Taylor & Francis
Taylor & Francis Group

Journal of Quality Technology
A Quarterly Journal of Methods, Applications and Related Topics

ISSN: 0022-4065 (Print) 2575-6230 (Online) Journal homepage: https://www.tandfonline.com/loi/ujqt20

Variable selection for kriging in computer experiments

Hengzhen Huang, Dennis K. J. Lin, Min-Qian Liu & Qiaozhen Zhang

To cite this article: Hengzhen Huang, Dennis K. J. Lin, Min-Qian Liu & Qiaozhen Zhang (2020) Variable selection for kriging in computer experiments, Journal of Quality Technology, 52:1, 40-53, DOI: 10.1080/00224065.2019.1569959

To link to this article: https://doi.org/10.1080/00224065.2019.1569959

JOURNAL OF QUALITY TECHNOLOGY
2020, VOL. 52, NO. 1, 40–53
https://doi.org/10.1080/00224065.2019.1569959

Variable selection for kriging in computer experiments

Hengzhen Huang[a], Dennis K. J. Lin[b], Min-Qian Liu[c], and Qiaozhen Zhang[c]

[a]College of Mathematics and Statistics, Guangxi Normal University, Guilin, China; [b]Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania; [c]School of Statistics and Data Science, LPMC & KLMDASR, Nankai University, Tianjin, China

ABSTRACT

An efficient variable selection technique for kriging in computer experiments is proposed. Kriging models are popularly used in the analysis of computer experiments. The conventional kriging models, the *ordinary kriging,* and *universal kriging* could lead to poor prediction performance because of their prespecified mean functions. Identifying an appropriate mean function for kriging is a critical issue. In this article, we develop a Bayesian variable-selection method for the mean function and the performance of the proposed method can be guaranteed by the convergence property of Gibbs sampler. A real-life application on piston design from the computer experiment literature is used to illustrate its implementation. The usefulness of the proposed method is demonstrated via the practical example and some simulative studies. It is shown that the proposed method compares favorably with the existing methods and performs satisfactorily in terms of several important measurements relevant to variable selection and prediction accuracy.

KEYWORDS

Bayesian variable selection; experimental design; Gaussian process; Gibbs sampler; prediction

1. Introduction

As science and technology have advanced to a higher level, computer experiments are becoming increasingly prevalent surrogates for physical experiments because of their economy. A computer experiment, however, can still be time consuming and costly. One primary goal of a computer experiment is to build an inexpensive metamodel that approximates the original expensive computer model well. The kriging model first proposed by Sacks et al. (1989) in computer experiments is desirable because of its convenience, flexibility, and broad generality.

A kriging model mainly contains two parts: a mean function and a stationary Gaussian process (GP). Typically, a kriging model either uses a constant as the mean function (known as the *ordinary kriging*, OK) or assumes some prespecified variables in the mean function (known as the *universal kriging*, UK). A computer experiment usually contains a large number of input variables, so it is important to identify those (relatively few) variables having significant impact on the response. The idea of variable-selection for computer experiments is initialized by Welch et al. (1992). In their landmark work, attention was focused on selecting variables with significant impact on the GP while keeping the mean function as a constant. Their strategy can be used as a screening tool in the early stage of experimentation. Following the spirit of Welch et al. (1992), Li and Sudjianto (2005) developed a penalized likelihood method and Linkletter et al. (2006) developed a Bayesian method to screen variables with significant impact on the GP. After the screening stage, a more elaborate kriging model with a complex mean function is desirable.

The focus here is on the next stage—how to appropriately select variables (terms) for the mean function of kriging. It is noteworthy that the aim of selecting the mean function of kriging is to establish an initial model with high quality, which would facilitate advanced analysis such as the sensitivity analysis, model calibration, model validation, and heuristic optimization. The popular initial model is the OK model, and sometimes a UK model with some prespecified variables in the mean function is adopted as the initial model. Some literature (Hung 2011; Joseph, Hung, and Sudjianto 2008; Martin and Simpson 2005), however, has shown that both the OK and UK can be underperformed in terms of prediction accuracy when strong trends exist. The main reason is that a constant mean in the OK may be inadequate to capture the overall trend, while the prespecified mean

CONTACT Min-Qian Liu mqliu@nankai.edu.cn School of Statistics and Data Science, Nankai University, Tianjin 300071, P. R. China; Hengzhen Huang hhzgxnu@126.com College of Mathematics and Statistics, Guangxi Normal University, Guilin, 541004, China.

