used for inference about f. Note that we assume that the inputs are nonnegative real numbers because most physical quantities are represented by nonnegative numbers.

In many practical problems, there is additional information about the simulator. One such type of information concerns the output value as a set of inputs

approaches a boundary of the region on which they are defined. Such information can be obtained from an asymptotic analysis or simple physical considerations. For example, a parabolic PDE models the temperature of a solid body immersed in a fluid held at a fixed temperature as a function of space and time (Cengel and Ghajar (2011)). Clearly, as the time goes to infinity, the temperature at every point of the solid body converges to the temperature of the fluid. As another example, in the bending of a plate (Wang, Reddy and Lee (2000)), the maximum displacement approaches zero as the thickness of the plate approaches infinity. Consider modeling the temperature y at the midplane of a 1mm thick platinum plate with large length and width that is immersed in a fluid as a function of time x. The function is plotted in Figure 1 with crosses. We see that y quickly converge to the steady state asymptote of 1,200. The posterior mean and 98% prediction intervals of the stationary GP model constructed with the data marked as circles are plotted as dotted lines. It can be seen that the prediction intervals for $x > 300$ are too wide and they are widest when $x > 1{,}600$. Moreover, the posterior mean decreases and converge to the estimated prior mean, which is less than 1,200, when $x > 1{,}200$. These behaviors contradict the known fact that y is close to 1,200 for large x. The posterior mean and 98% prediction intervals for the proposed GP model, which we call the *Boundary Modified Gaussian Process* (BMGP) model, are plotted as solid lines. We see that the posterior mean and prediction intervals for the BMGP converge to 1,200 for $x > 300$, which is behavior expected of a valid model for this problem.

In the above examples, the available information can be written as

$$\lim_{\mathbf{x}^s \to \mathbf{c}^s} [f(\mathbf{x}) - a(\mathbf{x})] = 0, \tag{1.1}$$

where $\mathbf{x}^s = (x_{s_1}, \ldots, x_{s_l}), \{s_1, \ldots, s_l\} \subset \{1, \ldots, d\}, \mathbf{c}^s = (c_1^s, \ldots, c_l^s) \in (\mathbb{R}_+ \cup \{\infty\})^l$, and $a : \mathbb{R}_+^d \to \mathbb{R}$ is continuous. Note that $\mathbf{x}^s \to \mathbf{c}^s$ refers to a sequence $\{\mathbf{x}(g) \in \mathbb{R}_+^d : g = 1, 2, \ldots\}$ of $\mathbf{x}$ values such that $x_{s_i}(g) \to c_i^s$ and each $x_j(g), j \notin \{s_1, \ldots, s_l\}$ is a fixed positive value. An example of (1.1) is $\lim_{x_1 \to \infty}[f(x_1, x_2) - a] = 0$. Cases where $\lim_{\mathbf{x}^s \to \mathbf{c}^s} f(\mathbf{x}) = \infty$ can be handled within the framework given by (1.1) by transforming y so that the limit is finite (e.g., $1 - \exp(-ey)$, where e is a constant). When $d = 1$, asymptotes of f (in the sense of Definition 2.4 in Giblin (1972)) are special cases of (1.1). If $\mathbf{c}^s = \infty$ and $a(\mathbf{x}) = a \in \mathbb{R}$, the line $y = a$ is called a horizontal asymptote. If $\mathbf{c}^s = \infty$ and $a(\mathbf{x})$ is the equation of a line, the line represented by $a(\mathbf{x})$ is called an asymptote. Because $\mathbf{c}^s$ often lies at the boundary of the region in which $\mathbf{x}^s$ is defined (e.g., each c_i^s is either 0 or ∞ when each x_{s_i} can take on any nonnegative value), we shall use the term boundary

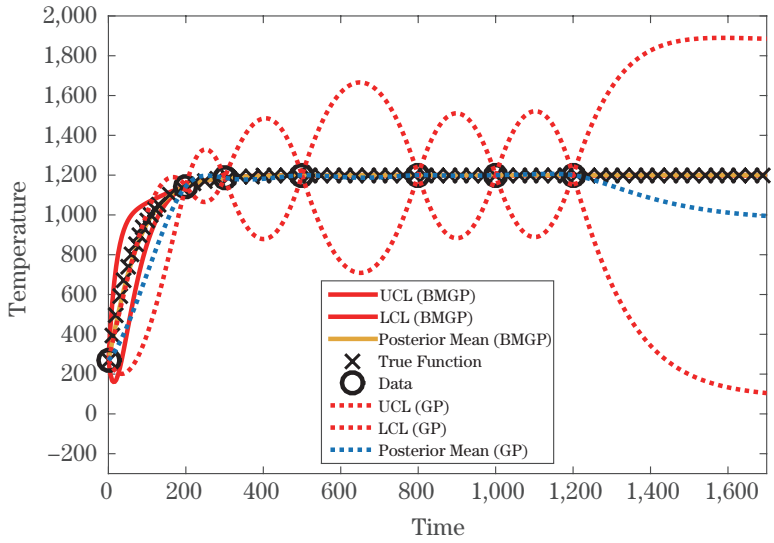

Figure 1. Point (posterior mean) and interval predictions [LCL, UCL] of BMGP and GP models constructed with a seven-point design for the platinum plate temperature model.

information to refer to the information given by (1.1). In a small number of cases, the output value as a set of inputs approaches a point in the interior of the region on which they are defined is known. For instance, the amplitude of vibration approaches infinity as the frequency of vibration approaches the resonant frequency. Problems such as this can also be handled by (1.1).

Aside from physical considerations, sources of information of the form (1.1) can be found from various simplifications commonly made to physical models. A physical quantity $\mathcal{P}$ is a function of three dimensional space and time, and is modeled as the solution of a PDE (Farlow (1982)). The solution of the equation is often simplified by certain assumptions. For instance, the behavior at time infinity of $\mathcal{P}$, called the steady state behavior, can be simpler to study mathematically because the time dimension is eliminated. For instance, both heat and wave equations reduce to the Poisson equation (Gockenbach (2011, p. 444)), which reduces to an ordinary differential equation if $\mathcal{P}$ depends only on one spatial dimension. The PDE can also be simplified when one assumes that physical properties that appear in the equation are spatially uniform, or certain terms in the equation are negligible. For example, the thermal conductivity in the parabolic heat equation can be assumed to be constant to simplify the equation so that explicit solution is possible (Cengel and Ghajar (2011, Chap. 4)). However, in a realistic uncertainty quantification setting, the spatial variation

of the thermal conductivity is approximated by a polynomial function (Xiu and Karniadakis (2003)). Clearly, a constant thermal conductivity is a limiting case of a polynomial function. As a second example, the telephone equation models the vibration of a string subject to friction (Coleman (2013, p. 59)). Neglecting the friction term reduces the telephone equation to the wave equation, which has a well-known explicit solution. As a third example, the complex Navier-Stokes equation can be simplified to the simple Bernoulli equation using assumptions such as zero viscosity (Humphrey and DeLange (2013, Chap. 8)). Finally, the solution of a PDE is also simplified by assuming that the size of the system is large or small in one or two spatial dimensions so that $\mathcal{P}$ is only a function of the remaining spatial dimensions. In structural mechanics, the governing equations for three-dimensional elasticity are simplified in this manner (plane stress and plane strain assumptions) (Johnson (2000)).

When $\{s_1, \ldots, s_l\}$ is a proper subset of $\{1, \ldots, d\}$, i.e., $\mathbf{x}^s$ does not include all components of $\mathbf{x}$, and $\mathbf{c}^s \in \mathbb{R}_+^l$, the information given by (1.1) is infinitely more than the values of the output at one or a finite number of points in the input space, which is the form of information obtained in a computer experiment. This is because (1.1) gives the value of f on a subset of a line, plane, hyperplane, or other linear manifolds (linear manifolds are sets of points that can be translated into a subspace of dimension $< d$). When some components of $\mathbf{c}^s$ equal ∞, (1.1) clearly provides an approximation to the value of f on infinitely many points. Thus, it is clear that the information given by (1.1) can be very valuable to incorporate in constructing a surrogate for the computer model and for extrapolation beyond the experiment region. This paper proposes a modified GP model to take into account the boundary information contained in (1.1). To the best of our knowledge, this problem has not been formulated in the computer experiments literature. Only the case where $\mathbf{x}^s = \mathbf{x}$ and $\mathbf{c}^s \in \mathbb{R}_+^d$ is trivial. For such a case, (1.1) gives the value of f at a point (since f is assumed continuous) and the information is easily incorporated as an additional point in the computer experiment data. Existing works on constrained models for computer experiments focus on monotonicity constraints. Some methods for this purpose are proposed in Golchi et al. (2015), Tan (2015) and Wang (2012). The BMGP model also appears to be a novel statistical approach to improve extrapolation accuracy. However, it is not a panacea to the extrapolation problem as it is possible to have correct boundary information but poor extrapolation performance.

The remainder of the paper is organized as follows: Section 2 reviews the stationary GP model. Section 3 gives the proposed BMGP model. Three real

examples are given in Section 4 to illustrate the improved performance achieved with the proposed model. Concluding remarks are given in Section 5.

2. Review of Gaussian Process Modeling

In the widely employed GP modeling approach, the prior for the functional relationship $f : \mathbb{R}_+^d \to \mathbb{R}$ between output and inputs is a stationary GP. The GP prior is

$$Y(\mathbf{x}) = \beta_0 + \mathcal{G}(\mathbf{x}), \tag{2.1}$$

where $\mathcal{G}(\mathbf{x})$ is a zero mean stationary GP. Given points $\mathbf{x}$ and $\mathbf{x}'$, the covariance of $Y(\mathbf{x})$ and $Y(\mathbf{x}')$ is $\text{cov}[Y(\mathbf{x}), Y(\mathbf{x}')] = \sigma_0^2 R(\mathbf{x}, \mathbf{x}'|\boldsymbol{\Theta}^0)$, where σ_0^2 is the variance and R is the correlation function with parameter $\boldsymbol{\Theta}^0 = (\theta_1^0, \ldots, \theta_d^0)$. It is common and sensible to choose R so that $R(\mathbf{x}, \mathbf{x}'|\boldsymbol{\Theta}^0) \to 0$ as $\|\mathbf{x} - \mathbf{x}'\|_2 \to \infty$ (Santner, Williams and Notz (2003)). This assumption implies that the output deviation from the mean β_0 at a point in the input region carries little information about the deviation at points far away. We shall adopt this assumption throughout the paper.

