Indicator-based Bayesian Variable Selection for Gaussian Process Models in Computer Experiments

Fan Zhang
School of Mathematical and Statistical Science, Arizona State University, Arizona, USA
Ray-Bing Chen
Department of Statistics, National Cheng Kung University, Tainan, Taiwan
Institute of Data Science, National Cheng Kung University, Tainan, Taiwan
Ying Hung
Department of Statistics and Biostatistics, Rutgers University, New Jersey, USA
Xinwei Deng
Department of Statistics, Virginia Tech, Virginia, USA

Abstract: Gaussian process (GP) models are commonly used in the analysis of computer experiments. Variable selection in GP models is of significant scientific interest but existing solutions remain unsatisfactory. For each variable in a GP model, there are two potential effects with different implications: one is on the mean function, and the other is on the covariance function. However, most of the existing research on variable selection for GP models has focused only on one of the effects. To tackle this problem, we propose an indicator-based Bayesian variable selection procedure to take into account the effects from both the mean and covariance functions. A variable is defined to be inactive if both effects are not significant, and an indicator is used to represent the variable being active or not. For active variables, the proposed method adopts different prior assumptions to capture the two effects. The performance of the proposed method is evaluated by both simulations and real applications in computer experiments.

Keywords: Bayesian Variable Selection; Emulator, Kriging, Median Probability Criterion.

1. Introduction

Physical experiments are often expensive, time-consuming, and dangerous to perform, especially for the study of complex systems. An effective and more efficient alternative is computer experiment, which refers to the study of real systems using complex mathematical models. However, computer experiments typically require a great deal of computing time to produce simulation results, especially for complex problems. Therefore, it is desirable to build a statistical model as an emulator for the actual computer experiments for prediction, optimization, and calibration. The construction of emulators for the study of computer

experiments has received great attention in the past decades (Sacks et al., 1989; Santner et al., 2003). Gaussian process (GP) models, also called kriging models, are widely used to construct the emulators due to their flexibility in capturing the underlying nonlinearity and quantifying the prediction uncertainty. Their interpolation property is also suitable for the study of deterministic computer simulations. Examples of different modifications and applications of GP can be found in Joseph (2006), Gramacy and Lee (2008), Levy and Steinberg (2010), Reich et al. (2009), Plumlee and Joseph (2018), Chen et al. (2018), etc.

An important issue in GP modeling is to identify variables with significant impacts on the simulation responses. For complex systems, there are usually a large number of variables involved in computer experiments. These variables can have very different impacts on the responses. Correct identification of significant variables not only provides scientific insights into the underlying systems but also improves the prediction accuracy of the emulator (Joseph et al., 2008). Therefore, the focus of this research is to achieve simultaneous estimation and variable selection for GP models.

In general, a GP model contains two parts: a mean function $\mu(\mathbf{x})$ and a Gaussian process $Z(\mathbf{x})$. The input variables $\mathbf{x}$ will affect both the mean function and the Gaussian process. The mean function captures the global trend through the unknown coefficients, and the Gaussian process captures the local structure through the correlation parameters. For each variable in a GP model, there are potential effects on the two parts of the GP models with different implications. Thus it is crucial to identify the important variables by simultaneously considering both of the possible effects. It is worth noting that the ordinary kriging is a popular GP model with a grand mean. It generally works well, but it is known that failing to account for important variables in the mean function can cause poor performance in prediction (Joseph et al., 2008). Furthermore, including unimportant variables in the mean function can also deteriorate the prediction performance.

Most of the existing works on variable selections for GP models often focus only on one of the effects. For example, Welch et al. (1992) introduces variable screening methods to identify important correlation parameters sequentially. The idea of sequential variable selection in the correlation function was also used for the high-dimensional Gaussian process (Chen et al., 2012). Linkletter et al. (2006) propose a Bayesian procedure for the selection of significant correlation parameters. These methods allow the variables to have different impacts on the smoothness of the underlying system but overlook the potential impacts on the global trend. On the other hand, Joseph et al. (2008), Hung (2011) and Huang et al. (2020) propose modifications of GP models to perform variable selection only through the mean function coefficients in GP models. When the sample size is huge, Zhao et al. (2018) proposed subsample aggregating (subagging) approach to deal with the variable selection in the mean function of GP models.

To conduct variable selection for GP models based on both effects, this work proposes a unified Bayesian variable selection procedure, which is different from the conventional variable selection methods in GP modeling where selections are performed in either the mean or the correlation function. Note that when a variable is called active, it may not be necessary to affect in both the mean function and the correlation function. For example, in our real data case study shown in Section 6, we have found the variable "zone-to-zone transition" has a significant effect on the mean function but not on the part of the Gaussian process. Thus for each variable, an indicator is defined to represent the variable being active or not. That is, the proposed method simultaneously considers the potential impacts on the mean and correlation function from each variable. For active variables, their impacts

are further distinguished by two Bayesian priors. A variable is inactive only if both effects are not significant. Furthermore, an active variable can be active due to the contributions to the mean, the correlation function, or both. The major motivation is to enhance the interpretability of GP model by disentangling the two impacts from each variable, one on mean and the other on the correlation function, under a hierarchical Bayesian structure. Finally, the proposed method can be modified as a two-indicator approach for detecting the activities of the effects in the mean function and the correlation parameter separately.

The remaining of this paper is organized as follows. Section 2 briefly reviews the Gaussian process model. Section 3 details the proposed method. Simulation studies are conducted to examine the proposed method in Sections 4 and 5. Section 6 contains an illustration using real data. In Section 7, the proposed method is extended to a two-indicator approach for the mean function and the correlation parameter separately. We conclude this work with some discussion in Section 8.

2. Gaussian Process Model

This section gives a brief review on the Gaussian process model. Denote $\mathbf{x} = (x_1, ..., x_p)$ as a p-dimensional input and $y(\mathbf{x}) \in \mathcal{R}$ as the response. Suppose the observed data is denoted by $\{(\mathbf{x}_i, y_i), i = 1, ..., n\}$. A Gaussian process model can be written as

$$y(\mathbf{x}) = \mu(\mathbf{x}) + Z(\mathbf{x}), \tag{2.1}$$

where $\mu(\mathbf{x})$ is a mean function and $Z(\mathbf{x})$ is a Gaussian process with zero mean and covariance function $\phi(\mathbf{x})$. Here $\mu(\mathbf{x})$ is expressed as

$$\mu(\mathbf{x}) = \sum_{k=1}^{p} \beta_k x_k = \mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta},$$

where $f(\boldsymbol{x}) = (x_1, \ldots, x_p)^\top$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ is a vector of unknown coefficients. There are various choices of covariance functions (Rasmussen, 2003) such as the Matern function and the powered exponential function. The powered exponential function is expressed as

$$\phi(Z(\mathbf{x}_i), Z(\mathbf{x}_j)) = \sigma^2 \exp(-\sum_{k=1}^{p} \theta_k |x_{ik} - x_{jk}|^\kappa),$$

where θ_k is the correlation parameter, $\boldsymbol{\theta} = (\theta_1, \ldots \theta_p)^\top$, σ^2 is the variance, and $0 < \kappa \leq 2$ controls the underlying smoothness.

The unknown parameters, $\boldsymbol{\beta}, \boldsymbol{\theta}$, and σ^2, can be estimated by the maximum likelihood approach with

$$\hat{\boldsymbol{\beta}} = (F\Phi^{-1}F^\top)^{-1}F\Phi^{-1}\mathbf{y}, \quad \hat{\sigma}^2 = \frac{(\mathbf{y} - F\hat{\beta})^\top \Phi^{-1}(\mathbf{y} - F\hat{\beta})}{n}, \tag{2.2}$$

and

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta_i > 0, \forall i} \left\{ n \log(\widehat{\sigma}^2) + \log(|\Phi|) \right\},$$

where $F = [f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n)]^\top$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ with $y_i = y(\mathbf{x}_i)$, and $\mathbf{\Phi}$ is an $n \times n$ correlation matrix with elements $\phi(Z(\mathbf{x}_i), Z(\mathbf{x}_j))$. Plugging in the estimators, prediction at $\mathbf{x}$ can be obtained by

$$\hat{y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^\top \hat{\beta} + \boldsymbol{\psi}(\boldsymbol{x})^\top \mathbf{\Phi}^{-1}(\mathbf{y} - \mathbf{f}(\mathbf{x})^\top \hat{\beta}), \tag{2.3}$$

where $\boldsymbol{\psi} = (\phi(Z(\mathbf{x_1}), Z(\mathbf{x})), \ldots, \phi(Z(\mathbf{x_n}), Z(\mathbf{x})))^\top$.