function in the UK may contain some redundant predictors. As a result, identifying an appropriate mean function is crucial for building an accurate kriging model. The existing literature suggests two major approaches to accomplish this purpose: the blind kriging (BK) method, which uses a Bayesian forward selection procedure (Joseph, Hung, and Sudjianto 2008); and the penalized blind kriging (PBK) method, which uses the penalized likelihood approach for variable selection (Hung 2011). These methods first identify mean functions as a variable selection problem and then the built initial models can effectively reduce the prediction error. The BK and PBK also recommended that one should treat the mean function and the GP as an integration rather than dealing with them separately. Both methods, however, suffer from their own limitations: the theoretical properties of the BK method are difficult to derive, while the appealing oracle property of the PBK method only holds when the design forms a complete lattice. When the experimental resources are limited, a complete lattice design is typically infeasible. In this article, we will focus on variable selection for the mean function. Although there are some universal approaches to simultaneously select the active variables for both the GP and the mean function (see, for example, Marrel et al. 2008), as will be seen, these universal approaches provide a preliminary basis for selecting the mean function.

This article develops a new variable selection method, the stochastic searched blind kriging (SSBK), for identifying the proper mean function of kriging. The SSBK is a Bayesian approach, which allows the number of candidate variables to be much larger than the number of runs. The SSBK fully accounts for the spatial correlation that is inherent to a kriging model when selecting the mean function. It is shown that correctly accounting for the correlation has an impact on the variable-selection result. Moreover, the proposed method is rational and flexible in the sense that it possesses the Markov convergence property under a wide class of experimental designs and, hence, does not suffer from the limitations of the BK and PBK methods.

The remainder of this article is organized as follows. Section 2 introduces a hierarchical kriging model that provides the basis for the SSBK. Section 3 presents a step-by-step algorithm for the SSBK and describes its implementation details. In Section 4, a real-life application on piston design analysis (Hoffman et al. 2003) is used to illustrate the implementation and performance of the SSBK. Some simulation studies are conducted in Section 5 to further demonstrate the usefulness of the SSBK. Finally, concluding remarks and further discussions are given in Section 6.

2. A hierarchical kriging model

In this section, we first describe the UK model, then a hierarchical setting for variable selection in which the UK model details are presented. The hierarchical setting introduced here is the same as in Le and Zidek (2006), with an exception that the idea of variable selection is incorporated. To make a better presentation, a concise description without loss of important elements is given below. Note that the hierarchical setting introduced here is different from that of Han and Görtz (2012) because multi-fidelity experiments are beyond the scope of this article. As alluded to earlier, however, our method can be used to establish high-quality initial models for this kind of experiment.

2.1. The universal kriging model

Suppose the input vector $\mathbf{x} \in \mathcal{X}, \mathcal{X} \subset \mathbb{R}^p$ is the design region with p distinct factors and $\mathcal{C} = \{f_1(\mathbf{x}), ..., f_k(\mathbf{x})\}$ is the set containing all candidate variables (not including the intercept) for the mean function, where k is the number of candidate variables that are functions (e.g., linear and quadratic effects and two-factor interactions) of the p factors. The UK model considered here at the screening stage for the mean function assumes that the output of a computer code $Y(\mathbf{x})$ is represented by

$$Y(\mathbf{x}) = \sum_{i=1}^{k} \mu_i f_i(\mathbf{x}) + Z(\mathbf{x}), \qquad [1]$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)^T$ is the vector of unknown regression parameters and $Z(\mathbf{x})$ is a centered stationary GP. Given any two points $\mathbf{x}_i$ and $\mathbf{x}_j$, the covariance of $Y(\mathbf{x}_i)$ and $Y(\mathbf{x}_j)$ is given by $\text{cov}[Y(\mathbf{x}_i), Y(\mathbf{x}_j)] = \text{cov}[Z(\mathbf{x}_i), Z(\mathbf{x}_j)] = \sigma^2 R(\mathbf{x}_i, \mathbf{x}_j)$, where $R(\mathbf{x}_i, \mathbf{x}_j)$ is the correlation function and σ^2 is the process variance. The most commonly used correlation function is perhaps the power exponential correlation function:

$$R(\mathbf{x}_i, \mathbf{x}_j) = \prod_{u=1}^{p} \rho_u^{|x_{iu} - x_{ju}|^{a_u}}, \qquad [2]$$

where x_{iu} is the uth element of $\mathbf{x}_i, 0 < \rho_u < 1$ and $0 < a_u \leq 2, u = 1, ..., p$. Throughout this article, we take $a_u = 2$ for all $u = 1, 2, ..., p$. Taking a_u's $= 2$ is suggested by a number of works (cf., Chen et al. 2016;