In a computer experiment, the output is evaluated at n values of inputs in the design $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, which yields a vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ of observed outputs. The prior process is updated with the data, giving a posterior GP (Sacks et al. (1989); Currin et al. (1991))

$$Y(\cdot)|(\mathbf{Y}, \beta_0, \sigma_0^2, \boldsymbol{\Theta}^0) \sim \text{GP}(M(\cdot|\beta_0, \boldsymbol{\Theta}^0), C(\cdot, \cdot|\sigma_0^2, \boldsymbol{\Theta}^0)), \tag{2.2}$$

with mean function $M(\cdot|\beta_0, \boldsymbol{\Theta}^0)$ and covariance function $C(\cdot, \cdot|\sigma_0^2, \boldsymbol{\Theta}^0)$. The mean function is

$$M(\mathbf{x}|\beta_0, \boldsymbol{\Theta}^0) = \beta_0 + \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{Y} - \beta_0 \mathbf{1}), \tag{2.3}$$

where $\mathbf{r}(\mathbf{x}) = (R(\mathbf{x}, \mathbf{x}_1|\boldsymbol{\Theta}^0), \ldots, R(\mathbf{x}, \mathbf{x}_n|\boldsymbol{\Theta}^0))^T$, $\mathbf{R} = (R(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\Theta}^0))_{1 \le i,j \le n}$ (a matrix with element $R(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\Theta}^0)$ in the ith row and jth column), and $\mathbf{1}$ is an $n \times 1$ vector of 1's. The covariance function is

$$C(\mathbf{x}, \mathbf{x}'|\sigma_0^2, \boldsymbol{\Theta}^0) = \sigma_0^2[R(\mathbf{x}, \mathbf{x}'|\boldsymbol{\Theta}^0) - \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1} \mathbf{r}(\mathbf{x}')]. \tag{2.4}$$

In a computer experiment, it is only possible to obtain a finite set of data, which should be concentrated within the region of interest $\mathcal{X}$. This is because $M(\mathbf{x}|\beta_0, \boldsymbol{\Theta}^0) \to \beta_0$ and $C(\mathbf{x}, \mathbf{x}|\sigma_0^2, \boldsymbol{\Theta}^0) \to \sigma_0^2$ as $\min\{\|\mathbf{x} - \mathbf{x}_1\|_2, \ldots, \|\mathbf{x} - \mathbf{x}_n\|_2\} \to \infty$, i.e., the posterior mean and variance of the GP converges to the prior mean and variance respectively as $\mathbf{x}$ gets further away from the design points. Thus, we shall assume that $\mathcal{D} \subset \mathcal{X}$.

The mean and covariance functions (2.3)-(2.4) depend on the parameters β_0, σ_0^2, and $\mathbf{\Theta}^0$, which can be estimated with the maximum likelihood method. Given $\mathbf{\Theta}^0$, $\beta_0 = \hat{\beta}_0 = (\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{R}^{-1}\mathbf{Y}$, and $\sigma_0^2 = \hat{\sigma}_0^2 = (\mathbf{Y}-\hat{\beta}_0\mathbf{1})^T\mathbf{R}^{-1}(\mathbf{Y}-\hat{\beta}_0\mathbf{1})/n$ maximize the likelihood. The maximum likelihood estimate $\hat{\mathbf{\Theta}}^0$ of $\mathbf{\Theta}^0$ is obtained by minimizing

$$n\log(\sigma_0^2) + \log(|\mathbf{R}|). \tag{2.5}$$

It is common to perform statistical inference on f using $\hat{\mathbf{\Theta}}^0$ in place of $\mathbf{\Theta}^0$, $\hat{\beta}_0$ in place of β_0, and $\hat{\sigma}_0^2$ in place of σ_0^2, i.e., using the GP

$$Y(\cdot)|(\mathbf{Y}, \hat{\beta}_0, \hat{\sigma}_0^2, \hat{\mathbf{\Theta}}^0) \sim \mathrm{GP}(M(\cdot|\hat{\beta}_0, \hat{\mathbf{\Theta}}^0), C(\cdot,\cdot|\hat{\sigma}_0^2, \hat{\mathbf{\Theta}}^0)). \tag{2.6}$$

This is an empirical Bayes approach. In GP modeling of computer experiments, the product Gaussian and product Matern correlation functions are often employed. In this paper, we employ the product Matern correlation function with smoothness parameter 1.5:

$$R(\mathbf{x}, \mathbf{x}'|\mathbf{\Theta}^0) = \prod_{i=1}^{d} \exp(-\rho_i)(\rho_i + 1), \rho_i = \frac{\sqrt{6}|x_i - x_i'|}{\theta_i^0}. \tag{2.7}$$

It is useful to standardize $x_1, \ldots, x_d$ in numerical optimization of (2.5) because it makes changes of the same magnitude in each component of $\mathbf{\Theta}^0$ to be of comparable importance.

3. GP Modeling with Boundary Information

This section describes our proposed modification of the GP model to incorporate boundary information given by

$$\lim_{\mathbf{x}^i \to \mathbf{c}^i} [f(\mathbf{x}) - a^i(\mathbf{x})] = 0, i = 1, \ldots, k, \tag{3.1}$$

where $\mathbf{x}^i = (x_{i_1}, \ldots, x_{i_{l(i)}})$, $\{i_1, \ldots, i_{l(i)}\} \subset \{1, \ldots, d\}$, $\mathbf{c}^i = (c_1^i, \ldots, c_{l(i)}^i) \in (\mathbb{R}_+ \cup \{\infty\})^{l(i)}$, and a^i is continuous. We assume all $\mathbf{c}^i, i = 1, \ldots, k$ are *distinct*, i.e., if $x_{i_h} = x_{j_m}$ for $i \neq j$ ($\mathbf{x}^i$ and $\mathbf{x}^j$ in the ith and jth equations in (3.1) include a common input), then $c_h^i \neq c_m^j$. The point $\mathbf{c}^i$ is typically on a boundary of the region $\mathbb{R}_+^{l(i)}$ in which $\mathbf{x}^i$ is defined, i.e., it includes components that equal 0 and ∞. Thus, $\{\mathbf{x} \in (\mathbb{R}_+ \cup \{\infty\})^d : \mathbf{x}^i = \mathbf{c}^i\}$ is often an edge of the boundary of $\mathbb{R}_+^d$. In some engineering problems, an input may be defined over a region other than $[0, \infty)$ (e..g, the percentage of a substance in a mixture). In those cases, $\mathbf{c}^i$ is also often on the boundary of the region in which $\mathbf{x}^i$ is defined. The case $\lim_{\mathbf{x}^i \to \mathbf{c}^i} f(\mathbf{x}) = \infty$ can be transformed into a special case of (3.1) by using the transformation $1 - \exp(-ey)$ for some e that may be estimated with

data. Unlike $1/y$, this transformation allows us to handle cases where there exists a vertical asymptote $\lim_{\mathbf{x}^i \to \mathbf{c}^i} f(\mathbf{x}) = \infty$ and a zero horizontal asymptote $\lim_{\mathbf{x}^j \to \mathbf{c}^j} f(\mathbf{x}) = 0$. Although a $\mathbf{c}^i$ with a component that equals ∞ can seem far from the region of interest $\mathcal{X}$, it is often the case that f converges quickly enough so that the information in (3.1) is useful for improving prediction within $\mathcal{X}$. Moreover, the information is useful for improving extrapolation accuracy.

As reviewed in Section 2, stationary GP models can incorporate information about known function values at a finite set of points (which is the kind of information obtained in a computer experiment). Thus, when each $\mathbf{x}^i = \mathbf{x}$ and $\mathbf{c}^i \in \mathbb{R}^d_+$ (does not have ∞ as a component), we can incorporate the information in (3.1) by including $(\mathbf{c}^i, a^i(\mathbf{c}^i)), i = 1, \ldots, k$ as data points. However, the information in (3.1) cannot be easily incorporated in the stationary GP model when $\mathbf{x}^i$ is a proper subset of $\mathbf{x}$ or when some components of $\mathbf{c}^i$ is ∞. For example, when $d = 2$ and $\lim_{x_1 \to 0} f(x_1, x_2) = a$, the value of y on an entire ray is known. Even if we update the GP model with several data points $(\mathbf{x}, y)$ of the form $((0, x_2), a)$, the GP model can still fail to predict accurately on the edge $x_1 = 0$. Moreover, since we know y converges to a constant as $x_1 \to 0$, we should have decreasing uncertainty about the function f as $x_1 \to 0$. This fact is not taken into account in a stationary GP model because the prior variance is constant. As another example, suppose x_1 represents time, and $\lim_{x_1 \to \infty} f(x_1, x_2) = a$, i.e., the steady state value of y is known. There is no existing method for exploiting this information. If a stationary GP model is used, the posterior mean when $x_1 \to \infty$ is the prior mean and the posterior variance when $x_1 \to \infty$ is the prior variance. This gives poor extrapolation behavior.

We propose to incorporate the information given by (3.1) into the BMGP model by choosing its prior mean function $\mu(\mathbf{x})$ so that

$$\lim_{\mathbf{x}^i \to \mathbf{c}^i} [\mu(\mathbf{x}) - a^i(\mathbf{x})] = 0, i = 1, \ldots, k, \tag{3.2}$$

and its prior variance function $\sigma^2(\mathbf{x})$ so that

$$\lim_{\mathbf{x}^i \to \mathbf{c}^i} \sigma^2(\mathbf{x}) = \gamma_i, i = 1, \ldots, k. \tag{3.3}$$

In this case, the GP prior that we use has *nonstationary mean and variance*. If $\gamma_i = 0$, this ensures that the prior process Y for the BMGP model converges in mean square to $a^i(\mathbf{x})$ as $\mathbf{x}^i \to \mathbf{c}^i$, i.e., $\lim_{\mathbf{x}^i \to \mathbf{c}^i} E\{[Y(\mathbf{x}) - a^i(\mathbf{x})]^2\} = 0, i = 1, \ldots, k$. However, to make the BMGP model robust to misspecifications in (3.1), we shall allow γ_i to be a small positive number.

We do not know a way to incorporate the boundary information through the

covariance function alone (as suggested by a referee). Let M and C denote the posterior mean and covariance functions of the GP, and let μ and Σ denote the prior mean and covariance functions. Note that it is common and sensible to choose the prior correlation function R so that $R(\mathbf{x}, \mathbf{x}') \to 0$ as $\|\mathbf{x} - \mathbf{x}'\|_2 \to \infty$ (see first paragraph of Section 2) and to choose a bounded prior variance function $\Sigma(\mathbf{x}, \mathbf{x})$ (i.e., we are not completely ignorant about the values that the output can take at any $\mathbf{x}$). If such a choice is made, $M(\mathbf{x}) \to \mu(\mathbf{x})$ and $C(\mathbf{x}, \mathbf{x}') \to \Sigma(\mathbf{x}, \mathbf{x}')$ as $\min\{\|\mathbf{x} - \mathbf{x}_1\|_2, \ldots, \|\mathbf{x} - \mathbf{x}_n\|_2, \|\mathbf{x}' - \mathbf{x}_1\|_2, \ldots, \|\mathbf{x}' - \mathbf{x}_n\|_2\} \to \infty$, i.e., the posterior mean and covariance functions of the GP converge to the prior mean and covariance functions respectively as $\mathbf{x}$ and $\mathbf{x}'$ get further away from the design points. Thus, if we choose $\mu = \beta_0$ as in standard GP modeling, the GP prediction will not converge to the known boundary values and the posterior variance at the boundary will not be small or zero.