Based on (2.1), it is clear that each variable has two potential impacts, one is the linear effects through the mean function $\boldsymbol{\beta}$, and the other is the effects on smoothness through correlation parameters $\boldsymbol{\theta}$. Therefore, in performing variable selection, it is important to clearly identify their effects.

3. Indicator-based Bayesian Variable Selection

In this section, we develop an indicator-based Bayesian variable selection for Gaussian process models. The idea is to introduce a latent indicator for each input variable in the Gaussian process model to represent whether the variable is active or not. For active variables, different priors are assumed for the mean function coefficients and correlation parameters. Moreover, to enhance the computational efficiency, we adopt some techniques from the empirical Bayes (Yuan and Lin, 2005) to obtain a meaningful approximation of the corresponding posterior density by setting proper priors of the indicators and the unknown parameters.

3.1 Priors and Posteriors

Denote γ_k as the latent indicator for the kth variable, with one indicating active and zero otherwise. The kth variable is inactive if $\beta_k = 0$ and $\theta_k = 0$. The prior distributions of β_k and θ_k are specified as mixture distributions dependent on γ_k. Using this hierarchical Bayes formulation, γ_k, β_k and θ_k are well associated with each other. Then a numerical algorithm can be used to generate the samples of γ_k, and these samples can be used to infer which variables are active.

Let us start from setting priors for the indicators. For notation convenience, denote $\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}$ as the corresponding quantities $\boldsymbol{\beta}, \boldsymbol{\theta}$ under $\boldsymbol{\gamma}$. Let $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)'$ to be the vector of indicator parameters. For the prior of γ_k, because of two status of each γ_k, we consider the commonly used Bernoulli prior, $Bern(q)$, where q is the probability of $\gamma_k = 1$. By assuming the independence among γ_k's, the prior of $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_p)$ can be written as

$$P(\boldsymbol{\gamma}) \propto q^{|\boldsymbol{\gamma}|}(1-q)^{p-|\boldsymbol{\gamma}|},$$

where $|\boldsymbol{\gamma}| = \sum_{k=1}^{p} \gamma_k$ and q is the prior probability of $\gamma_k = 1$.

Now we specify the priors for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Straightforwardly, we set $\beta_k = 0$ if $\gamma_k = 0$. When $\gamma_k = 1$, the double exponential distribution is chosen as the prior distribution for β_k. Thus, the prior of β_k is

$$\pi(\beta_k|\gamma_k) = (1-\gamma_k)\delta(0) + \gamma_k DE(0, \tau_k),$$

where $DE(0, \tau_k)$ is the double exponential distribution and it has a density function as $(1/2)\tau_k \exp(-\tau_k|\beta_k|)$ with the positive parameter, τ_k. Note that the double-exponential (DE)

prior can be accommodated for large coefficients because of its heavier tail property (Casella and Park, 2008). For θ_k, we set it as

$$\pi(\theta_k|\gamma_k) = (1-\gamma_k)\delta(0) + \gamma_k Exp(\lambda_k),$$

where $Exp(\lambda_k)$ is an exponential distribution with density function, $\lambda_k \exp(-\lambda_k\theta_k)$ and λ_k is the positive hyper-parameter. In addition, we assume the independence among the priors of β_k and θ_k. To simplify the technique presentation, we assume that the hyper-parameters $\tau_k = \tau$ and $\lambda_k = \lambda$ for all k. Finally, we set the prior for σ^2 to be an inverse χ-squared distribution $Inv-\chi^2(\nu_0)$, i.e., $\sigma^2 = (\sigma^2)^{-\nu_0/2-1}\exp(-1/(2\sigma^2))$. Based on above formulation, we can write the posterior density, $P(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2|\mathbf{y})$ as

$$\begin{aligned}
&P(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2|\mathbf{y}) \\
&\propto \exp(-\frac{1}{2}[n\log\sigma^2 + \log|\Phi(\boldsymbol{\theta})| + \frac{(\mathbf{y} - \boldsymbol{F}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^\top \Phi^{-1}(\boldsymbol{\theta})(\mathbf{y} - \boldsymbol{F}_{\gamma}\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{\sigma^2}]) \\
&\times \prod_{k=1}^{p}[\frac{\tau}{2}\exp(-\tau|\beta_k|)]^{\gamma_k} \times \prod_{k=1}^{p}[\lambda\exp(-\lambda|\theta_k|)]^{\gamma_k} \\
&\times q^{|\boldsymbol{\gamma}|}(1-q)^{p-|\boldsymbol{\gamma}|} \times (\sigma^2)^{-\nu_0/2-1}\exp(-1/(2\sigma^2)) \\
&\propto \exp(-\frac{1}{2}[\log|\Phi(\boldsymbol{\theta})| + \frac{(\mathbf{y} - \boldsymbol{F}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^\top \Phi^{-1}(\boldsymbol{\theta})(\mathbf{y} - \boldsymbol{F}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}) + \tau\sigma^2\sum_{k\in\boldsymbol{\gamma}}|\beta_k| + \lambda\sigma^2\sum_{k\in\boldsymbol{\gamma}}\theta_k}{\sigma^2}]) \\
&\times (\sigma^2)^{-(n+\nu_0)/2-1}\exp(-1/(2\sigma^2)) \times (\frac{q}{1-q}\tau\lambda)^{|\boldsymbol{\gamma}|}.
\end{aligned}$$

Clearly, the posterior density here has a complicated expression. To facilitate the computation, we borrow strength from empirical Bayes to obtain a good approximation of the posterior density.

3.2 Posterior Approximation

Denote $\rho_1 = \tau\sigma^2, \rho_2 = \lambda\sigma^2$, and $\omega = \frac{q}{1-q}\tau\lambda$. The posterior distribution can be represented as

$$P(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y}) \propto \exp(-\frac{1}{2}L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}))\omega^{|\boldsymbol{\gamma}|},$$

where $L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ is defined as

$$L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \log|\Phi(\boldsymbol{\theta})| + \frac{(\mathbf{y} - \mathbf{F}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}})^\top \Phi^{-1}(\boldsymbol{\theta})(\mathbf{y} - \mathbf{F}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}) + \rho_1\sum_{k\in\boldsymbol{\gamma}}|\beta_k| + \rho_2\sum_{k\in\boldsymbol{\gamma}}\theta_k}{\sigma^2}, \quad (3.4)$$

and the posterior marginal likelihood of $\boldsymbol{\gamma}$ is

$$P(\boldsymbol{\gamma}|\mathbf{y}) = C(\mathbf{y})\omega^{|\boldsymbol{\gamma}|}\int\int\exp(-\frac{1}{2}L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}))d\boldsymbol{\beta}_{\boldsymbol{\gamma}}d\boldsymbol{\theta}_{\boldsymbol{\gamma}}.$$

The major difficulty for obtaining this marginal posterior is the high-dimensional integration. To overcome this drawback, we focus on a subset of models with the highest posterior probability, of which the posterior probability can be well approximated. Then we introduce

a numerical algorithm to generate the samples of $\boldsymbol{\gamma}$ from this approximation density for Bayesian inference.

Here we focus on a subset of models with the highest posterior probability, which can be well approximated. Such an idea is similar to the idea of the maximizing-a-posterior (MAP), which is used in approximating the posterior for variable selection in linear models (Yuan and Lin, 2005). Note that $(\boldsymbol{\beta}, \boldsymbol{\theta})$ are dependent on $\boldsymbol{\gamma}$. Without loss of generality, we hereafter omit this dependency for notation convenience. We define $(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*)$ as

$$(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*) = \arg\min_{(\boldsymbol{\beta},\boldsymbol{\theta})} L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}) \tag{3.5}$$