Haaland and Qian 2011; Han, Santner, and Rawlinson 2009; Huang et al. 2016, 2018; Kennedy and O'Hagan 2000; Li and Sudjianto 2005; Linkletter et al. 2006; Qian and Wu 2008; Zhou, Qian, and Zhou 2011; Ranjan, Haynes, and Karsten 2011; Tan and Wu 2012). The case a_u's $= 2$ gives a process with infinitely differentiable paths and is useful when the response is analytic. Other choices may also be used, for example, Hung (2011) suggested using a_u's $,= 1$ which helps with Markovian properties for variable selection.

Suppose the output is observed at n values of the input factors given by the rows of $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$. This yields a vector $\mathbf{Y} = (Y(\mathbf{x}_1), ..., Y(\mathbf{x}_n))^T$ of observed outputs. In general, not all the variables in $\mathcal{C}$ have significant impacts on the output. If the set of variables $\{\omega_1(\mathbf{x}), ..., \omega_m(\mathbf{x})\} \subset \mathcal{C}$ is found to be significant, then the following model, which includes the intercept term, is considered for subsequent analysis:

$$Y(\mathbf{x}) = \beta_0 + \sum_{i=1}^{m} \beta_i \omega_i(\mathbf{x}) + Z(\mathbf{x}). \quad [3]$$

Under Eq. [3], the best linear unbiased predictor (BLUP) at an untried site $\mathbf{x}^*$ is

$$\hat{y}(\mathbf{x}^*) = \boldsymbol{\omega}(\mathbf{x}^*)^T \hat{\boldsymbol{\beta}} + \mathbf{r}(\mathbf{x}^*)^T \mathbf{R}^{-1} \left(\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\beta}}\right), \quad [4]$$

where $\boldsymbol{\omega}(\mathbf{x}^*) = (1, \omega_1(\mathbf{x}^*), ..., \omega_m(\mathbf{x}^*))^T$, $\hat{\boldsymbol{\beta}} = (\mathbf{F}^T\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{R}^{-1}\mathbf{Y}$, $\mathbf{R} = (R(\mathbf{x}_i, \mathbf{x}_j))_{n\times n}$, $\mathbf{r}(\mathbf{x}^*) = (R(\mathbf{x}^*, \mathbf{x}_1), ..., R(\mathbf{x}^*, \mathbf{x}_n))^T$, and $\mathbf{F} = (\boldsymbol{\omega}(\mathbf{x}_1), ..., \boldsymbol{\omega}(\mathbf{x}_n))^T$. In Eq. [4], $\mathbf{R}$, $\mathbf{r}$, and $\hat{\boldsymbol{\beta}}$ depend on the correlation function $R(\cdot)$, which depends on the vector of correlation parameters $\boldsymbol{\rho} = (\rho_1, ..., \rho_p)^T$. Both the maximum-likelihood method and Bayesian method are frequently used to estimate the parameter vector $\boldsymbol{\rho}$ (see, e.g., Fang, Li, and Sudjianto 2006; Santner, Williams, and Notz 2003).

2.2. Specifying the priors for variable selection

As previously discussed, some of the k variables in $\mathcal{C}$ may be negligible. This can be captured by augmenting the mean function of the UK of Eq. [1] with a k-indicator vector $\boldsymbol{\delta}$ of 0's and 1's: 0 corresponds to a negligible variable and 1 corresponds to an active one. This idea is motivated from the multiple linear regression context (Chen et al. 2011; George and McCulloch 1993). However, the spatial correlations among the observations that are inherent to a kriging model will be taken into account.