Assuming that the computer model is known to have continuous partial derivatives (differentiable) within a region, we should specify $\mu(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ so that the GP prior has mean square partial derivatives that are mean square continuous (*mean square differentiable*) within the region. This would give a GP prior that is consistent with the prior knowledge that the true function is differentiable. The GP prior is mean square differentiable if the following conditions hold:

1. The mean function $\mu(\mathbf{x})$ has continuous partial derivatives.
2. The covariance function $C(\mathbf{x}, \mathbf{x}') = \sigma(\mathbf{x})\sigma(\mathbf{x}')R(\mathbf{x}, \mathbf{x}')$, where R is the correlation function, has continuous mixed partial derivatives $\partial^2 C(\mathbf{x}, \mathbf{x}')/\partial x_i \partial x_i'$ at all points $(\mathbf{x}, \mathbf{x}') = (\mathbf{x}, \mathbf{x})$ (Adler (2010, p. 27)).

Similar conditions can be stated to guarantee that the BMGP model has mean square partial derivatives of higher order that are mean square continuous. Clearly, if information about the rate of convergence in (3.1) is available, that information should be used to specify $\mu(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ for the BMGP model. Nevertheless, such information is difficult to obtain.

It is our opinion that a mean square differentiable GP is a good choice for a differentiable computer code due to two reasons. First, mean square differentiability is mathematically simple to verify. Second, it implies differentiability in probability, which means that a finite difference approximation of the derivative is close to the mean square derivative with high probability. Nevertheless, the BMGP model proposed in Section 3.1 yields differentiable sample paths with probability one because the mean is differentiable, and the centered BMGP is a differentiable scale multiple of a stationary GP with Matern correlation func-

tion (2.7).

3.1. Choice of prior mean and variance functions

We propose the following method to specify $\mu(\mathbf{x})$ and $\sigma^2(\mathbf{x})$. Let $p^i(\mathbf{x}) = \mathbf{x}^i$ and $d_i(p^i(\mathbf{x}), \mathbf{c}^i) = d_i(\mathbf{x}^i, \mathbf{c}^i)$ be a measure of distance between $\mathbf{x}^i$ and $\mathbf{c}^i$ (e.g., distance metric on $\mathbb{R}_+ \cup \{\infty\}$) that is bounded above by 1. Then, define

$$\lambda_0(\mathbf{x}) = \frac{\sum_{j=1}^k d_j^2(p^j(\mathbf{x}), \mathbf{c}^j)}{\sum_{j=1}^k d_j^2(p^j(\mathbf{x}), \mathbf{c}^j) + \sum_{j=1}^k \alpha_j / d_j^2(p^j(\mathbf{x}), \mathbf{c}^j)},$$

$$\lambda_i(\mathbf{x}) = \frac{\alpha_i / d_i^2(p^i(\mathbf{x}), \mathbf{c}^i)}{\sum_{j=1}^k d_j^2(p^j(\mathbf{x}), \mathbf{c}^j) + \sum_{j=1}^k \alpha_j / d_j^2(p^j(\mathbf{x}), \mathbf{c}^j)},$$

$$i = 1, \ldots, k, \tag{3.4}$$

where $\alpha_j, j = 1, \ldots, k$ are positive constants to be estimated. We set

$$\mu(\mathbf{x}) = \sum_{i=0}^{k} a^i(\mathbf{x}) \lambda_i(\mathbf{x}), \tag{3.5}$$

where a^0 is a constant. Thus, the mean is a convex combination of $a^0, a^1(\mathbf{x}), \ldots, a^k(\mathbf{x})$. The weight function $\lambda_i(\mathbf{x})$ given to the ith piece of boundary information increases when α_i increases or when the distance to $\mathbf{c}^i$ decreases. If some or all α_i's get smaller, then the weight given to a^0 gets larger. This choice of weight function is motivated by the inverse distance interpolator (Gordon and Wixom (1978)). Note that if $\mathbf{x}$ is changed in such a way that $p^i(\mathbf{x})$ gets further from $\mathbf{c}^i$ for all $i = 1, \ldots, k$, then $\mu(\mathbf{x})$ will be closer to a^0. Thus, if $\mathbf{x}$ is far from the boundaries in (3.1) and the α_i's are small enough, the prior mean is nearly constant, which is a choice motivated by the stationary GP model. Moreover, the following results imply that the prior mean (3.5) reproduces the known boundary behavior (3.1).

Theorem 1. *Assume for all $j = 1, \ldots, k$ that $d_j^2(p^j(\mathbf{x}), \mathbf{c}^j)$ is a continuous function (with respect to the sequence definition of continuity) of $\mathbf{x}$ with the extended real hyperplane as domain and range in $[0, 1]$, $d_j^2(p^j(\mathbf{x}), \mathbf{c}^j) = 0$ if and only if $p^j(\mathbf{x}) = \mathbf{c}^j$, and $d_j^2(p^j(\mathbf{x}), \mathbf{c}^j)$ is a strictly increasing function of the absolute value of each component of $p^j(\mathbf{x}) - \mathbf{c}^j$ (for $c_m^j = \infty$, we assume d_j^2 is strictly decreasing in x_{j_m}). Consider a sequence $\{\mathbf{x}(g) \in \mathbb{R}_+^d : g = 1, 2, \ldots\}$ such that $\mathbf{x}^i(g) \to \mathbf{c}^i$ (where $\mathbf{x}^i(g) \to \mathbf{c}^i$ refers to component-wise convergence), all $x_m(g), m \in \{1, \ldots, d\} \backslash \{i_1, \ldots, i_{l(i)}\}$ are fixed, and $\mathbf{x}^j(g) \neq \mathbf{c}^j$ for all $j \neq i$ and g. Then, $\lambda_i(\mathbf{x}(g)) \to 1$ and for all $j \neq i$, $\lambda_j(\mathbf{x}(g)) \to 0$.*

Corollary 1. *Consider a sequence* $\{\mathbf{x}(g) \in \mathbb{R}_+^d : g = 1, 2, \dots\}$ *of* $\mathbf{x}$ *values such that* $\mathbf{x}^i(g) \to \mathbf{c}^i$, *all* $x_m(g), m \in \{1, \dots, d\} \backslash \{i_1, \dots, i_{l(i)}\}$ *are fixed, and* $\mathbf{x}^j(g) \neq \mathbf{c}^j$ *for all* $j \neq i$ *and* g. *Assume* $\limsup_{g\to\infty} |a^j(\mathbf{x}(g))| < \infty, j = 1, \dots, k$ *and the conditions on* $d_j^2(p^j(\mathbf{x}), \mathbf{c}^j)$ *in Theorem* 1 *hold. Then,* $\mu(\mathbf{x}(g)) - a^i(\mathbf{x}(g)) \to 0$.

Our choice of the prior variance function is

$$\sigma^2(\mathbf{x}) = s^2 \prod_{i=1}^{k} \{[d_i^2(p^i(\mathbf{x}), \mathbf{c}^i)]^{\eta_i} + \delta\}^2, \tag{3.6}$$

where s^2 and $\eta_j, j = 1, \dots, k$ are positive constants to be estimated. We see that $s^2\delta^{2k}$ is a lower bound for $\sigma^2(\mathbf{x})$. Moreover, if $\{\mathbf{x}(g) : g = 1, 2, \dots\}$ is a sequence such that $\mathbf{x}(g) \to \mathbf{x}^*$ and $\mathbf{x}^j \to \mathbf{c}^j$, then $\sigma^2(\mathbf{x}) \to s^2\delta^2 \prod_{i=1, i\neq j}^{k} \{[d_i^2(p^i(\mathbf{x}^*), \mathbf{c}^i)]^{\eta_i} + \delta\}^2$. Note that η_i affects the rate of decrease of $\sigma^2(\mathbf{x})$ as $\mathbf{x}$ approaches the ith boundary. The larger the η_i, the faster the rate of decrease. Setting $\delta = 0$ gives $\lim_{\mathbf{x}^i \to \mathbf{c}^i} \sigma^2(\mathbf{x}) = 0, i = 1, \dots, k$, which implies that we are certain a priori that $f(\mathbf{x}) - a^i(\mathbf{x}) \to 0$ as $\mathbf{x}^i \to \mathbf{c}^i$. However, we do not set $\delta = 0$ due to a few reasons. If $\delta = 0$, the covariance matrix of the response at the design points can be degenerate if $d_i^2(p^i(\mathbf{x}_j), \mathbf{c}^i) = 0$ for some design point $\mathbf{x}_j$. Moreover, not setting $\delta = 0$ provides the model some robustness to misspecifications of some of the limits in (3.1).

The assumption that $d_i^2(p^i(\mathbf{x}), \mathbf{c}^i)$ is a continuous measure of distance between $\mathbf{x}^i$ and $\mathbf{c}^i$ with range in $[0, 1]$ as stated in Theorem 1 is important for three reasons. First, Theorem 1 and Corollary 1 hold under the assumptions, which guarantee that the prior mean reproduces known boundary behavior. Second, as we need the prior variance to converge to a small value when $\mathbf{x}$ approaches a boundary in (3.1), the prior variance should be a function of a distance measure that satisfies the conditions in Theorem 1. Third, because we use (3.6), it is sensible to restrict $d_i^2(p^i(\mathbf{x}), \mathbf{c}^i) \in [0, 1]$. Otherwise, the effect of η_i on $\sigma^2(\mathbf{x})$ will be different depending on whether $d_i^2(p^i(\mathbf{x}), \mathbf{c}^i) > 1$ or $d_i^2(p^i(\mathbf{x}), \mathbf{c}^i) < 1$. Moreover, it seems more reasonable to have bounded prior variance, as with a stationary GP prior.

Note that (3.5) and (3.6) contain a total of $2k + 3$ parameters that need to be estimated. To reduce computational burden, we set $\alpha_1 = \cdots = \alpha_k = \alpha$ and $\eta_1 = \cdots = \eta_k = \eta$. When the parameters are estimated with the maximum likelihood method (Section 3.2), this makes the performance of the BMGP model more reliable than the case where the α_i's and η_j's are allowed to be different. It seems that when the α_i's or η_j's are allowed to be different, there can be many local optimizers in the likelihood function, which makes discovery of the

global maximum difficult. Moreover, different local optimizers can give quite different results. Intuitively, the BMGP model should be more sensitive to the choice of parameters $\eta_1, \ldots, \eta_k$ than to the choice of $\alpha_1, \ldots, \alpha_k$. A change in the α_i's will change the prior mean function. However, the posterior mean function will interpolate the data and satisfy (3.2) whatever positive values of α_i's are employed. In contrast, a change in η_i changes the value of the prior variance, which is approximately the posterior variance at points sufficiently far away from the data points. Thus, coverage far from the design region will be affected by the choice of η_i. When the correlations are weak, coverage within the design region will be affected as well.