Let us denote $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \boldsymbol{u}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^* + \boldsymbol{v}$. In the formulation of $L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta})$, we can have $\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta} = \boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta}^* - \boldsymbol{F}\boldsymbol{u}$. Moreover, we consider the Taylor expansion as

$$\begin{aligned}
\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}) &= \boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^* + \boldsymbol{v}) \\
&\approx \boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^*) - \boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^*)[\boldsymbol{v}^\top \circ \frac{\partial \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}]\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^*); \qquad (3.6) \\
\log|\boldsymbol{\Phi}(\boldsymbol{\theta})| &= \log|\boldsymbol{\Phi}(\boldsymbol{\theta}^* + \boldsymbol{v})| \\
&\approx \log|\boldsymbol{\Phi}(\boldsymbol{\theta}^*)| + \mathrm{tr}\big(\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^*)[\boldsymbol{v}^\top \circ \frac{\partial \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}]\big) \\
&\quad + \mathrm{tr}\big(\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^*)[\boldsymbol{v}^\top \circ \frac{\partial^2 \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} \circ \boldsymbol{v}]\big) \\
&\quad - \mathrm{tr}\big(\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^*)[\boldsymbol{v}^\top \circ \frac{\partial \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}]\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^*)[\boldsymbol{v}^\top \circ \frac{\partial \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}]\big) \qquad (3.7) \\
&\triangleq \log|\boldsymbol{\Phi}(\boldsymbol{\theta}^*)| + \mathrm{tr}(L(\boldsymbol{v})) + \mathrm{tr}(Q(\boldsymbol{v})),
\end{aligned}$$

where $L(\boldsymbol{v})$ and $Q(\boldsymbol{v})$ are linear terms and quadratic terms in approximating $\log|\boldsymbol{\Phi}(\boldsymbol{\theta})|$. Here the term $\boldsymbol{v}^\top \circ \frac{\partial \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}$ is a matrix with its (i,j) entry as $\boldsymbol{v}^\top \frac{\partial \phi_{ij}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}$, where $\phi_{ij}(\boldsymbol{\theta})$ is the (i,j)th entry of $\boldsymbol{\Phi}(\boldsymbol{\theta})$. Similarly, the term $\boldsymbol{v}^\top \circ \frac{\partial^2 \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} \circ \boldsymbol{v}$ is a matrix with its (i,j) entry as $\boldsymbol{v}^\top \frac{\partial^2 \phi_{ij}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\boldsymbol{v}$. The detailed expression for $\boldsymbol{v}^\top \circ \frac{\partial \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}$ and $\boldsymbol{v}^\top \circ \frac{\partial^2 \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} \circ \boldsymbol{v}$ can be found in the Supplementary. Then we can write the posterior marginal as

$$\begin{aligned}
P(\boldsymbol{\gamma}|\boldsymbol{y}) &= C(\boldsymbol{y})\omega^{|\boldsymbol{\gamma}|} \int\int \exp\left(-\frac{1}{2}L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta})\right) d\boldsymbol{\beta}_{\boldsymbol{\gamma}} d\boldsymbol{\theta}_{\boldsymbol{\gamma}} \\
&\approx C(\boldsymbol{y})\omega^{|\boldsymbol{\gamma}|} \exp\left(-\frac{1}{2}L_\rho(\boldsymbol{\beta}^*, \boldsymbol{\theta}^*)\right) \\
&\quad \times \int\int \exp\left(-\left[\frac{1}{2}\mathrm{tr}\big(L(\boldsymbol{v}) + Q(\boldsymbol{v})\big) + \frac{1}{2\sigma^2}f(\boldsymbol{u}, \boldsymbol{v})\right]\right) d\boldsymbol{u} d\boldsymbol{v}. \qquad (3.8)
\end{aligned}$$

Here $f(\boldsymbol{u}, \boldsymbol{v})$ is defined as

$$\begin{aligned}
f(\boldsymbol{u}, \boldsymbol{v}) &= -\boldsymbol{\epsilon}^{*\top}\boldsymbol{\Omega}\boldsymbol{\epsilon}^* - 2\boldsymbol{\epsilon}^{*\top}\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^* + \boldsymbol{v})\boldsymbol{F}\boldsymbol{u} \\
&\quad + (\boldsymbol{F}\boldsymbol{u})^\top\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^* + \boldsymbol{v})(\boldsymbol{F}\boldsymbol{u}) + \rho_1 \sum_{k\in\boldsymbol{\gamma}}(|\beta_k^* + u_i| - |\beta_k^*|) + \rho_2 \sum_{k\in\boldsymbol{\gamma}} v_k,
\end{aligned}$$

where $\boldsymbol{\epsilon}^* = \boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta}^*$, $\boldsymbol{\Omega} = \boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^*)[\boldsymbol{v}^\top \circ \frac{\partial \boldsymbol{\Phi}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}]\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta}^*)$. Now our main task is to evaluate

$$h(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2}\mathrm{tr}\big(L(\boldsymbol{v}) + Q(\boldsymbol{v})\big) + \frac{1}{2\sigma^2} f(\boldsymbol{u}, \boldsymbol{v}).$$

Based on the definition of $\boldsymbol{u}$ and $\boldsymbol{v}$, it is known that $h(\boldsymbol{u}, \boldsymbol{v})$ is minimized at $\boldsymbol{u}^* = 0$ and $\boldsymbol{v}^* = 0$ such that $h(u^*, v^*)$ is proportional to some constant.

Note that under different scenarios of γ, there can be two different types of models.

Definition 1. A model $\boldsymbol{\gamma}$ is called a *regular* model if and only if all coefficients of $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ (and $\boldsymbol{\theta}^*_{\boldsymbol{\gamma}}$) is nonzero.

Definition 2. A model $\boldsymbol{\gamma}$ is called a *nonregular* model if at least one coefficient of $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ (or $\boldsymbol{\theta}^*_{\boldsymbol{\gamma}}$) is zero.

We will discuss the approximation with respect to the two different model scenarios separately in the following propositions.

Proposition 1. *For the regular model in Definition 1, with sample size n large enough and using linearization approximation on both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, then one can approximate $P(\boldsymbol{\gamma}|\boldsymbol{y})$ as*

$$P(\boldsymbol{\gamma}|\boldsymbol{y}) = C_2 C(\boldsymbol{y})(\sqrt{\sigma^2}\omega)^{|\boldsymbol{\gamma}|} \times \exp\left(-\frac{1}{2}\min_{(\boldsymbol{\beta},\boldsymbol{\theta})} L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta})\right) + o(n). \tag{3.9}$$

The detailed derivation can be found in the Supplementary. Note that the approximation with $o(n)$ (3.9) comes from the Laplace approximation at the $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$. Here $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$ are essentially the MAP (maximum-a-posterior) estimators as shown in (3.5) and (3.4). When the sample size n grows, it is expected that the variance of the posterior distribution will become smaller, making the Laplace approximation to be adequate. However, for computer experiments with relatively small sample sizes, such an approximation may not be accurate. Based on our extensive simulation study in this work, it is found this approximation works fairly reasonable for the sample size $n = 30$ and $p = 5$. We have also tried the case of $n = 50$ or 100 for $p = 10$ or 20. The numerical results are also reasonable. One possible explanation is the design points are from a space-filing design and our approximation (i.e., Taylor series expansion) is expanded at the maximizing-a-posterior (MAP) estimators. It would be interesting to conduct a rigorous investigation of the approximation bound, which could be beyond the scope of this work.

The approximation obtained in (3.9) does not apply to the nonregular model. It is because that $h(\boldsymbol{u}, \boldsymbol{v})$ may not be differentiable at $\boldsymbol{u} = \boldsymbol{u}^*, \boldsymbol{v} = \boldsymbol{v}^*$ in nonregular model (see Definition 2). In this situation, we show that one can concentrate the regular model for the model selection procedure. Specifically, we compare a nonregular model $\boldsymbol{\gamma}$ with a regular submodel included. Assume that the $\boldsymbol{\gamma}$ has the form $\boldsymbol{\gamma} = (1, \ldots, 1, 0, \ldots, 0)$ with the first $|\boldsymbol{\gamma}|$ entries are 1's, and only the first s out of $|\boldsymbol{\gamma}|$ components of $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ and $\boldsymbol{\theta}^*_{\boldsymbol{\gamma}}$ are nonzeros, respectively. It means that $s < |\boldsymbol{\gamma}|$. Denote $\boldsymbol{\gamma}^*$ to be a p-dimensional binary vector as a submodel of $\boldsymbol{\gamma}$ with only the first s components being 1. The task here is to compare $P(\boldsymbol{\gamma}|\boldsymbol{y})$ and $P(\boldsymbol{\gamma}^*|\boldsymbol{y})$.

Proposition 2. *For the nonregular model in Definition 2, for sample size n large enough and using the linearization approximation, one can obtain,*

$$\frac{P(\boldsymbol{\gamma}|\boldsymbol{y})}{P(\boldsymbol{\gamma}^*|\boldsymbol{y})} \leq C(\sqrt{\sigma^2}\omega)^{|\boldsymbol{\gamma}|-s} + o(n). \tag{3.10}$$

The detailed derivation can be found in the Supplementary. One can see that if $\omega \leq 1$, the data would not give more support to the bigger model $\boldsymbol{\gamma}$ (Yuan and Lin, 2005). Therefore, we can focus on the regular model to avoid computing $P(\boldsymbol{\gamma}|\boldsymbol{y})$ for the nonregular model. Although the quality of such an approximation would rely on the large sample size, it is found in our simulation that this approximation works fairly well even for the small sample size. It would be interesting to understand the effect of $o(n)$ under the context of computer experiments when n is relatively small. One possible explanation is that the regular model, based on the Taylor expansion around the MAP estimators, received more support from the data than the complex nonregular model. We also would like to remark that the condition of $\boldsymbol{\theta} > 0$ could affect the accuracy of approximation in (3.10) since the first-order Laplace approximation at $\boldsymbol{\theta}^*$ is not established under the condition of $\boldsymbol{\theta} > 0$. But our empirical study does not encounter such a problem. A possible explanation is that $\boldsymbol{\theta} > 0$ is satisfied in the neighborhood of $\boldsymbol{\theta}^*$.