First, we treat Eq. [1] as the prior distribution for the true model, that is, the output vector $\mathbf{Y}$ follows a two-stage hierarchical model whose first stage is given by

$$[\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{\rho}] \sim N_n(\tilde{\mathbf{F}}\boldsymbol{\mu}, \sigma^2\mathbf{R}), \quad [5]$$

where N_n denotes an n-dimensional multivariate normal distribution, $\tilde{\mathbf{F}} = (\tilde{\mathbf{f}}(\mathbf{x}_1), ..., \tilde{\mathbf{f}}(\mathbf{x}_n))^T$ with $\tilde{\mathbf{f}}(\mathbf{x}) = (f_1(\mathbf{x}), ..., f_k(\mathbf{x}))^T$. Second, we place a prior on $[\boldsymbol{\mu}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{\rho}]$. Here, the prior on $\boldsymbol{\mu}$ depends on $\boldsymbol{\delta}$ and σ^2, and the priors on $\boldsymbol{\delta}, \sigma^2$ and $\boldsymbol{\rho}$ are independent, that is,

$$[\boldsymbol{\mu}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{\rho}] = [\boldsymbol{\mu}|\boldsymbol{\delta}, \sigma^2][\boldsymbol{\delta}][\sigma^2][\boldsymbol{\rho}]. \quad [6]$$

By introducing the indicator variable $\delta_i = 0$ or 1, an independent normal mixture prior is placed on each component of $\boldsymbol{\mu}$:

$$[\mu_i|\delta_i, \sigma^2] \sim (1-\delta_i)N(0, \sigma^2\tau_i^2) + \delta_i N(0, c_i^2\sigma^2\tau_i^2), \text{ for } i = 1, ..., k. \quad [7]$$

The value of τ_i is set to be small so that, when $\delta_i = 0$, μ_i is tightly centered on 0 and will not have a large effect. The much larger variance (by taking $c_i \gg 1$) when $\delta_i = 1$ allows the higher possibility of a large μ_i. As in Box and Meyer (1986), we set $c_1 = \cdots = c_k = 10$. This indicates that an important effect has an order of magnitude larger than an unimportant one. The choice of τ_i follows the fact that a small coefficient has a standard deviation $\sigma\tau_i$ and will lie within $0 \pm 3\sigma\tau_i$ with a very high probability. Thus, setting $\tau_i = 1/(3 \times \text{range}(f_i(\mathbf{x}))$ implies that even a large change in $f_i(\mathbf{x})$, as compared with range$(f_i(\mathbf{x}))$, will result in a change in $Y(\mathbf{x})$ no more than σ.

The prior on $\boldsymbol{\delta}$ reflects a prior on the mean function. A simple choice might have the elements of $\boldsymbol{\delta}$ being independent, so that $[\boldsymbol{\delta}] = \prod_{i=1}^{k}[\delta_i]$. When no information is available, a uniform prior $\Pr(\delta_i = 1) = \Pr(\delta_i = 0) = 0.5$ is chosen for each i. The prior for σ^2 is chosen to be a noninformative one $[\sigma^2] \propto 1/\sigma^2$. Finally, each component of $\boldsymbol{\rho}$ is given an independent standard uniform prior, that is, $\rho_i \sim U(0, 1)$ for each i.

3. A Bayesian variable selection method for the mean function of kriging

For the hierarchical setup previously introduced, the posterior distribution of $\boldsymbol{\delta}$, denoted by $[\boldsymbol{\delta}|\mathbf{Y}]$, contains the information relevant to variable selection. Based on the output vector $\mathbf{Y}$, the posterior $\boldsymbol{\delta}$ updates the prior probabilities on each of the 2^k possible values of $\boldsymbol{\delta}$. Identifying a mean function with each $\boldsymbol{\delta}$ via $(\delta_i = 1) \Longleftrightarrow (f_i(\mathbf{x})$ is included), those $\boldsymbol{\delta}$'s with higher posterior probabilities identify the promising mean functions. However, it is usually intractable to derive a closed form for $[\boldsymbol{\delta}|\mathbf{Y}]$; thus, how to extract

information from $[\boldsymbol{\delta}|\mathbf{Y}]$ is an important issue. In this section, we first demonstrate that posterior samples from $[\boldsymbol{\delta}|\mathbf{Y}]$ can be conveniently generated by the Gibbs sampler and present a step-by-step algorithm for variable selection. The implementation details surrounding the algorithm are then described.

3.1. An algorithm of the stochastic searched blind kriging

Based on Eqs. [5] through [7] and the priors on $\boldsymbol{\delta}, \sigma^2$ and $\boldsymbol{\rho}$, the following results can be obtained in the usual way. For clarity, the proof is deferred to the Appendix.