We use the following distance measure (which satisfies the conditions in Theorem 1)

$$d_i^2(\mathbf{x}^i, \mathbf{c}^i) = \frac{\varphi(x_{i_1}, c_1^i)^2 + \cdots + \varphi(x_{i_{l(i)}}, c_{l(i)}^i)^2}{l(i)}, \tag{3.7}$$

where $l(i)$ is the number of components of $\mathbf{x}^i$ as in (3.1),

$$\varphi(x_{i_m}, c_m^i) = \left| \frac{U_{i_m}}{U_{i_m} + x_{i_m}} - \frac{U_{i_m}}{U_{i_m} + c_m^i} \right| \tag{3.8}$$

and U_{i_m} = mean of x_{i_m} values in the design, which is strictly positive. The rationale for choosing (3.7) and (3.8) is as follows. First, they satisfy all conditions stated in Theorem 1. Note that $\varphi(x_{i_m}, c_m^i)$ has range $[0, 1]$ and because we divide the sum in (3.7) by $l(i)$, we have $0 \le d_i^2(\mathbf{x}^i, \mathbf{c}^i) \le 1$. Second, $\varphi(x_{i_m}, c_m^i)$ in (3.8) is a valid metric on $\mathbb{R}_+ \cup \{\infty\}$ (in particular, the triangle inequality is satisfied), and $d_i(\mathbf{x}^i, \mathbf{c}^i)$ in (3.7) is a metric on $(\mathbb{R}_+ \cup \{\infty\})^{l(i)}$. We set U_{i_m} equal to the mean of the x_{i_m} values in the design because in this case, $\varphi(U_{i_m}, 0) = 1/2$, $\varphi(U_{i_m}, \infty) = 1/2$, and $\varphi(0, \infty) = 1$, i.e., the distance between U_{i_m} and 0 is half of the distance between ∞ and 0 (which equals the maximum of 1). Based on the above discussion, we see that (3.7) and (3.8) provide a sensible measure of distance between $\mathbf{x}^i$ and $\mathbf{c}^i$. For the case where x_{i_m} is defined on a bounded interval, we can simply use a standardized Euclidean metric instead of (3.8). Finally, when x_{i_m} is defined on $\mathbb{R}$, we can use the metric $|x_{i_m}/(U_{i_m} + |x_{i_m}|) - c_m^i/(U_{i_m} + |c_m^i|)|$, which is given in Haaser and Sullivan (1991, p. 59).

Because $\varphi(x_{i_m}, c_m^i)^2$ in (3.8) is continuously differentiable with respect to each x_{i_m} at all $x_{i_m} \in \mathbb{R}_+$, each $\lambda_i(\mathbf{x})$ given by (3.4) is continuously differentiable with respect to x_l at all points $\mathbf{x}$ in $\mathbb{R}_+^d$ such that $p^j(\mathbf{x}) = \mathbf{c}^j$ for at most one j. To prove this, we simply need to multiply the numerator and denominator of (3.4) by $\prod_{j=1}^k d_j^2(p^j(\mathbf{x}), \mathbf{c}^j)$. Thus, $\mu(\mathbf{x})$ given by (3.5) is continuously differentiable

with respect to x_l at all such points if each $a^i(\mathbf{x})$ also has the same property. It is also easy to see that $\sqrt{\sigma^2(\mathbf{x})}$ given by (3.6) has continuous partial derivatives at all $\mathbf{x}$ in $\mathbb{R}^d_+$ such that $d_i^2(p^i(\mathbf{x}), \mathbf{c}^i) \neq 0$ for all $i = 1, \ldots, k$. Thus, the BMGP model with the mean and variance functions given by (3.4)-(3.8), and correlation function R in (2.7) is mean square differentiable at all $\mathbf{x} \in \mathbb{R}^d_+$ such that $p^i(\mathbf{x}) \neq \mathbf{c}^i$ for all $i = 1, \ldots, k$.

An example of an alternative choice to (3.4) and (3.5) is $\mu(\mathbf{x}) = \sum_{i=1}^k a^i(\mathbf{x})$ $\lambda_i(\mathbf{x})$, where $\lambda_i(\mathbf{x}) = d_i^{-\alpha_i}(p^i(\mathbf{x}), \mathbf{c}^i)/[\sum_{j=1}^k 1/d_j^{\alpha_j}(p^j(\mathbf{x}), \mathbf{c}^j)]$, and $\alpha_1, \ldots, \alpha_k$ are positive numbers. An alternative choice to (3.6) is

$$\sigma^2(\mathbf{x}) = s^2 \prod_{i=1}^k \{G[d_i^2(p^i(\mathbf{x}), \mathbf{c}^i); \eta_i] + \delta\}^2, \tag{3.9}$$

where $G(z; \eta_i)$ is a strictly increasing CDF with parameter η_i and $G(0; \eta_i) = 0$. For example, we can take G to be the CDF of a Beta distribution with density proportional to $z^{\eta_i - 1}(1 - z)^{\eta_i - 1}$. We recommend the prior mean and variance functions given by (3.4)-(3.8) as we have found that they give reliable and excellent performance in general.

3.2. Parameter estimation

This section discusses estimation of the parameters $a^0, \alpha, s^2, \delta, \eta$ in the mean and variance functions, and the vector of parameters $\boldsymbol{\Theta}$ in the correlation function using the method of maximum likelihood. Cases with α_i's or η_j's that are not constrained to be the same are similarly handled. Clearly, a fully Bayesian approach can also be implemented with the use of Markov Chain Monte Carlo methods.

The design is $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and the output is $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. Define $\boldsymbol{\Lambda} = (\lambda_0(\mathbf{x}_1), \ldots, \lambda_0(\mathbf{x}_n))^T$, $\boldsymbol{\Phi} = (\sum_{j=1}^k a^j(\mathbf{x}_1)\lambda_j(\mathbf{x}_1), \ldots, \sum_{j=1}^k a^j(\mathbf{x}_n)\lambda_j(\mathbf{x}_n))^T$, and

$$\mathbf{Q} = \left(\prod_{l=1}^k \{[d_l^2(p^l(\mathbf{x}_i), \mathbf{c}^l)]^\eta + \delta\} \prod_{l=1}^k \{[d_l^2(p^l(\mathbf{x}_j), \mathbf{c}^l)]^\eta + \delta\} R(\mathbf{x}_i, \mathbf{x}_j | \boldsymbol{\Theta}) \right)_{1 \leq i,j \leq n}.$$

We shall use the product Matern correlation function given in (2.7). The posterior process, which we use to construct point and interval predictions, is

$$Y(\cdot)|(\mathbf{Y}, a^0, \alpha, s^2, \delta, \eta, \boldsymbol{\Theta}) \sim \mathrm{GP}(M(\cdot|a^0, \alpha, \delta, \eta, \boldsymbol{\Theta}), C(\cdot, \cdot|s^2, \delta, \eta, \boldsymbol{\Theta})), \tag{3.10}$$

with mean function $M(\cdot|a^0, \alpha, \delta, \eta, \boldsymbol{\Theta})$ and covariance function $C(\cdot, \cdot|s^2, \delta, \eta, \boldsymbol{\Theta})$. The mean function is

$$M(\mathbf{x}|a^0, \alpha, \delta, \eta, \boldsymbol{\Theta}) = \mu(\mathbf{x}) + \mathbf{q}(\mathbf{x})^T \mathbf{Q}^{-1}(\mathbf{Y} - \boldsymbol{\Phi} - a^0 \boldsymbol{\Lambda}), \tag{3.11}$$

where $\mathbf{q}(\mathbf{x}) = (\prod_{i=1}^{k}\{[d_i^2(p^i(\mathbf{x}),\mathbf{c}^i)]^\eta + \delta\}\prod_{i=1}^{k}\{[d_i^2(p^i(\mathbf{x}_j),\mathbf{c}^i)]^\eta + \delta\}R(\mathbf{x},\mathbf{x}_j|\boldsymbol{\Theta}))_{j=1,\ldots,n}$. The covariance function is given by

$$C(\mathbf{x},\mathbf{x}'|s^2,\delta,\eta,\boldsymbol{\Theta}) = \sigma(\mathbf{x})\sigma(\mathbf{x}')R(\mathbf{x},\mathbf{x}'|\boldsymbol{\Theta}) - s^2\mathbf{q}(\mathbf{x})^T\mathbf{Q}^{-1}\mathbf{q}(\mathbf{x}'). \tag{3.12}$$

Given α, δ, η, and $\boldsymbol{\Theta}$, it can be shown that $a^0 = \hat{a}^0 = (\boldsymbol{\Lambda}^T\mathbf{Q}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\mathbf{Q}^{-1}(\mathbf{Y}-\boldsymbol{\Phi})$ and $s^2 = \hat{s}^2 = (\mathbf{Y}-\boldsymbol{\Phi}-\hat{a}^0\boldsymbol{\Lambda})^T\mathbf{Q}^{-1}(\mathbf{Y}-\boldsymbol{\Phi}-\hat{a}^0\boldsymbol{\Lambda})/n$ maximize the likelihood. Thus, the maximum likelihood estimates $\hat{\alpha}$, $\hat{\delta}$, $\hat{\eta}$, and $\hat{\boldsymbol{\Theta}}$ of α, δ, η, and $\boldsymbol{\Theta}$ are obtained by minimizing

$$n\log(\hat{s}^2) + \log(|\mathbf{Q}|). \tag{3.13}$$

We perform statistical inference on f using $\hat{\boldsymbol{\Theta}}, \hat{\eta}, \hat{\delta}, \hat{\alpha}, \hat{s}^2, \hat{a}^0$ in place of $\boldsymbol{\Theta}, \eta, \delta, \alpha, s^2, a^0$. In optimizing (3.13), we use the BFGS algorithm (Nocedal and Wright (2006)) implemented in the Matlab fminunc function. We actually optimize the logarithm of the positive quantities α, δ, η, and $\boldsymbol{\Theta}$ because fminunc is for unconstrained optimization. We use the starting values of 0.5,1,2,3 for η, a single starting value of e^{-2} for δ, and two starting values of $(0.91)\mathbf{1}$ and $(2.55)\mathbf{1}$ for $\boldsymbol{\Theta}$. The starting value for α is $\min\{\sum_{j=1}^{k} d_j^2(p^j(\mathbf{x}_i),\mathbf{c}^j) : i = 1,\ldots,n\}/\min\{\sum_{j=1}^{k} 1/d_j^2 (p^j(\mathbf{x}_i),\mathbf{c}^j) : i = 1,\ldots,n\}$, which yields rather balanced weights to a^0 and each a^i in (3.5) at the design points. We should check that $\hat{\delta}$ is small. Otherwise, the accuracy of the information (3.1) is suspect. In all examples in Section 4, we obtain $\hat{\delta} < 0.1$.