Remark 1. The issue of identifiability is common in Gaussian process modeling when parameters are estimated for both the mean function and the correlation function. The identifiability problem is relatively less an issue in the proposed model due to three reasons. First, the proposed method only includes linear terms in the mean function without higher-order terms. As a result, the correlation between the mean function coefficients and the correlation parameters is generally smaller and thus the identifiability issue has less impact. Second, the proposed Bayesian framework is closely related to a constraint estimation which imposes penalties to the parameter estimation, both for the mean function coefficient as well as for the correlation parameters, and therefore the identifiability issue can be further alleviated through a penalization. Third, based on equation (3.4) in Section 3.2, it can be shown from the empirical Bayes perspective that the regularization term $\sum \theta_i$ plays a similar role as containing the correlation length. Therefore, the estimated length parameters are penalized to avoid the identifiability issue. Furthermore, the proposed method is of interest when a large number of variables are involved in the computer experiments but a few of them have significant impacts. Thus those larger correlation lengths create a penalty to shrink the small lengths to zero and avoid the identifiability issue.

3.3 Bayesian Inference Procedure

Based on the analysis in Section 3.2, we can focus on the marginal posterior density function for the regular model. According to the expression in (3.4) and (3.9), we have

$$\begin{aligned} P(\boldsymbol{\gamma}|\mathbf{y}) &\approx C(\mathbf{y})(\sqrt{\sigma^2}\omega)^{|\boldsymbol{\gamma}|} \times \exp(-\frac{1}{2}\min_{\beta,\theta} L_\rho(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})) \\ &= C(\mathbf{y})(\sqrt{\sigma^2}\omega)^{|\gamma|} \times \exp(-\frac{1}{2}\tilde{L}_\rho(\gamma)), \end{aligned} \tag{3.11}$$

where $\tilde{L}_\rho(\boldsymbol{\gamma}) = \min_{(\boldsymbol{\beta},\boldsymbol{\theta})} L_\rho(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})$. Note that it is not trivial to find the model for the density in (3.11). To address this challenge, we take advantage of the sampling technique to

generate the corresponding Monte Carlo samples as the estimation of $P(\boldsymbol{\gamma}|\mathbf{y})$ for Bayesian inference. The details of the numerical algorithm are described in Algorithm 1. Here the σ^2 is chosen to be a pre-specified constant.

Algorithm 1 Numerical sampling algorithm for $\boldsymbol{\gamma}$

Step 1: Set initial values of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Step 2: Fix $\boldsymbol{\gamma}$ and update $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ by solving the minimization problem, $\min_{(\boldsymbol{\beta},\boldsymbol{\theta})} L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$.

Step 3: Fix $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, and then sequentially sample γ_i based on

$$P(\gamma_i = 1|\mathbf{y}, \boldsymbol{\gamma}_{-i}) = \frac{P(\gamma_i = 1, \boldsymbol{\gamma}_{-i}|\mathbf{y})}{P(\gamma_i = 0, \boldsymbol{\gamma}_{-i}|\mathbf{y}) + P(\gamma_i = 1, \boldsymbol{\gamma}_{-i}|\mathbf{y})},$$

for each $i = 1, 2, ..., p$. Here $\boldsymbol{\gamma}_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_p)^t$ represents the vector of all γ's except γ_i.

Step 4: Repeat Step 2 - 3 till convergence or the maximal number of iterations.

The proposed numerical algorithm is similar to a Monte Carlo Expectation Conditional Maximization (ECM) algorithm (Trevezas et al., 2014). Here we may treat the γ_i's as latent variables. In addition, there are non-explicit forms for both E- and M-steps. Thus the numerical optimization is adopted to identify the current best values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and then a Gibbs sampling type method is used to generate the samples of $\boldsymbol{\gamma}$ as shown in Hastie et al. (2001).

The Algorithm 1 is implemented in MATLAB. In Step 2, the minimization problem is solved respective to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ iteratively by taking "patternsearch" function in MATLAB. Suppose that we iterate the algorithm K times. We will discard the first few samples, say T, and then collect the remaining $(K - T)$ samples of γ vectors as the posterior samples of the indicators, γ_k. Based on our empirical experience, we usually set $K = 2,000$ and $T = 1,000$.

Having the Monte Carlo samples, we adopt the median probability criterion (Barbieri and Berger, 2004) for variable selection of active variable X_k's. Specifically, we estimate the marginal probability of $\gamma_k = 1$, $\hat{P}(\gamma_k = 1)$, for each variable X_k from the Monte Carlo samples and then we consider the kth variable to be active if $\hat{P}(\gamma_i = 1) \geq 0.5$. Note that in the literature, the highest posterior probability criterion is also commonly used in Bayesian variable selection, where the model is selected by maximizing the model posterior probabilities among all 2^p possible models. Barbieri and Berger (2004) have shown that under certain conditions, one can identify the same model under these two criteria for the linear regression. According to our numerical experience, the model identified by the median probability criterion would be in the top ranking in terms of the model posterior probabilities. Thus with the consideration of computational efficiency, the median probability criterion is used here. Once we determine the active variables, $\boldsymbol{\beta}_\gamma$ and $\boldsymbol{\theta}_\gamma$ can be estimated by solving the optimization problem in Eq. (3.5) with respect to the selected active variables. Thus the prediction value of $y(\mathbf{x})$ can be obtained according to Eq. (2.3).

In Algorithm 1, we treat σ^2 as a tuning parameter. To include σ^2 into the Algorithm 1, one can add a step for sampling σ^2 from its conditional distribution. Usually, the sampling of σ^2 is not needed in each iteration. One can update it after a few iterations for Steps 2 and 3.

Table 1: Different values of true $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in groups 1 to 5

	$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$	$\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$
scenario 1 :	(-0.2, 0, 0, 0, 0.4)	(0.3, 0, 0, 0.2, 0)
scenario 2 :	(-1.0, 0, 0, 0, 2.0)	(0.3, 0, 0, 0.2, 0)
scenario 3 :	(-2.0, 0, 0, 0, 4.0)	(0.3, 0, 0, 0.2, 0)
scenario 4 :	(-2.0, 0, 0, 0, 4.0)	(1.5, 0, 0, 1.0, 0)
scenario 5 :	(-2.0, 0, 0, 0, 4.0)	(3.0, 0, 0, 2.0, 0)

4. Simulation Study

In this section, we examine the performance of the proposed method by a five-dimensional simulation study with data generated from a pre-specified Gaussian process. In the Supplementary, we also demonstrate a numerical example based on the typical setting of a computer experiment and discuss the parameter tuning issue.

Here we compare the proposed method with the blind kriging method (Joseph et al., 2008). The blind kriging method, modified from the ordinary kriging, has an unknown mean function to be identified through some data-analytic procedures. Joseph et al. (2008) considered the Bayesian forward selection technique for the unknown mean model under the maximum likelihood estimates of the correlation parameters. Here, the blind kriging is implemented using a MATLAB toolbox called "ooDACE" (Couckuyt et al., 2012) which integrates the correlation parameter estimation and the estimation of the unknown mean model. In addition, we consider a variant of the selection approach in Linkletter et al. (2006) such that it can identify active variables in the mean function and the covariance function.