Proposition 1. *The full conditional distributions of* $\boldsymbol{\mu}, \sigma^2$, δ_i $(i = 1, \ldots, k)$ *and* $\boldsymbol{\rho}$ *can be expressed as follows:*

$$\left[\boldsymbol{\mu}|\sigma^2, \boldsymbol{\delta}, \boldsymbol{\rho}, \mathbf{Y}\right] \sim N_k(\mathbf{Av}, \mathbf{A}), \tag{8}$$

$$\left[\sigma^2|\boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\rho}, \mathbf{Y}\right] \sim \mathrm{IG}\left(\frac{n+k}{2}, \frac{(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu}) + \boldsymbol{\mu}^T\mathbf{D}^{-1}\boldsymbol{\mu}}{2}\right), \tag{9}$$

$$\left[\delta_i = x|\boldsymbol{\delta}_{(-i)}, \boldsymbol{\mu}, \sigma^2, \boldsymbol{\rho}, \mathbf{Y}\right] = \left(\frac{a}{a+b}\right)^x \left(\frac{b}{a+b}\right)^{1-x}, \quad \text{for } i = 1, \ldots, k, x = 0 \text{ or } 1, \tag{10}$$

$$\left[\boldsymbol{\rho}|\boldsymbol{\mu}, \sigma^2, \boldsymbol{\delta}, \mathbf{Y}\right] \propto |\mathbf{R}|^{-\frac{1}{2}} \exp\left\{-\frac{(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu})}{2\sigma^2}\right\} I_{(0,1)^k}(\boldsymbol{\rho}), \tag{11}$$

where IG *denotes an inverse gamma distribution,* $\mathbf{A} = \sigma^2(\tilde{\mathbf{F}}^T\mathbf{R}^{-1}\tilde{\mathbf{F}} + \mathbf{D}^{-1})^{-1}$ *with* $\mathbf{D} = \mathrm{diag}\{(c_1^{\delta_1}\tau_1)^2, \ldots, (c_k^{\delta_k}\tau_k)^2\}$, $\mathbf{v} = \sigma^{-2}\tilde{\mathbf{F}}^T\mathbf{R}^{-1}\mathbf{Y}$, $a = [\boldsymbol{\mu}|\delta_i = 1, \boldsymbol{\delta}_{(-i)}, \sigma^2, \mathbf{Y}]$, $b = [\boldsymbol{\mu}|\delta_i = 0, \boldsymbol{\delta}_{(-i)}, \sigma^2, \mathbf{Y}]$, $(-i)$ *denotes all of the components except for the ith one, and* $I_A(\cdot)$ *is the indicator function for set* A *with* $I_A(x) = 1$ *if* $x \in A$ *and* $I_A(x) = 0$ *otherwise.*

From Proposition 1, the prior distributions provide the regularization that the matrix $\mathbf{A}$ may be invertible even if $\tilde{\mathbf{F}}$ is not of full column rank. With an initial value $(\boldsymbol{\mu}^{(0)}, \sigma^{2^{(0)}}, \boldsymbol{\delta}^{(0)}, \boldsymbol{\rho}^{(0)})$, the Gibbs sampler, which is a well-known Markov chain Monte Carlo (MCMC) method (Shao 2003), can be used to generate the posterior samples of $\boldsymbol{\mu}, \sigma^2, \boldsymbol{\delta}$ and $\boldsymbol{\rho}$ by repeated successive sampling from Eqs. [8] through [11]. Because we focus on variable selection for the mean function, the posterior samples of $\boldsymbol{\delta}$ are of particular interest. It is known (e.g., Theorem 4.4 of Shao 2003) that, under regularity conditions, the Markov chain generated by the Gibbs sampler is *aperiodic*, *invariant*, *irreducible*, *Harris recurrent* and, hence, converges to a limiting distribution. This property leads to the fact that, with the length of the chain increasing, the empirical distribution of the realized values of $\boldsymbol{\delta}$ converge to the posterior distribution $[\boldsymbol{\delta}|\mathbf{Y}]$, which contains the information relevant to variable selection. In other words, decisions about the selection of the mean functions can be made based on the posterior samples of $\boldsymbol{\delta}$. Specifically, if a sample $\boldsymbol{\delta}^*$ appears in the posterior samples with a high frequency in a relative sense, the corresponding mean function could be promising. This is because, once the chain generated by the Gibbs sampler mixes well, the value with a relatively large probability under $[\boldsymbol{\delta}|\mathbf{Y}]$ is likely to be realized with a relatively high frequency. Of course, there may be several values realized with relatively high frequencies and all of the corresponding mean functions should be kept for further investigations. As previously mentioned, the matrix $\tilde{\mathbf{F}}$ in Proposition 1 is not necessarily of full rank, which implies that a wide class of experimental designs can be used for the Gibbs sampler. For the sake of convenience, a step-by-step variable selection algorithm for the mean function of kriging is formulated as follows.