Finally, we shall provide some guidelines for specifying the starting values of α and η. The starting value of α should be less than 100 as a value of 100 means that even at a point $\mathbf{x}$ with the maximum distance of 1 from $\mathbf{c}^1,\ldots,\mathbf{c}^k$, the weight given to each piece of boundary information in (3.5) is $100/k$ times the weight given to a^0 and the total weight given to all k pieces of boundary information is 100 times the weight given to a^0. As the variance is approximately proportional to the distance to a boundary raised to the power of η when δ is very small, the starting value for η should be between 1/7 and 7 so that the increase in variance is not too fast when the distance to the boundary is near zero or near one. Note that ranges above can also be used as bounds in the optimization to exclude nonsensical estimates.

3.3. Model checking and validation

The BMGP model proposed in Sections 3.1-3.2 can yield significant improvements in interpolation and extrapolation accuracy over the stationary GP model. However, it may not always work well. Thus, the model should be validated if possible. In model validation, prediction performance within the experiment re-

gion and extrapolation performance are quite different issues. While the former can be checked via crossed validation (which is discussed in this section), the latter can only be checked by comparing the model's prediction against the true output in a test set contained in the region of extrapolation.

Bastos and O'Hagan (2009) provide a few diagnostic statistics for checking the adequacy of the GP model. Those statistics can be applied in a straightforward manner to the BMGP model. However, a test set not used to build the model is needed to check for model adequacy with the diagnostic statistics proposed by Bastos and O'Hagan (2009). We check the adequacy of the BMGP model for predicting within the experiment region using leave-one-out cross validation (Mitchell and Morris (1992)). Assuming that none of the model parameters are re-estimated after leaving each data point out, the vector of leave-one-out cross validation errors is given by the short-cut formula

$$\mathbf{E} = \text{diag}\{\mathbf{Q}^{-1}\}^{-1}\mathbf{Q}^{-1}(\mathbf{Y} - \mathbf{\Phi} - a^0\mathbf{\Lambda}), \tag{3.14}$$

where $\text{diag}\{\mathbf{Q}^{-1}\}$ is a diagonal matrix with the same diagonal as $\mathbf{Q}^{-1}$. The leave-one-out error cannot be used to determine whether the proposed model gives sufficiently accurate extrapolation prediction. It is also important to check whether the prediction interval constructed without a data point contains the data point. Let us partition

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & Q_{22} \end{pmatrix}, \tag{3.15}$$

where $Q_{22}s^2$ is the prior variance of the data point that is left out. Partition

$$\mathbf{Q}^{-1} = \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & S_{22} \end{pmatrix} \tag{3.16}$$

correspondingly. Then, given $\mathbf{Q}^{-1}$, a short cut formula for computing the inverse of $\mathbf{Q}_{11}$, which is the prior covariance matrix (discounting the factor s^2) of the data without the point that is left out, is

$$\mathbf{Q}_{11}^{-1} = \mathbf{S}_{11} - \frac{\mathbf{S}_{12}\mathbf{S}_{21}}{S_{22}}. \tag{3.17}$$

This formula, together with (3.12), allows quick computation of the prediction interval for a data point given that the point is left out from the data used to fit the BMGP model.

Note that while the BMGP model may produce inaccurate predictions within the experiment region $\mathcal{X}$ in the finite sample case, it should not fail for large sample sizes. It should be true that the posterior mean of the BMGP model

converges to the true function and the posterior variance converges to zero at all points within $\mathcal{X}$ when the data get dense in $\mathcal{X}$. This is because the BMGP model interpolates the data and has continuous posterior mean and variance functions. Thus, if the BMGP model is not sufficiently accurate, one can improve its predictions within $\mathcal{X}$ by adding more design points instead of by changing its mean or variance function. Moreover, because the stationary GP model is a special case of the BMGP model ($\alpha = \eta = 0, \delta = 1$), the BMGP model can only perform poorer than the stationary GP model if the estimates of the BMGP model parameters obtained by numerically maximizing the likelihood function are poor choices for the parameters. Fortunately, this problem often manifests itself in a huge estimate for s^2 (see the last paragraph of Section 4.3), which will not go without notice. In some cases, due to insufficient starting points for optimization, an innocent-looking local optimizer of the likelihood function that gives poorer performance than the global optimizer is found. This problem can only be avoided with the use of many starting points. Lastly, because the stationary GP model is a special case of the BMGP model and the former model often has excellent prediction performance within $\mathcal{X}$, the choices (3.4)-(3.8) of the mean and variance functions only need careful scrutiny and modification when we are interested to achieve better extrapolation performance with the BMGP model.

4. Examples

This section presents three realistic examples to compare the performance of the BMGP model (Section 3) and the stationary GP model (Section 2).

4.1. Example 1: Platinum plate temperature

We revisit the platinum plate temperature example described in the introduction. The initial temperature of the plate is 270K, the fluid temperature is 1,200K, and the heat transfer mechanism is convection and conduction. The thermal diffusivity and thermal conductivity of platinum are obtained from tabulated values in Cengel and Ghajar (2011). The convective heat transfer coefficient is taken to be $20\text{W}/(\text{m}^2\text{K})$. The temperature at the midplane of the plate $y = f(x)$ as a function of time x is obtained from the solution of a parabolic PDE, which is given explicitly as an infinite series in Cengel and Ghajar (2011). We know that

$$\lim_{x\to\infty} [f(x) - 1200] = 0. \tag{4.1}$$

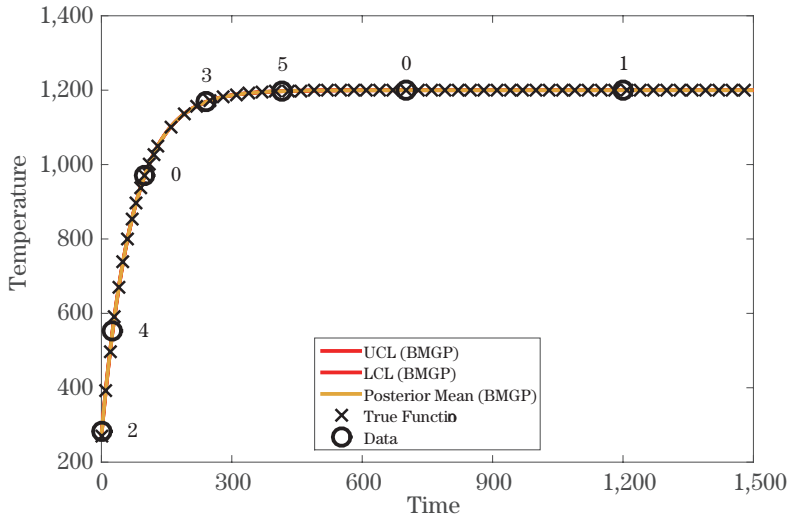

Figure 2. Point (posterior mean) and interval predictions [LCL, UCL] of BMGP model constructed with sequential design, platinum plate temperature example.

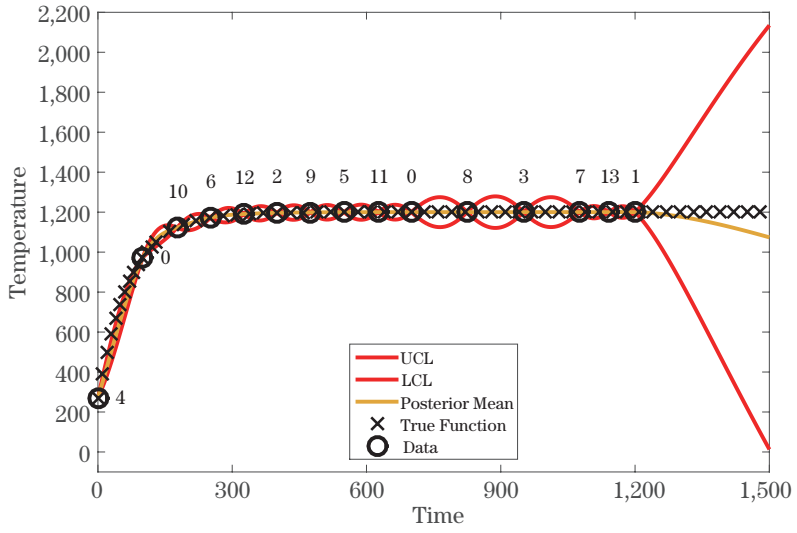

Figure 3. Point (posterior mean) and interval predictions [LCL, UCL] of GP model constructed with sequential design, platinum plate temperature example.

Figure 1 illustrates the vastly superior performance of the BMGP model over the stationary GP model when the design is given by $\mathcal{D} = \{0, 200, 300, 500, 800, 1{,}000, 1{,}200\}$. Note that we shall call the stationary GP model, the GP model for simplicity. We also consider a sequential design approach in which the

point with maximum prediction variance within the interval [0, 1200] is added into the design at each iteration. The addition of points is stopped when the width of the 98% prediction interval for the new added point is less than or equal to 200 (prior to updating the model) and the interval contains the observed response. We use the initial design $\mathcal{D}_0 = \{100, 700\}$. Figure 2 plots the design points (circle), posterior mean, and 98% prediction intervals [LCL,UCL] for the BMGP model. The two initial design points are labelled with 0, the first added design point is labelled with 1, and similarly for the other design points. We see that the sequential design procedure is terminated after five points are added, and the BMGP point and prediction interval limits are nearly identical to the true function. In contrast, the GP model terminates only after adding 13 points (Figure 3), and many of the points are in the interval [300, 1,200], which is wasteful because if we have observed that the response is approximately 1,200 at $x = 300$, we would know that it is close to 1,200 for all $x > 300$. Moreover, unlike the BMGP model, the point and interval predictions of the GP model deteriorate outside the experiment region [0, 1200].