We consider the input $\mathbf{x}$ with dimensionality to be $p = 5$. Without loss of generality, we set the experimental region as $[0, 1]^p$. To generate the simulation data, we first sample the input data points, $\mathbf{x}_i$, $i = 1, 2, \ldots, n$, from a Latin hypercube design. Here different sample sizes $n = 5, 10, 15, 20, 30$ are considered. The responses, y_i, are generated by following the Gaussian process model in (2.1) with $\kappa = 2$ and pre-specified $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

To investigate the effects of different scenarios of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ on selecting active variables, we consider five scenarios of active variables as shown in Table 8. Here, we consider the variable being active under three different situations: active in the mean part, active in thecorrelation part, and active in both mean part and correlation part. In all set-ups, X_1, X_4 and X_5 are active variables. For X_1, both β_1 and θ_1 are non-zeros. For X_4, its correlation parameter θ_4 is non-zero but β_4 is fixed as zero. For X_5, the β_5 is non-zero but θ_5 is fixed as zero. The differences among the five scenarios are the scales of true $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. As the Gaussian correlation function is used, it is interesting to investigate the scale effects with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

To evaluate the accuracy of the proposed method, we consider the following performance measures: the True Classification Rate (TCR), the True Positive Rate (TPR), the False

Table 2: The TCR, TPR and FPR for different n and scenarios

		$n=5$	$n=10$	$n=15$	$n=20$	$n=30$
Scenario 3	avg. TCR	0.864	0.972	0.996	1	1
	avg. TPR	0.913	0.993	1	1	1
	avg. FPR	0.21	0.06	0.01	0	0

Positive Rate (FPR),

$$\begin{aligned} \text{TCR} &= \frac{\text{number of correctly selected variables}}{\text{number of variables}}; \\ \text{TPR} &= \frac{\text{number of correctly selected active variables}}{\text{number of active variables}}; \\ \text{FPR} &= \frac{\text{number of falsely selected active variables}}{\text{number of inactive variables}}. \end{aligned}$$

The TCR is an overall evaluation of the accuracy in the identification of the active and inactive variables. TPR is the average rate of active variables identified correctly and is used to measure the power of the method. FPR is the average rate of inactive variables that are included in the regression and can be considered as type I error of the selected approach. Larger values of TCR and TPR indicate better performance, whereas smaller values of FPR indicate better performance than larger values.

Here we fix the tuning parameters as $q = 0.5, \sigma = 1, \tau = 1, \lambda = 1$, and $\kappa = 2$ for the proposed method. For each scenario with sample size n, we independently repeat this simulation 50 times, and the 50 selection results are summarized in Table S1 in Supplementary. Table 9 is the selection results for Scenario 3. From the results, it is seen that the values of TCR, TPR and FPR in all five scenarios are acceptable when n is relatively large. Specifically, the values of TCR and TPR are close to 1 and the values of FPR are close to 0, which indicates good accuracy in selecting active variables. In fact, we have tried the cases with a larger sample size, n. For example, when $n = 500$, based on Scenario 3 and the same tuning parameters, the values of TCR and TPR are also equal to 1 and the value of FPR is 0. To save space, the results with larger sample sizes are omitted here. Moreover, the comparison of Scenarios 1, 2 and 3 indicates that the proposed method can obtain better selection performance when β_k is larger under fixed $\boldsymbol{\theta}$. The comparison of Scenarios 3, 4 and 5 indicates that when fixing $\boldsymbol{\beta}$, the smaller $\boldsymbol{\theta}$ is, the better selection accuracy of the proposed method can achieve.

Furthermore, we examine the effect of the tuning parameters, λ and τ, on selecting active variables. Taking the setting of Scenario 3 with $n = 15$ for illustration, we consider a set of possible (λ, τ) to be $\{(0.5, 0.5), (0.5, 1), (1, 0.5), (1, 1), (1, 5), (5, 1), (5, 5)\}$, while σ^2 is set as 1. We repeat the simulation 50 times under each setting of tuning parameters. Table 3 reports the selection results of the proposed method over 50 replications. Table 3 shows that when the values of λ and τ become larger, the values of TPR keep on 1, but the values of FPR become larger accordingly. Larger FPR values mean that more inactive variables are identified as active ones by our approach. Thus the results from Table 3 could imply that larger values of λ and τ could result in over-selecting the variables. The selection of

Table 3: The TCR, TPR and FPR for different values of λ and τ

	(λ, τ)						
	(0.5, 0.5)	(0.5, 1)	(1, 0.5)	(1, 1)	(1, 5)	(5, 1)	(5, 5)
avg. TCR	1	0.996	1	0.996	0.996	0.992	0.98
avg. TPR	1	1	1	1	1	1	1
avg. FPR	0	0.01	0	0.01	0.01	0.02	0.05

Table 4: Results of the blind kriging for simulations with $n = 15$

	avg. TCR	avg. TPR	avg. FPR
scenario 1	0.852	0.8867	0.1267
scenario 2	0.856	0.9267	0.1667
scenario 3	0.868	0.9467	0.1667
scenario 4	0.832	0.8667	0.1467
scenario 5	0.808	0.8267	0.1467

tuning parameters will be discussed in Supplementary. Finally, due to the median probability criterion, we choose 0.5 as the threshold value for the posterior probability to decide whether the variable is active or not. We have tried the other threshold values, like $0.1, 0.2, \ldots, 0.8$ and 0.9 and no matter what the threshold value is, the values of TPR and FPR are similar. Thus we simply fix this threshold value as 0.5 for our proposed method.

Note that it is important to show the convergence of the proposed numerical algorithm. Here the Monte Carlo standard error (MCSE), introduced in Jones et al. (2006) and Flegal et al. (2008), can be used to check the convergence of the Monte Carlo samples. When the corresponding MCSE value is sufficiently small, it indicates the convergence of the Monte Carlo samples. Here we compute MCSE of the samples of indicator parameters by choosing the whole samples as one batch, and the threshold value of MCSE is set as 0.04 as suggested in Flegal et al. (2008). Take the five scenarios in Section 4.1 with $n = 15$ as an illustration. Among the total of 250 cases, there is only one case whose maximal MCSE value of the five variables is 0.065, and for the other cases, all standard deviation values are less than the threshold values. The samples of indicator parameters in the process appeared to be stuck on 0 and 1 for the variables being active or not. Under 2000 iterations, after burning in the first 1000 samples, the MCSE of the remaining indicators samples is less than the threshold value, 0.04. This provides proper evidence of the convergence of the MCMC process.

In addition, we examine the blind kriging method under the five scenarios with $n = 15$. We compute the average TCR, TPR and FPR in each scenario based on 50 independent replications, and the results are reported in Table 4. By comparing the results of the proposed method with results in Table 4, the proposed method can generally be more accurate than the blind kriging method on variable selection. In particular, it is seen that when true β is large and true θ is small, the blind kriging tends to over-select the variables because of the non-zero FPR values.

The blind kriging selects the active variables from the mean function. While the proposed selection approach not only targets on the active variables in the mean function but also

Table 5: The results based on the approach generalized from Linkletter et al. (2006).

	avg. TCR	avg. TPR	avg. FPR
scenario 1	1.0000	1.0000	0.0000
scenario 2	1.0000	1.0000	0.0000
scenario 3	1.0000	1.0000	0.0000
scenario 4	0.9840	1.0000	0.0400
scenario 5	0.9520	1.0000	0.1200

considers the non-zero correlation parameters. For benchmark comparison, we also include a variant of the selection approach in Linkletter et al. (2006). Linkletter et al. (2006) introduced a Bayesian selection approach with a focus on the correlation function, which needs to generate a new inert variable in the analysis. Since this inert variable must be non-active, we consider a variant of their approach by generating the posterior samples of the coefficient and correlation parameter for this inert factor as reference distributions to check whether the other variables are active or not. We take the same five scenarios with $n = 15$ as an illustration, and the selection results are summarized based on 50 independent replications. We first generate the inert variables in each replication and set β_6 and θ_6 as their corresponding coefficients. Then we iterate our algorithm 2000 times and compute the posterior medians for β_6 and θ_6 based on the last 1000 iterations. Thus, a variable will be considered active if at least one of its posterior medians is larger than the median values for β_6 and θ_6. Table 5 reports the selection performance of this variant method. Generally, this method has a similar performance to our proposed approach in terms of TCR, TPR and FPR and outperforms the blind kriging. It provides certain evidence that it is useful to consider both effects in the mean function and the correlation function.

5. Simulations with 10 and 20 variables

In this section, we consider a large variable dimension in the simulation studies. In computer experiments, the variable dimension is not very large due to the concern of the curse of dimensionality and expensive computational cost. Thus we set the number of the variables, $p = 10$ and 20, by extending Scenario 3 shown in Section 4.1 through adding 5 or 15 inter variables, respectively. That is, $\beta_{true} = (-2, 0, 0, 0, 4, 0, ..., 0)$ and $\theta_{true} = (0.3, 0, 0, 0.2, 0, 0, ..., 0)$. It means that there are only three active variables, X_1, X_4 and X_5.

First, we consider the cases with $p = 10$ and $n = 50$ and 100. In each replication, we choose n points from a LHD in $[-1, 1]^{10}$, and generated the responses based on β_{true} and θ_{true}. Based on 50 independent replications, the selection results for $n = 50$ are summarized as TCR = 1.0000, TPR = 1.0000 and FPR = 0.0000. For the case of $n = 100$, the selection results from 50 replications are TCR = 0.9800, TPR = 0.9933, FPR = 0.0257. Generally, the proposed method can identify active variables with very high probability and only over-select few inactive variables in very low frequencies.