Algorithm A (Stochastic searched blind kriging (SSBK)).

Step 1. Identify all the input factors and the set of candidate variables $\mathcal{C}$.

Step 2. Initialize $\boldsymbol{\mu}^{(0)}, \sigma^{2^{(0)}}, \boldsymbol{\delta}^{(0)}, \boldsymbol{\rho}^{(0)}$, and $j_{\max}$ (the maximum number of iterations). Set $j = 1$.

Step 3. Sample $\boldsymbol{\mu}^{(j)}$ from the multivariate normal distribution $[\boldsymbol{\mu}^{(j)}|\sigma^{2^{(j-1)}}, \boldsymbol{\delta}^{(j-1)}, \boldsymbol{\rho}^{(j-1)}, \mathbf{Y}]$ given by the right-hand side of Eq. [8].

Step 4. Sample $\boldsymbol{\sigma}^{2^{(j)}}$ from the inverted gamma distribution $[\boldsymbol{\sigma}^{2^{(j)}}|\boldsymbol{\mu}^{(j)}, \boldsymbol{\delta}^{(j-1)}, \boldsymbol{\rho}^{(j-1)}, \mathbf{Y}]$ given by the right-hand side of Eq. [9].

Step 5. Sample $\delta_i^{(j)}$ from the Bernoulli distribution $[\delta_i^{(j)} = x|\boldsymbol{\delta}_{(-i)}^{(j)}, \boldsymbol{\mu}^{(j)}, \sigma^{2^{(j)}}, \boldsymbol{\rho}^{(j-1)}, \mathbf{Y}]$, whose density is given by the right-hand side of formula [10], for $i = 1, \ldots, k$, $x = 1$ or 0, where $\boldsymbol{\delta}_{(-i)}^{(j)} = (\delta_1^{(j)}, \ldots, \delta_{i-1}^{(j)}, \delta_{i+1}^{(j-1)}, \ldots, \delta_k^{(j-1)})$.

Step 6. Sample $\boldsymbol{\rho}^{(j)}$ from the distribution $[\boldsymbol{\rho}^{(j)}|\boldsymbol{\mu}^{(j)}, \sigma^{2^{(j)}}, \boldsymbol{\delta}^{(j)}, \mathbf{Y}]$, whose density up to a normalization constant is given by the right-hand side of formula [11]. Set $j = j + 1$.

Step 7. If $j > j_{\max}$, stop; otherwise, return to *Step 3*.*Step 8.* Identify the promising mean functions based on $\boldsymbol{\delta}^{(0)}, \boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(j_{\max})}$.

A similar algorithm is called the stochastic search variable selection in the regression context. Because Algorithm 1 focuses on the variable selection problem

for the mean function of kriging, we call it the stochastic searched blind kriging (SSBK).

3.2. Details for implementation of the SSBK

In *Step 1*, the input factors and the set of candidate variables $\mathcal{C}$ are usually subjectively identified based on the experimenters' experience. Sometimes, one might be able to use information from preliminary screening (e.g., Marrel et al. 2008; Morris 1991; Welch et al. 1992) to help identify the input factors and the set $\mathcal{C}$. In such a case, the SSBK serves as a follow-up/enhanced screening procedure.