4.2. Example 2: Kirchhoff plate bending problem

The bending of a square Kirchhoff plate of length $\mathcal{L}$ subject to a distributed load is described by a pair of Poisson equations (Wang, Reddy and Lee (2000, Chap. 7)):

$$\begin{aligned}
&\triangle\mathcal{M}(\tau_1,\tau_2) = -Q[\tau_1(\mathcal{L}-\tau_1)+\tau_2(\mathcal{L}-\tau_2)], (\tau_1,\tau_2)\in[0,\mathcal{L}]^2,\\
&\quad \mathcal{M}(\tau_1,\tau_2) = 0 \ \ \forall(\tau_1,\tau_2)\in\partial[0,\mathcal{L}]^2 = \{(u,v): u=0\}\cup\{(u,v): u=\mathcal{L}\},\\
&\cup\{(u,v): v=0\}\cup\{(u,v): v=\mathcal{L}\},\\
&\triangle\mathcal{W}(\tau_1,\tau_2) = -\frac{\mathcal{M}(\tau_1,\tau_2)}{\mathcal{F}}, (\tau_1,\tau_2)\in[0,\mathcal{L}]^2, \mathcal{W}(\tau_1,\tau_2) = 0 \ \ \forall(\tau_1,\tau_2)\in\partial[0,\mathcal{L}]^2.
\end{aligned} \tag{4.2}$$

The solution of the first equation is the Marcus moment $\mathcal{M}$ while the solution of the second equation is the vertical displacement of the plate $\mathcal{W}$. Both are functions of the plane coordinates (τ_1, τ_2). In the above equations, $\triangle$ is the Laplace operator, $Q[\tau_1(\mathcal{L}-\tau_1)+\tau_2(\mathcal{L}-\tau_2)]$ is the magnitude of the distributed load, and $\mathcal{F}$ is the flexural rigidity of the plate. The equations in (4.2) are solved with the finite difference scheme given in Li and Chen (2009, pp. 57-59). The vector of inputs is $(x_1, x_2, x_3) = (\mathcal{F}, Q, \mathcal{L})$, the experiment region is $\mathcal{X} = [1\times10^6, 1.4\times10^6]\times[3\times10^5, 7\times10^5]\times[0.7, 2]$, and the response is $y = \max\{|\mathcal{W}(\tau_1,\tau_2)| : (\tau_1,\tau_2)\in[0,\mathcal{L}]^2\}$. The ranges of inputs are chosen realistically.

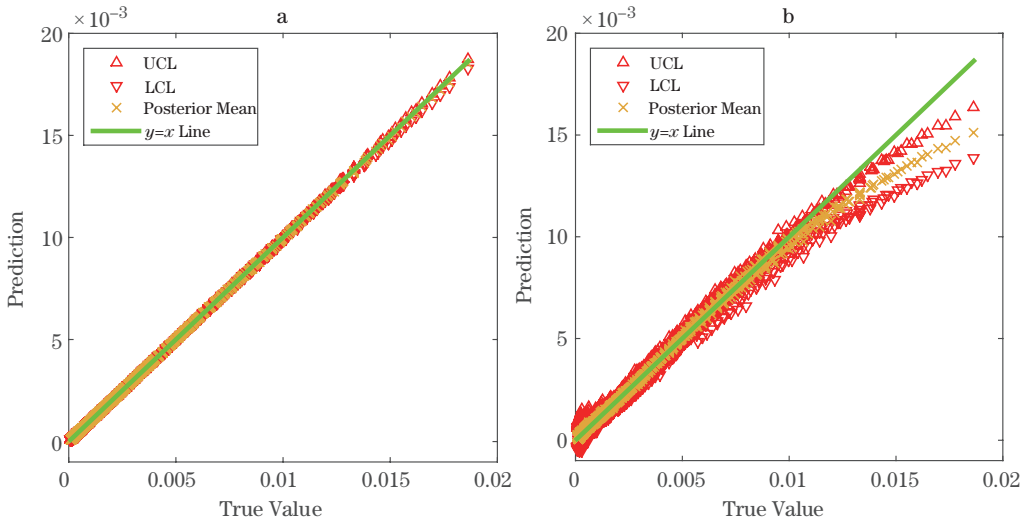

Figure 4. Plot of prediction versus true value in the within-experiment-region grid test set for BMGP (a) and GP (b) models obtained with a 32-point Sobol sequence design, Example 2.

We know that y is small for all $\mathbf{x}$ in the experiment region because the minimum flexural rigidity of 1×10^6 is approximately the flexural rigidity of a titanium plate 4.63 centimeters thick (Young's modulus = 110GPa, Poisson ratio = 0.3). Based on physical considerations, we know that y approaches zero as the flexural rigidity gets larger, or the load gets smaller, or the plate length goes to zero:

$$\lim_{x_1 \to \infty} [f(\mathbf{x})] = 0, \lim_{x_2 \to 0} [f(\mathbf{x})] = 0, \lim_{x_3 \to 0} [f(\mathbf{x})] = 0. \tag{4.3}$$

We use the first 32 points of the Sobol sequence as the design, and fit the BMGP and GP models. We evaluate the performance of the models using a regular 9^3 grid in the experiment region. The grid is constructed by taking Cartesian product of equally spaced levels in each input, where the levels include the minimum and maximum levels. In Figure 4a, we plot the 0.99 quantile (UCL), 0.01 quantile (LCL), and the posterior mean versus the true response values in the grid test set for the BMGP model. Figure 4b is a similar plot for the GP model. We see that the BMGP model gives very accurate predictions and narrow prediction intervals. In contrast, the GP model appears to suffer from bias in the prediction of large response values. The mean absolute error (MAE), average length of 98% prediction intervals, and coverage of the prediction intervals for the BMGP and GP models are given in the first row below the header row in Table 1. We see that the BMGP model gives more accurate predictions, far shorter

Table 1. MAE, average prediction interval length, and coverage for BMGP and GP models constructed with Sobol sequence and sequential design for four test sets, Kirchhoff plate example.

Test Set	Design	BMGP			GP		
		Mean Absolute Error (MAE) $\times 10^5$	Average Prediction Interval Length $\times 10^5$	Coverage	Mean Absolute Error (MAE) $\times 10^5$	Average Prediction Interval Length $\times 10^5$	Coverage
9^3 regular grid on $\mathcal{X}$	32 Sobol points	1.45	6.36	0.949	16.82	72.77	0.911
	Sequential	0.73	4.08	0.957	7.56	77.58	0.992
9^3 regular grid on $\mathcal{X}_1$	32 Sobol points	1.50	12.11	0.966	25.93	300.42	0.995
	Sequential	1.05	13.48	0.985	12.56	396.92	1.000
9^3 regular grid on $\mathcal{X}_2$	32 Sobol points	17.37	7.69	0.379	29.86	353.68	0.999
	Sequential	20.39	9.74	0.429	26.36	432.32	0.999
9^3 regular grid on $\mathcal{X}_3$	32 Sobol points	3.98	53.61	1.000	33.54	616.75	0.999
	Sequential	5.31	56.00	0.999	31.38	811.91	0.999

prediction intervals, and better coverage than the GP model.

We evaluate a sequential design approach that chooses the point within the experiment region with largest prediction variance as the next design point in each iteration. The initial design is the first eight points of the Sobol sequence whereas the final design is of size 32. The performance of the BMGP and GP models obtained with the sequential design approach is presented in the second row of Table 1. We see that the BMGP model gives better accuracy and shorter interval length but poorer coverage than the GP model. Moreover, it can be seen that the sequential design is an improvement over the Sobol sequence. Figure 5 plots the projections of the sequential design for the BMGP model onto the (x_1, x_2) plane (a) and the (x_1, x_3) plane (c). The eight initial design points are plotted with the symbol 0, the first added point is plotted with the symbol 1 and so on. If we divide the bounded square design region for (x_1, x_2) into four quadrants, many of the added design points concentrate in the upper left quadrant. Similarly, we see that many of the added design points fall in the upper left quadrant of the design region for (x_1, x_3). This phenomenon is due to the fact that (4.3) provides prior information about the behavior of the response y when x_1 gets large, and when x_2 or x_3 gets small. Since there is prior information that y is small when x_1 is large, many added design points should have a value of x_1 closer to the minimum level. In contrast, the sequential design for the GP model (Figure 5b and 5d) distributes points quite uniformly in the four quadrants and the points tend to be near the boundary of the experiment region.

We also investigate the extrapolation performance of the BMGP and GP models. We are interested in the performance of the BMGP model for values of x_1 larger than its maximum in the experiment region and values of x_2 and x_3 smaller than their minimums in the experiment region because (4.3) contains information that helps extrapolation in these cases. To investigate extrapolation performance for large values of x_1, we use a 9^3 regular grid on $\mathcal{X}_1 = [1.4 \times 10^6, 1.8 \times 10^6] \times [3 \times 10^5, 7 \times 10^5] \times [0.7, 2]$ as test set. Thus, the range of x_1 is changed from $[1 \times 10^6, 1.4 \times 10^6]$ to $[1.4 \times 10^6, 1.8 \times 10^6]$. To investigate extrapolation performance for small values of x_2, we use a 9^3 regular grid on $\mathcal{X}_2 = [1 \times 10^6, 1.4 \times 10^6] \times [1 \times 10^5, 3 \times 10^5] \times [0.7, 2]$. Finally, to investigate extrapolation performance for small values of x_3, we use a 9^3 regular grid on $\mathcal{X}_3 = [1 \times 10^6, 1.4 \times 10^6] \times [3 \times 10^5, 7 \times 10^5] \times [0.1, 0.7]$. The MAE, average 98% prediction interval length, and coverage of the BMGP and GP models constructed with the first 32 points of the Sobol sequence, and the sequential design described above are given in Table 1. The GP model gives intervals that are very wide, which implies that it is rather noninformative about y, and poor prediction accuracy for the third test set $\mathcal{X}_3$ (the ranges of y in the test sets are $1325 \times 10^{-5}, 796 \times 10^{-5}$, and 28×10^{-5} respectively). The BMGP model gives good prediction accuracy, narrow intervals, and good coverage for large x_1. For the small x_3 test set, the accuracy and coverage are good but the average interval width is larger than the range of the response. The prediction accuracy for the small x_2 test set is good but the coverage of the prediction intervals is poor.

For the BMGP model constructed with the Sobol sequence and sequential design, the leave-one-out crossed validated MAE (0.5×10^{-5} and 1.1×10^{-5} respectively) and coverage (1 for both designs) are excellent. This indicates, without the need for an independent test set, good prediction performance in the experiment region. As shown in Table 1, extrapolation performance may not be as good. It seems that the only way to validate the model's extrapolation capability is to use a test set in the extrapolation region.

In Table 2, we give results for three alternatives to the proposed BMGP model. First, we allow the α_i's in (3.4) to be different. Second, we allow the η_i's in (3.6) to be different. Third, we use the alternative variance function in (3.9) with $\eta_1 = \cdots = \eta_k = \eta$ and $G(z; \eta) = [1 - (1 - z)^\eta]^\eta$. Performance of the three models evaluated on the regular 9^3 grid in the experiment region $\mathcal{X}$ is given in Table 2. We see that it is advantageous to allow the η_i's in (3.6) to be different. The third model (which has a different variance function) performs worst. The first model and the original BMGP model perform similarly.