The distributions of β_i and θ_j among 50 replications for the case with $n = 50$ are in Figure S1 in the Supplementary. Here β_i and θ_i are the optimal solutions in each iteration of our algorithm. The variables $X_6, \ldots, X_{10}$ can be treated as five inert variables. For these

inert variables, the corresponding β_i and θ_i are all stuck on zeros. For the active variables, X_1, X_4 and X_5, the medians of the corresponding samples are significantly far from zeros and are all close to the true values. These results indicate that the proposed approach can have high TPR and low FPR values. In addition, we also compute the 95% HPD intervals for all parameters via the R function `hdi` in the package "HDInterval" and report these values in Table S3 in the Supplementary. Overall, the lengths of the HPD intervals are quite small because the parameters, β_i and θ_i, are obtained via an optimization approach based on the training data and the current indicators.

Furthermore, we also conduct variance-based sensitivity analysis to compare with the proposed method. A function called "sobol2002" from an R-package "sensitivity" is used for sensitivity analysis, which implements the Monte Carlo estimation of the Sobol indices for both first-order and total indices at the same time. Here we consider the first-order index value generated from this method for identifying active variables. Only the variables with first-order indices that are larger than a threshold will be selected as active variables. The case with $p = 10$ and $n = 100$ is studied here. To find the best selection results, we went over 14 thresholds as 0.01, 0.02,..., 0.09, 0.1, 0.2,..., 0.5. Among these thresholds, the maximum TCR is 0.7960, which appears when the threshold is set at 0.05, and the maximum TPR is 0.7933, which appears when the threshold is set at 0.01. Both of these two values are smaller than those of the proposed approach. The FPR values decrease as rthe threshold increases: the FPR=0.31 if the threshold is set at 0.01 and the FPR=0.09 if the threshold is set at 0.05. Note that only X1 and X5 will be selected when we set the threshold as 0.5. Thus it can be seen that our proposed approach overall has better performance in terms of TPR and FPR.

Now we consider the cases of $p = 20$. The simulation setting is the same as the case of $p = 10$ except having 15 inert variables. Here we set the sample size, n, as 100 and 200, and generate the experimental points from an LHD over $[-1, 1]^{20}$. To avoid the numerical problems in MATLAB, we multiplied the sample points by eight instead of inputting the original sample points. The selection results are summarized from the 50 independent replications. When $n = 100$, the proposed method has TCR = 0.9670, TPR = 0.8933 and FPR = 0.0200. For the case of $n = 200$, three measurements are TCR = 0.9460, TPR = 0.9133, FPR = 0.0482. Overall the proposed method also works quite well even though there are more inert variables.

6. A Real-Data Case Study

This real example has been studied in Fang et al. (2005) and Joseph et al. (2008). Here we provide a brief background of the computer experiment. The engine block and head joint sealing assembly is a fundamental structural design in the automotive internal combustion engine. Design decisions need to be made upfront prior to the availability of a physical prototype. The design of the joint sealing affects downstream design decisions for other engine components and can significantly impact the long lead time tooling and machining facility setup. It is very expensive in time and expense to conduct such designs. The use of a computer simulation model is indispensable (Chen et al., 2002). The engine block and head joint sealing assembly is very complex due to multiple functional requirements (e.g., combustion gas, high-pressure oil, oil drain, and coolant sealing) and complicated geometry. The interactions among design parameters in this assembly (block and head structures,

Table 6: RMSE under different parameters

σ	λ	τ	RMSE	σ	λ	τ	RMSE
1	0.1	0.1	1.0463	1.2	1	1	0.8109
1	0.5	0.5	0.6917	1.5	1	1	0.9659
0.6	1	1	0.7330	1	1.5	1.5	0.7741
0.8	1	1	0.7221	1	2	2	0.7333
1	1	1	0.6749	1	5	5	0.7333

gasket, and fasteners) have significant effects. Usually, a finite element model was used to capture the complexity of part geometry, the compliance in the components, non-linear material properties, and the contact interface between the parts. To address the performance robustness of the joint sealing, manufacturing variability of the mating surfaces and head bolt tensional load are included in the analysis for which design parameters are optimized.

Here, eight factors are selected for experimentation. These are gasket thickness (x_1), number of contour zones (x_2), zone-to-zone transition (x_3), bead profile (x_4), coining depth (x_5), deck face surface flatness (x_6), load/deflection variation (x_7), and head bolt force variation (x_8). Because of the complexity in the simulation setup and the excessive computing requirements, a 27-run orthogonal array is used and is shown in Supplementary Section S7. In this example, the gap lift (y) is the response variable.

For using the proposed variable selection method to consider both effects in the mean function and the correlation parameters, we only involve the main effects in the mean part, i.e., $\boldsymbol{f}(x)^T\boldsymbol{\beta} = \boldsymbol{x}^T\boldsymbol{\beta}$. Note that in Joseph et al. (2008), the main effects and interaction effects are used to construct the mean model for the blind kriging. The predictive RMSE, based on leave-one-out cross validation (LOOCV), is used as a performance measure. A smaller value of RMSE indicates better performance on prediction. The values of RMSE under different settings of tuning parameters are reported in Table 6

Clearly, it is seen that $(\sigma, \lambda, \tau) = (1, 1, 1)$ gives the smallest RMSE than other settings in the table. This smallest RMSE is also smaller than the RMSE obtained by the ordinary kriging model which is 0.7333. This result indicates that the proposed method can obtain a better prediction through the proper variable selection. According to the table in Supplementary Section S8, it shows that the posterior probabilities of $\gamma_i = 1$ for $i = 1, \ldots, 8$ in each trail in the LOOCV by fixing (σ, λ, τ) = (1, 1, 1). Based on the median probability criterion, once the posterior probability of $\gamma_i = 1$ is great than or equal to 0.5, x_i is treated as active. Thus x_1, x_3, x_6 and x_8 are the active variables. The other variables should be inactive because in most trails, the corresponding posterior probabilities are less than 0.5. For these four active variables, we estimate the corresponding β_i and θ_i via the MLE method. The results show that x_1, x_6 and x_8 have significant effects in both the mean function and correlation function. While for x_3, we have $\beta_3 = 0.0166$ and $\theta_3 = 7.04 \times 10^{-4}$ close to 0. It implies that x_3 may only affect the mean function.

7. Extension to Two-indicator Approach

When considering the different meanings of β and θ, instead of single indicator, we define two indicators for β_k and θ_k separately. Unlike the single indicator approach used above,

Table 7: The averages of TCR, TPR and FPR for β and θ for the Algorithm 2

	TCR	TPR	FPR
β	0.7880	0.7800	0.2067
θ	0.7960	1.0000	0.3400

two binary vectors are used to denote respectively whether β_i and θ_i are zeros or not. Let $\gamma_\beta = (\gamma_{\beta,1}, ..., \gamma_{\beta,p})$. Specifically $\gamma_{\beta,k} = 1$ if β_k is non-zero, and $\gamma_{\beta,k} = 0$ otherwise. Similarly, let $\gamma_\theta = (\gamma_{\theta,1}, ..., \gamma_{\theta,p})$. Thus, $\gamma_{\theta,i} = 1$ if θ_i is non-zero, and $\gamma_{\theta,i} = 0$ otherwise.