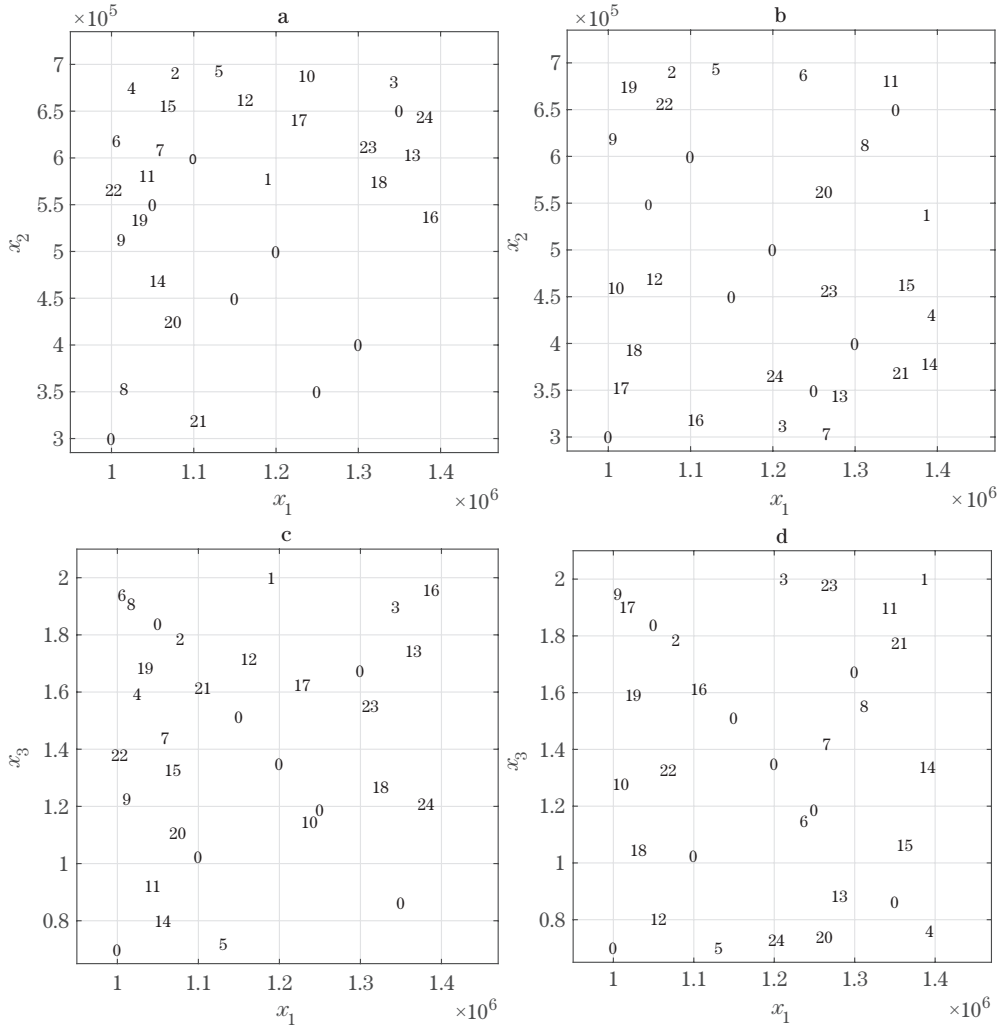

Figure 5. Two dimensional projections of sequential design for the BMGP model (a) and (c), and the GP model (b) and (d), Example 2. The initial design points are marked with 0, and the design point added at the ith iteration is marked with i.

4.3. Example 3: Two-dimensional heat transfer

In this example, we consider a two-dimensional heat transfer model of a solid object with cross section Ω shown in Figure 6 (Alberty, Carstensen and Funken (1999)). The boundary of the object consists of five smooth edges. Three of the edges are labelled Dirichlet edge ($\partial\Omega_1$), and two are labelled Neuman edge ($\partial\Omega_2$). The temperature $T(\tau_0, \tau_1, \tau_2)$ at coordinates (τ_1, τ_2) of the cross section at time τ_0 is the solution of a parabolic PDE given by

Table 2. MAE, average prediction interval length, and coverage for within-experiment-region test set for alternative BMGP models constructed with the first 32 Sobol points, Example 2.

	BMGP		
	Mean Absolute Error (MAE) $\times 10^5$	Average Prediction Interval Length $\times 10^5$	Coverage
Different α_i's	1.46	6.24	0.959
Different η_i's	1.33	5.76	0.975
Different variance function	2.04	6.50	0.888

$$\frac{\partial}{\partial \tau_0} T(\tau_0, \tau_1, \tau_2) = \frac{\partial^2}{\partial \tau_1^2} T(\tau_0, \tau_1, \tau_2) + \frac{\partial^2}{\partial \tau_2^2} T(\tau_0, \tau_1, \tau_2) + F, (\tau_1, \tau_2) \in \Omega,$$
$$T(0, \tau_1, \tau_2) = T_0 \quad \forall (\tau_1, \tau_2) \in \Omega, T(\tau_0, \tau_1, \tau_2) = T_b \forall (\tau_1, \tau_2) \in \partial\Omega_1, \tau_0 > 0,$$
$$\frac{\partial}{\partial N} T(\tau_0, \tau_1, \tau_2) = 0 \quad \forall (\tau_1, \tau_2) \in \partial\Omega_2, \tau_0 > 0. \tag{4.4}$$

In the above equation, T_0 is the initial temperature; T_b is the fixed temperature at the Dirichlet edges; $\partial/(\partial N) T(\tau_0, \tau_1, \tau_2)$ denotes the derivative of $T(\tau_0, \tau_1, \tau_2)$ in the direction normal to the boundary and pointing away from Ω; and F is the internal heat generated per unit volume per unit time. To solve (4.4), we use the finite element Matlab code described in Alberty, Carstensen and Funken (1999). The triangular mesh used is plotted in Figure 6. Note that the code returns $T(\tau_0, \tau_1, \tau_2)$ for all $(\tau_1, \tau_2) \in \Omega$ given fixed τ_0. The computation involves a finite difference discretization with respect to time.

The variable inputs to the computer model are $(x_1, x_2, x_3, x_4) = (\tau_0, T_0, F, T_b)$ and the output of interest is $y = \max\{T(\tau_0, \tau_1, \tau_2) : (\tau_1, \tau_2) \in \Omega\}$, which can be obtained by simply taking the maximum of the temperature values at the vertices of the triangulation. The experiment region is $\mathcal{X} = [0.1, 2.1] \times$ [650, 1,050] $\times [0, 200] \times$ $[250, 450]$. Based on physical considerations, we know that when internal heat generation $x_3 = F = 0$, the temperature converges to the boundary temperature $x_4 = T_b$, which is the steady state temperature, when $x_1 = \tau_0$ gets large. Although the temperature profile $T(\tau_0, \tau_1, \tau_2)$ at time $x_1 = 0$ is $x_2 = T_0$ for all $(\tau_1, \tau_2) \in \Omega$, it will immediately change so that it is continuous as a function of (τ_1, τ_2) with value x_4 at the Dirichlet boundary for any time $x_1 > 0$ (Mattheij, Rienstra and ten Thije Boonkkamp (2005, p. 244)). For x_1 infinitesimally close to zero, there must be values $T(\tau_0, \tau_1, \tau_2)$ for $(\tau_1, \tau_2) \in \Omega$ that is infinitesimally close to x_2. Because heat generated over infinitesimal time cannot raise the temperature of the object, it follows that $T(\tau_0, \tau_1, \tau_2) \leq \max\{x_2, x_4\}$ as $x_1 \to 0$.

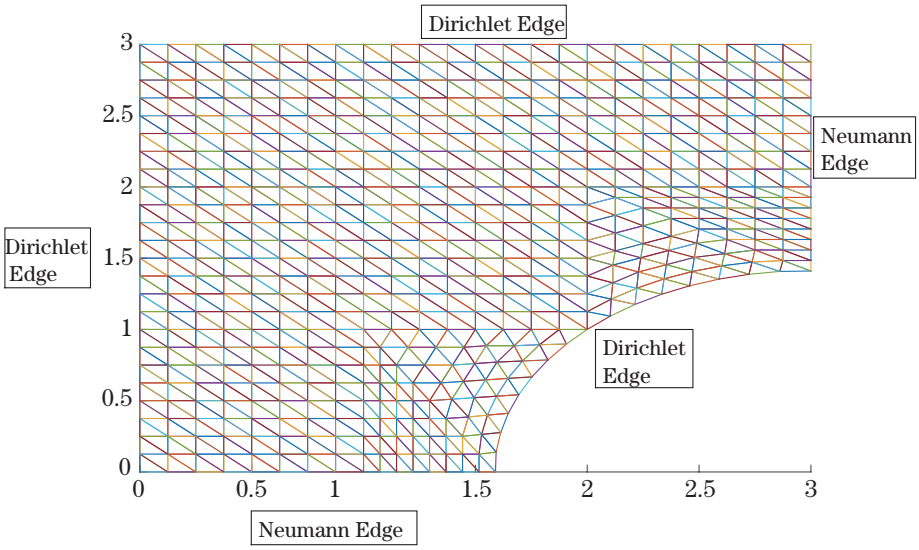

Figure 6. Triangular mesh of cross section geometry, Example 3.

Thus, we have

$$\lim_{(x_1,x_3)\to(\infty,0)}[f(\mathbf{x})-x_4]=0,\ \lim_{x_1\to 0}[f(\mathbf{x})-\max\{x_2,x_4\}]=0. \tag{4.5}$$

We consider using two designs to construct the BMGP and GP models. The first design consists of the first 40 points of the Sobol sequence. The second design is constructed by using the first 8 points of the Sobol sequence as initial design, and adding one point at a time, where the added point has maximum prediction variance. The sequential addition of points is terminated when the design size is 40. Two test sets are used, a 6^4 regular grid in the experiment region $\mathcal{X}$, and a 6^4 regular grid on $\mathcal{X}_1 = [2.1, 3.1]\times$ [650, 1,050] $\times[0, 200]\times[250, 450]$. The second test set is used to test the extrapolation performance of the BMGP and GP models near steady state (large x_1). For the BMGP model constructed with the Sobol sequence and sequential design, the leave-one-out crossed validated MAE (3.59 and 5.96 respectively) and coverage (0.975 for both designs) are excellent. This indicates good prediction performance within the experiment region (similarly for the GP model).

Table 3 gives the MAE, average 98% prediction interval length, and coverage evaluated with the two test sets for the BMGP model and GP model constructed using the two designs described above. We see that the BMGP model provides better prediction accuracy and shorter prediction intervals than the GP model within the experiment region. However, it gives slightly poorer coverage than the

Table 3. MAE, average prediction interval length, and coverage of BMGP and GP models constructed with Sobol sequence and sequential design for two test sets, two-dimensional heat transfer example.

Test Set	Design	BMGP			GP		
		Mean Absolute Error (MAE)	Average Prediction Interval Length	Coverage	Mean Absolute Error (MAE)	Average Prediction Interval Length	Coverage
6^4 regular grid on $\mathcal{X}$	40 Sobol points	4.65	24.62	0.864	7.95	46.94	0.919
	Sequential	2.75	26.52	0.995	3.40	42.43	0.996
6^4 regular grid on $\mathcal{X}_1$	40 Sobol points	16.35	211.50	1.000	28.95	561.76	1.000
	Sequential	19.13	318.32	0.996	28.10	601.00	0.998

Table 4. MAE, average prediction interval length, and coverage for within-experiment-region test set for alternative BMGP models constructed with the first 40 Sobol points, Example 3.