The priors of β_k and θ_k are also the mixture distributions, and the same as these used in Algorithm 1 by replacing γ_k with $\gamma_{\beta,k}$ and $\gamma_{\theta,k}$, separately. For the priors of γ_β and γ_θ, Bernoulli distributions with different probabilities are adopted here. By assuming the independence among factors, the priors can be written as

$$P(\boldsymbol{\gamma}_\beta) \propto q_1^{|\boldsymbol{\gamma}_\beta|}(1-q_1)^{p-|\boldsymbol{\gamma}_\beta|},$$

$$P(\boldsymbol{\gamma}_\theta) \propto q_2^{|\boldsymbol{\gamma}_\theta|}(1-q_2)^{p-|\boldsymbol{\gamma}_\theta|},$$

where $|\boldsymbol{\gamma}_\beta| = \sum_{k=1}^p \gamma_{\beta,k}$ and $|\boldsymbol{\gamma}_\theta| = \sum_{k=1}^p \gamma_{\theta,k}$. Denote $\omega_1 = \frac{q_1}{1-q_1}\lambda$ and $\omega_2 = \frac{q_2}{1-q_2}\tau$. In addition to the independent assumptions in Section 3.1, we also assume that the priors of γ_β and γ_θ are independent. Then we can write

$$P(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}_\beta, \boldsymbol{\gamma}_\theta | \mathbf{y}) \propto \exp(-\frac{1}{2} L_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}_\beta, \boldsymbol{\gamma}_\theta)) \omega_1^{\gamma_\beta} \omega_2^{\gamma_\theta},$$

where $L_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}_\beta, \boldsymbol{\gamma}_\theta)$ is defined as

$$\begin{aligned} L_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}_\beta, \boldsymbol{\gamma}_\theta) &\propto \log|\Phi(\boldsymbol{\theta}_{\gamma_\theta})| \\ &+ \frac{(\mathbf{y} - \boldsymbol{F}_{\boldsymbol{\gamma}_\beta}\boldsymbol{\beta}_{\boldsymbol{\gamma}_\beta})^\top \Phi^{-1}(\boldsymbol{\theta})(\mathbf{y} - \boldsymbol{F}_{\boldsymbol{\gamma}_\beta}\boldsymbol{\beta}_{\boldsymbol{\gamma}_\beta}) + \rho_1 \sum_{k \in \boldsymbol{\gamma}_\beta} |\beta_k| + \rho_2 \sum_{k \in \boldsymbol{\gamma}_\theta} \theta_k}{\sigma^2} \end{aligned}$$

Following the similar approximation procedures, the posterior marginal likelihood of γ_β and γ_θ can be approximated as

$$P(\gamma_\beta | \gamma_\theta, y) \propto \omega_1^{|\gamma_\beta|} \omega_2^{|\gamma_\theta|} exp(-\frac{1}{2} L_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}_\beta, \boldsymbol{\gamma}_\theta)).$$

$$P(\gamma_\theta | \gamma_\beta, y) \propto \omega_1^{|\gamma_\beta|} \omega_2^{|\gamma_\theta|} exp(-\frac{1}{2} L_p(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}_\beta, \boldsymbol{\gamma}_\theta)).$$

Here we also treat σ as a tuning parameter and it should be specified before implementing the following Algorithm 2.

Here Algorithm 2 is also implemented by MATLAB. To illustrate the performance of Algorithm 2, we revisit scenario 3 in Table 1 with $n = 15$. Based on the same simulation set-ups, the means of TCR, TPR and FPR for β and θ are used to evaluate the performance of Algorithm 2. The results are as shown in Table 7. The TPR of β equals 0.78, which means that only few true active variables might not be correctly selected as important variables,

Algorithm 2 Numerical sampling algorithm for $\boldsymbol{\gamma}_\beta$ and $\boldsymbol{\gamma}_\theta$

Step 1: Set initial values of $\boldsymbol{\gamma}_\beta$, $\boldsymbol{\gamma}_\theta$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Step 2: Fix $\boldsymbol{\gamma}_\beta$ and $\boldsymbol{\gamma}_\theta$ and update $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ by solving the minimization problem, $\min_{(\boldsymbol{\beta},\boldsymbol{\theta})} L_\rho(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}_\beta,\boldsymbol{\gamma}_\theta)$.

Step 3: Fix $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, and then sequentially sample $\gamma_{\beta,i}$ based on

$$P(\gamma_{\beta,i}=1|\mathbf{y},\boldsymbol{\gamma}_{\beta,-i})=\frac{P(\gamma_{\beta,i}=1,\boldsymbol{\gamma}_{\beta,-i}|\mathbf{y})}{P(\gamma_{\beta,i}=0,\boldsymbol{\gamma}_{\beta,-i}|\mathbf{y})+P(\gamma_{\beta,i}=1,\boldsymbol{\gamma}_{\beta,-i}|\mathbf{y})},$$

And sample $\gamma_{\theta,i}$ based on

$$P(\gamma_{\theta,i}=1|\mathbf{y},\boldsymbol{\gamma}_{\theta,-i})=\frac{P(\gamma_{\theta,i}=1,\boldsymbol{\gamma}_{\theta,-i}|\mathbf{y})}{P(\gamma_{\theta,i}=0,\boldsymbol{\gamma}_{\theta,-i}|\mathbf{y})+P(\gamma_{\theta,i}=1,\boldsymbol{\gamma}_{\theta,-i}|\mathbf{y})},$$

for each $i=1,2,...,p$. Here $\boldsymbol{\gamma}_{\beta,-i}=(\gamma_{\beta_1},\ldots,\gamma_{\beta_{i-1}},\gamma_{\beta_{i+1}},\ldots,\gamma_{\beta_p})^t$ represents the vector of all γ_β's except γ_{β_i}. And $\boldsymbol{\gamma}_{\theta,-i}=(\gamma_{\theta_1},\ldots,\gamma_{\theta_{i-1}},\gamma_{\theta_{i+1}},\ldots,\gamma_{\theta_p})^t$ represents the vector of all γ_θ's except γ_{θ_i}

Step 4: Repeat Step 2 - 3 till convergence or the maximal number of iterations.

Table 8: Different values of true $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in new scenarios

	$\boldsymbol{\beta}=(\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)$	$\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)$
scenario 3 :	(-2, 0, 0, 0.0, 4)	(0.3, 0, 0, 0.2, 0)
scenario 3.1 :	(-2, 0, 0, 0.3, 4)	(0.3, 0, 0, 0.2, 0)
scenario 3.2 :	(-2, 0, 0, 0.3, 4)	(0.3, 0, 0, 2.0, 0)
scenario 3.3 :	(-2, 0, 0, 3.0, 4)	(0.3, 0, 0, 2.0, 0)

and the FPR of β equals 0.2067, which means that the over-selection problem for the mean function does exist. For the correlation parameters, Algorithm 2 can identify all active θ_i, because the TPR value is equal to 1.0000. However, the over-selection problem for θ_i still exists and based on Table 7, more than one-third of the inactive θ_i are selected as important variables.

To compare the performance of Algorithms 1 and 2, in addition to scenario 3 in Section 4, more different scenarios are considered. In these scenarios, the different values of β_4 are chosen and these scenarios are shown in Table 8. To have a fair comparison, the results of two indicators in Algorithm 2 are re-summarized as the one indicator approach in Algorithm 1. That is that a variable, x_k, is active if $\gamma_{\beta,k}$ or $\gamma_{\theta,k}$ is equal to 1, and is non-active only when $\gamma_{\beta,k}=\gamma_{\theta,k}=0$. For these scenarios, we still fix $n=15$ and the other set-ups are the same as those in Section 4. The averages of TCR, TPR and FPR among 50 replications for both algorithms are reported in Table 9, and we also report the CPU time for each scenario with 50 replications. Here we run our MATLAB codes on the computer with Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz & 2.10 GHz and 128 GB RAM. According to Table 9, firstly, both algorithms have high TPR values, because the lowest value is 0.9933, which is quite close to one. It means that both algorithms can identify the true active variables. In addition, Algorithm 1 has higher TCR values and lower FPR values in all four scenarios.

Table 9: Average TCR, TPR and FPR for Algorithm 1 and 2, and different scenarios

		TCR	TPR	FPR	Time(s)
Algorithm 1	scenario 3	1.0000	1.0000	0.0000	148342
	scenario 3.1	0.9840	0.9933	0.0300	170667
	scenario 3.2	0.9800	1.0000	0.0500	190144
	scenario 3.3	0.9920	1.0000	0.0200	185888
Algorithm 2	scenario 3	0.8960	1.0000	0.2600	200687
	scenario 3.1	0.9000	1.0000	0.2500	208005
	scenario 3.2	0.8640	1.0000	0.3400	227651
	scenario 3.3	0.8640	1.0000	0.3400	251695

Thus Algorithm 2 has more serious over-selection problems. Finally Algorithm 2 takes more CPU time. This is because twice the number of indicators are used in Algorithm 2 and thus the larger model space is defined for Algorithm 2 to search the best active variable set.

8. Discussion

In this work, we proposed an indicator-based Bayesian variable selection method for Gaussian process model. To take into account the correlation in the regularization procedure, a hierarchical Bayesian structure is superimposed in this paper by the design of indicator functions and therefore, the identified active variables may have effects in the mean function and/or in the correlation function. The use of group selection in the proposed method rather than separate selection is to tie the regularization of two effects, one in the mean function and one in the correlation function, from the same variable by a hierarchical Bayesian structure, which is not only intuitive but also parsimonious. For active variables, their estimation may suffer from the identifiability issue. The use of empirical Bayesian procedure in the proposed method can potentially alleviate this issue through constraining the correlation lengths with a proper prior. Another possible mitigation is to consider the orthogonal GPs (Plumlee and Joseph, 2018) for the proposed selection framework.