	BMGP		
	Mean Absolute Error (MAE)	Average Prediction Interval Length	Coverage
Different α_i's	3.80	20.66	0.937
Different η_i's	4.32	22.65	0.868
Different variance function	4.67	24.98	0.863

GP model when the design used is the first 40 points of the Sobol sequence. The BMGP model has better extrapolation capability for large x_1 than the GP model because it gives more accurate predictions and shorter prediction intervals. The GP model gives noninformative extrapolation because the range of the response in the extrapolation test set is only 310, which is smaller than the average length of the prediction intervals given by the GP model. The BMGP model constructed with the sequential design gives somewhat noninformative extrapolations also.

In Table 4, we give results for three alternatives to the proposed BMGP model. First, we allow the α_i's in (3.4) to be different. Second, we allow the η_i's in (3.6) to be different. Third, we use the alternative variance function in (3.9) with $\eta_1 = \cdots = \eta_k = \eta$ and $G(z;\eta) = [1-(1-z)^{\eta}]^{\eta}$. Performance of the three models evaluated on the regular 6^4 grid in the experiment region is given in Table 4. We see that it is advantageous to allow the α_i's in (3.4) to be different. The third model (which has a different variance function) performs worst but it has performance similar to the original BMGP model.

A referee raised the question of the performance of the BMGP and GP models when Latin hypercube designs are used. We randomly simulate 100

maximin Latin hypercube designs, where each design is the best out of 5,000 Latin hypercube designs with respect to the maximin criterion. The BMGP and GP models are fitted with each of the 100 designs. We find that the mean and standard deviation of the MAE are 5.85 and 2.22 for the BMGP model, and 8.73 and 2.00 for the GP model. The mean and standard deviation of the average prediction interval length are 24.12 and 5.40 for the BMGP model, and 48.06 and 5.35 for the GP model. Finally, the mean and standard deviation of the coverage are 0.86 and 0.06 for the BMGP model, and 0.93 and 0.04 for the GP model.

Results for the model with different α_i's in this example are obtained with the restriction that $\alpha_1 + \alpha_2 \leq 200$ in optimization of the likelihood function, which restricts the weight given to the constant a^0 to be not too small. Without the restriction, the estimates for α_1 and α_2 would be very large ($> 10^9$), which would give very small weight to a^0 in (3.5), and the estimates for a^0 and s^2 would be of the order of 10^{11} and 10^{48} respectively. This clearly does not make sense. It is due to a flat likelihood and numerical errors. The MAE within the experiment region for this model is 9.63, which is somewhat larger than the MAE for the GP model. Finally, when using randomly generated maximin Latin hypercube designs to fit the BMGP model, we have found instances in which the MAE and average prediction interval length are very large ($> 10^9$) but these instances are all associated with extremely large estimates of s^2 ($> 10^{100}$) and η (> 100), and extremely small estimates of δ ($< 10^{-34}$). Again, these estimates do not make sense and is due to a flat likelihood and numerical errors. Thus, we rectified the problem by restricting η to be no greater than seven, which does not allow a steep increase in variance in (3.6) when the distance to a boundary is near one.

5. Conclusions

In this paper, we propose a modification of the stationary GP model, called BMGP, to take into account knowledge about the response of the form given in (1.1), which in practice is often information about the response behavior on an edge of the boundary of the input space. We propose a flexible prior mean that satisfies (1.1) based on the inverse distance interpolator. A prior variance function that converges to a small value as the input values get closer to the edge given in (1.1) is employed. Three real examples are given to illustrate that the BMGP model improves prediction accuracy and gives shorter prediction intervals for predicting within the experiment region. As we pointed out, the BMGP model includes the stationary GP model as a special case. Thus, it is expected

to perform better than the GP model unless its parameters are estimated poorly. In addition, the examples also show that the BMGP model has improved extrapolation performance if the extrapolation is performed in a region closer than the experiment region to the edge given in (1.1). However, extrapolation performance of the BMGP model is not always superior to the GP model although the BMGP model will produce accurate predictions at locations sufficiently close to the boundary in (1.1). This is to be expected since there is no available information about the behavior of the function "between" the experiment region and the boundary in (1.1).

Several problems need further research. First, alternative choices of mean and variance functions should be studied. Second, in the examples, the BMGP model is used to model responses that are scalar summaries of the solution of PDEs. However, it may also be used to model the entire solution, which is a function of space and time that satisfies some boundary conditions. The so called Dirichlet boundary conditions are precisely information of the form (1.1). In this case, the problem is to predict entire functions given data that are functions, as in Hung, Joseph and Melkote (2013). The author is currently working on this problem.

Acknowledgment

This research was supported by Early Career Scheme (ECS) project 21201414 funded by the Research Grants Council of Hong Kong and Strategic Research Grant 7004522 funded by City University of Hong Kong. The author thanks two referees and an associate editor for comments that helped improve the paper.

Appendix

Proof of Theorem 1: Since $d_j^2(p^j(\mathbf{x}), \mathbf{c}^j)$ is bounded and $\alpha_j, j = 1, \ldots, k$ are positive constants, $\lambda_0(\mathbf{x}) \to 0$ for any sequence of $\{\mathbf{x}(g) : g = 1, 2, \ldots\}$ such that $\mathbf{x}^i(g) \to \mathbf{c}^i$. Moreover, due to the assumptions on $d_j^2(p^j(\mathbf{x}), \mathbf{c}^j)$ and the fact that the $\mathbf{c}^j$'s are distinct, there exists g large enough so that $d_j^2(\mathbf{x}^j(g), \mathbf{c}^j) > \epsilon, j \neq i$ for any fixed ϵ small enough. Because $d_i^2(p^i(\mathbf{x}), \mathbf{c}^i)$ is a continuous function of $\mathbf{x}$, we have $d_i^2(\mathbf{x}^i(g), \mathbf{c}^i) \to 0$. Thus, $\lambda_i(\mathbf{x}(g)) \to 1$ and for all $j \neq i$, $\lambda_j(\mathbf{x}(g)) \to 0$.

Proof of Corollary 1: We have $\mu(\mathbf{x}) - a^i(\mathbf{x}) = [\lambda_i(\mathbf{x}) - 1]a^i(\mathbf{x}) + \sum_{m=0, m\neq i}^{k} a^m(\mathbf{x})\lambda_m(\mathbf{x})$. Because $\limsup_{g\to\infty}|a^j(\mathbf{x}(g))| < \infty, j = 1, \ldots, k$, there exists $B > 0$ and a g_1 such that $|a^j(\mathbf{x}(g))| < B, j = 0, \ldots, k$ for all $g \geq g_1$. By Theorem 1, we have $\lambda_i(\mathbf{x}(g)) - 1 \to 0$ and for all $j \neq i$, $\lambda_j(\mathbf{x}(g)) \to 0$. Thus, we

can find a g_2 such that for all $g \geq g_2$, we have $|\lambda_i(\mathbf{x}(g)) - 1| < \epsilon/[(k+1)B]$ and $|\lambda_j(\mathbf{x}(g))| < \epsilon/[(k+1)B]$ for all $j \neq i$. This implies that $|\mu(\mathbf{x}(g)) - a^i(\mathbf{x}(g))| < \epsilon$ for all $g \geq \max\{g_1, g_2\}$.

References

Adler, R. J. (2010). *The Geometry of Random Fields*. SIAM, Philadelphia.

Alberty, J., Carstensen, C. and Funken, S. A. (1999). Remarks around 50 lines of Matlab: short finite element implementation. *Numerical Algorithms* **20**, 117-137.

Bastos, L. S. and O'Hagan, A. (2009). Diagnostics for Gaussian process emulators. *Technometrics* **51**, 425-438.

Cengel, Y. A. and Ghajar, A. (2011). *Heat and Mass Transfer: Fundamentals and Applications* (4th Edition). McGraw-Hill, New York.

Coleman, M. P. (2013). *An Introduction to Partial Differential Equations with Matlab* (2nd Edition). CRC Press, Boca Raton.

Currin, C., Mitchell, T., Morris, M. 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.

Farlow, S. J. (1982). *Partial Differential Equations for Scientists and Engineers*. Wiley, New York.

Giblin, P. J. (1972). What is an asymptote? *The Mathematical Gazette* **56**, 274-284.

Gockenbach, M. S. (2011). *Partial Differential Equations: Analytical and Numerical Methods* (2nd Edition). Siam, Philadelphia.

Golchi, S., Bingham, D. R., Chipman, H. and Campbell, D. A. (2015). Monotone emulation of computer experiments, *SIAM/ASA Journal on Uncertainty Quantification* **3**, 370-392.

Gordon, W. J. and Wixom, J. A. (1978). Shepard's method of "metric interpolation" to bivariate and multivariate interpolation, *Mathematics of Computation* **32**, 253-264.

Haaser, N. B. and Sullivan, J. A. (1991). *Real Analysis*. Mineola: Dover Publications.

Humphrey, J. D. and DeLange, S. (2013). *An Introduction to Biomechanics: Solids and Fluids, Analysis and Design*. Springer Science & Business Media, New York.

Hung, Y., Joseph, V. R. and Melkote, S. (2013). Analysis of computer experiments with functional response, *Technometrics* **57**, 35-44.

Johnson, D. (2000). *Advanced Structural Mechanics: An Introduction to Continuum Mechanics and Structural Mechanics*. Thomas Telford, London.

Li, J. and Chen, Y. T. (2009). *Computational Partial Differential Equations Using MATLAB*. CRC Press, Boca Raton.

Mattheij, R. M., Rienstra, S. W. and ten Thije Boonkkamp, J. H. (2005). *Partial Differential Equations: Modeling, Analysis, Computation*. Siam, Philadelphia.

Mitchell, T. J. and Morris, M. D. (1992). Bayesian design and analysis of computer experiments: two examples, *Statistica Sinica* **2**, 359-379.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer Science & Business Media, New York.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer

experiments, *Statistical Science* **4**, 409-435.

Santner, T. J., Williams, B. J. and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York.

Tan, M. H. Y. (2015). Monotonic metamodels for deterministic computer experiments. *Technometrics* **59**, 1-10.

Wang, C. M., Reddy, J. N. and Lee, K. H. (2000). *Shear Deformable Beams and Plates: Relationships with Classical Solutions*. Elsevier Science, Oxford.

Wang, X. (2012). *Bayesian Modeling Using Latent Structures*. Doctoral dissertation, Department of Statistical Science, Duke University.

Xiu, D. and Karniadakis, G. E. (2003). A new stochastic approach to transient heat conduction modeling with uncertainty, *International Journal of Heat and Mass Transfer* **46**, 4681-4693.

Department of Systems Engineering and Engineering Management, City University of Hong Kong, Kowloon, Hong Kong.

E-mail: matthtan@cityu.edu.hk

(Received July 2015; accepted January 2016)