Note that the proposed method is also applicable to general power exponential correlation functions with different smoothness. Instead of pre-specifying the hyper-parameter for smoothness, one direction for future work is to incorporate the estimation of smoothness into the proposed Bayesian framework. In a preliminary study of Scenario 3 in Section 4.1, we have observed promising results of TCR = 0.988, TPR = 1.000 and FPR = 0.03 by using L_1-norm in the power exponential correlation function, i.e., $\kappa = 1$. We will further extend the proposed method to other correlation functions, such as variants of the Matern function (Gu et al., 2018), to enable meaningful variable selection. Another direction for future work is to seek a more effective procedure of parameter estimation under the empirical Bayes framework. Currently, the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be viewed as a penalized likelihood estimation. Alternatively, one can consider the restricted likelihood estimation (REML) approach (Lewis et al., 2021). It will be interesting to integrate the REML procedure with the proposed Bayesian selection method.

For the proposed algorithms, one needs to pre-specify the number of iterations. To determine the number of iterations automatically, we suggest using the MCSE value to

Table 10: The RMSE values under different parameters via the LOOCV

σ	λ	τ	RMSE
1	0.5	0.5	0.1721
1	0.5	1	0.3129
1	1	0.5	0.1664
1	1	1	0.3026
1	1	5	0.8326
1	5	1	0.7333
1	5	5	0.6663

define the stopping criterion. That is, for every certain iterations (e.g., 100 iterations), one can compute the MCSE values for all indicators. If the maximal MCSE value is less than the pre-specified threshold value, we can stop the algorithm; otherwise, we keep implementing the procedure. To implement the proposed algorithms, one also needs to pre-specify the tuning parameters, like λ and τ. The leave-one-out cross validation (LOOCV) can be a data-driven approach to determine the proper parameter values. We have implemented the LOOCV approach in our real data case study in Section 6 to choose the proper values. Moreover, take scenario 3 and $n = 15$ in Section 4 as an illustration. The values of RMSE generated from the LOOCV under different settings of tuning parameters, λ and τ, are reported in Table 10. The case with $(\lambda, \tau) = (1, 0.5)$ has the smallest RMSE value. This best parameter setup is the same as what we have used in Section 4. In addition to the LOOCV, Nguyen (2019) introduced the Bayesian optimization approach for parameter tuning. In the Bayesian optimization approach, one can consider the parameter tuning as a black-box optimization problem and a sequential design procedure is used to identify the best parameter values within a few iterations. Note that the proposed method only has two or three tuning parameters, the LOOCV approach can be a proper choice when the number of tuning parameters is small.

Moreover, the proposed variable selection for GPs can be extended for the group variable selection (Lai and Chen, 2020). We can use an indicator parameter to denote a group is active or not, and the proposed approach could be modified accordingly. One can also extend the proposed variable selection method for the Gaussian process models of computer experiments with both quantitative and qualitative factors (Zhou et al., 2011; Qian et al., 2008; Deng et al., 2017). Note that when the qualitative factors, discrete in nature, are presented in the model, it will be interesting to investigate how the indicator-based variable selection can be adopted for the variable selection of qualitative factors. Finally, an interesting direction is to consider a full Bayesian MCMC procedure for inference. Based on our empirical study, the key to efficiently implementing a fully Bayesian MCMC relies on an efficient sampling procedure for the correlation parameter θ. Some existing approaches, such as the "slice sampling" discussed by Huang et al. (2020), have not yet achieved sufficient efficiency in the estimation based on our preliminary study. It is also pointed out by Huang et al. (2020) that the direct use of the slice sampling is not generally recommended because of the computational issues in high-dimensional problems. As a future research, it will be interesting to investigate how to conduct an efficient sampling procedure for θ to enable a fully Bayesian approach.

References

Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *The Annals of Statistics 32*(3), 870–897.

Casella, G. and T. Park (2008). Bayesian lasso. *J. Amer. Statist. Assoc 103*(482), 681–686.

Chen, B., R. Castro, and A. Krause (2012). Joint optimization and variable selection of high-dimensional gaussian processes. *arXiv preprint arXiv:1206.6396*.

Chen, J. K., R.-B. Chen, A. Fujii, R. Suda, and W. Wang (2018). Surrogate-assisted tuning for computer experiments with qualitative and quantitative parameters. *Statistica Sinica 28*(2), 761–789.

Chen, T., J. Zwick, B. Tripathy, and G. Novak (2002, 03). 3D engine analysis and mls cylinder head gaskets design. In *SAE Technical Paper*. SAE International.

Couckuyt, I., A. Forrester, D. Gorissen, F. De Turck, and T. Dhaene (2012, July). Blind kriging: Implementation and performance analysis. *Adv. Eng. Softw. 49*, 1–13.

Deng, X., C. D. Lin, K.-W. Liu, and R. Rowe (2017). Additive gaussian process for computer models with qualitative and quantitative factors. *Technometrics 59*(3), 283–292.

Fang, K.-T., R. Li, and A. Sudjianto (2005). *Design and Modeling for Computer Experiments (Computer Science & Data Analysis)*. Chapman & Hall/CRC.

Flegal, J. M., M. Haran, and G. L. Jones (2008). Markov Chain Monte Carlo: Can We Trust the Third Significant Figure? *Statistical Science 23*(2), 250 – 260.

Gramacy, R. B. and H. K. H. Lee (2008). Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association 103*(483), 1119–1130.

Gu, M., J. Palomo, and J. O. Berger (2018). Robustgasp: Robust gaussian stochastic process emulation in r. *arXiv preprint arXiv:1801.01874*.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Huang, H., D. K. J. Lin, M.-Q. Liu, and Q. Zhang (2020). Variable selection for kriging in computer experiments. *Journal of Quality Technology 52*(1), 40–53.

Hung, Y. (2011). Penalized blind kriging in computer experiments. *Statistica Sinica 21*(3), 1171–1190.

Jones, G. L., M. Haran, B. S. Caffo, and R. Neath (2006). Fixed-width output analysis for markov chain monte carlo. *Journal of the American Statistical Association 101*(476), 1537–1547.

Joseph, V. R. (2006). Limit kriging. *Technometrics 48*(4), 458–466.

Joseph, V. R., Y. Hung, and A. Sudjianto (2008). Blind kriging: A new method for developing metamodels. *Journal of Mechanical Design 130*(3), 031102–031102–8.

Lai, W. and R. Chen (2020). A review of bayesian group selection approaches for linear regression models. *WIREs Computational Statistics*. to appear.

Levy, S. and D. M. Steinberg (2010). Computer experiments: a review. *AStA Advances in Statistical Analysis 94*(4), 311–324.

Lewis, J. R., S. N. MacEachern, and Y. Lee (2021). Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression. *Bayesian Analysis*. In press.

Linkletter, C., D. Bingham, N. Hengartner, D. Higdon, and K. Q. Ye (2006). Variable selection for gaussian process models in computer experiments. *Technometrics 48*(4), 478–490.

Nguyen, V. (2019). Bayesian optimization for accelerating hyper-parameter tuning. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pp. 302–305.

Plumlee, M. and V. R. Joseph (2018). Orthogonal gaussian process models. *Statistica Sinica 28*(2), 601–619.

Qian, P. Z. G., H. Wu, and C. J. Wu (2008). Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics 50*(3), 383–396.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer.

Reich, B. J., C. B. Storlie, and H. D. Bondell (2009). Variable selection in bayesian smoothing spline anova models: Application to deterministic computer codes. *Technometrics 51*(2), 110–120.

Sacks, J., S. B. Schiller, and W. J. Welch (1989). Designs for computer experiments. *Technometrics 31*(1), 41–47.

Santner, T. J., B. J. Williams, and W. I. Notz (2003). *The Design and Analysis of Computer Experiments*. Springer series in statistics. Springer.

Trevezas, S., S. Malefaki, and P.-H. Cournède (2014). Parameter estimation via stochastic variants of the ecm algorithm with applications to plant growth modeling. *Computational Statistics & Data Analysis 78*, 82–99.

Welch, W. J., R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris (1992). Screening, predicting, and computer experiments. *Technometrics 34*(1), 15–25.

Yuan, M. and Y. Lin (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association 100*(472), 1215–1225.

Zhao, Y., Y. Amemiya, and Y. Hung (2018). Efficient gaussian process modeling using experimental design-based subagging. *Statistica Sinica 28*(3), 1459–1479.

Zhou, Q., P. Z. Qian, and S. Zhou (2011). A simple approach to emulation for computer models with qualitative and quantitative factors. *Technometrics 53*(3), 266–273.