In *Step 2*, the initial values of $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\delta}^{(0)}$ are set to be zeros; that is, the SSBK starts from a model without any active effects in the mean function. The estimates of σ^2 and $\boldsymbol{\rho}$ obtained from the OK (denoted by $\hat{\sigma}^2_{\text{OK}}$ and $\hat{\boldsymbol{\rho}}_{\text{OK}}$) are used as $\sigma^{2^{(0)}}$ and $\boldsymbol{\rho}^{(0)}$, respectively. The $j_{\max}$, which is the maximum number of iterations for the Gibbs sampler, should be large enough so that convergence could be achieved. There are a variety of methods to detect the convergence of a Gibbs sequence. In Sections 4 and 5, $j_{\max}$ is set to be 100,000 and the ratio of the *Monte Carlo standard error* (MCSE) to the corresponding posterior sample mean is used to judge whether 100,000 iterations are sufficient (the computation issues of MCSE can be found in the Appendix). The examples in Sections 4 and 5 show that 100,000 iterations (thinned by discarding the first 10,000 samples and every four samples and retaining the next one in the sequence) result in a reasonably small ratio of the MCSE to the corresponding posterior sample mean.

Step 3 to *Step 7* are straightforward. Note that, in *Step 6*, the right-hand side of Eq. [11] only gives $[\boldsymbol{\rho}|\boldsymbol{\mu}, \sigma^2, \boldsymbol{\delta}, \mathbf{Y}]$ up to a normalization constant, which is typically difficult to obtain. One popular approach to solving this sampling problem is to insert the "slice sampling" into this step (Neal 2003). The slice sampling algorithm is another MCMC method that can be used to generate samples from a distribution, and it is also ergodic under some weak conditions. One appealing feature of the slice sampling is that it only requires the density function up to a normalization constant and thus can deal with the sampling problem encountered in Eq. [11]. Some other sampling techniques that only require the density function up to a normalization constant, such as the Metropolis-Hastings algorithm (Hastings 1970), can be adopted to deal with Eq. [11] as well. Although the distribution $[\boldsymbol{\mu}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{\rho}|\mathbf{Y}]$ (up to a normalization constant) may be obtained in the usual way, we do not recommend using the slice sampling to deal with it directly because of its high dimensionality. Also of note is that *Steps 4* and *6* can be modified if other correlation functions are used. In other words, the correlation matrix $\mathbf{R}$ appearing on the right-hand sides of Eqs. [9] and [11] is of a general form.

In *Step 8*, those mean functions that appear in $\boldsymbol{\delta}^{(0)}, \boldsymbol{\delta}^{(1)}, \ldots, \boldsymbol{\delta}^{(j_{\max})}$ with large frequencies (the top five, say) shall be identified as the promising ones. Sometimes, an "optimal" mean function needs to be determined, then some model selection criteria can be adopted. In this article, we use the *cross-validation prediction error* (CVPE) and *root-mean square prediction error* (RMSPE) criteria, which are popular for computer experiments, to further discriminate the promising mean functions. The CVPE criterion selects the mean function that minimizes

$$\text{CVPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y(\mathbf{x}_i) - \hat{y}_{(-i)}(\mathbf{x}_i)\right)^2},$$

where $\hat{y}_{(-i)}(\mathbf{x}_i)$ is the BLUP (see Eq. [4]) of the corresponding kriging model after removing the ith data point. If n_0 testing points $\mathbf{x}_1^*, \ldots, \mathbf{x}_{n_0}^*$ are available, the RMSPE evaluates

$$\text{RMSPE} = \sqrt{\frac{1}{n_0}\sum_{i=1}^{n_0}\left(y(\mathbf{x}_i^*) - \hat{y}(\mathbf{x}_i^*)\right)^2}.$$

Then the mean function that minimizes the RMSPE value can be selected. We will use the RMSPE criterion as the primary criterion, although other criteria can be used in a similar manner.

Remark 1. The computational cost of Algorithm 1 is dominated by *Steps 3* and *6* because they involve computation of the inverse of the information matrix and the correlation matrix. In this work, the Cholesky factorization algorithm is used to calculate the inverse of matrices. This implies that the computational complexity of the inverse of the information matrix and that of the correlation matrix are $O(k^3)$ and $O(n^3)$, respectively (Fang, Li, and Sudjianto 2006). Hence the computational complexity of Algorithm 1 is $O(k^3) + O(n^3)$ (j_{max} is taken to be 100,000 throughout this work), which could be time-consuming when k and/or n are not small. Single-site Gibbs sampling schemes (e.g., those proposed by Chang, Chen, and Chi 2016; Chen et al. 2011, 2013; and Huang et al. 2017), which sample the variables in a component-wise manner so that calculation of the inverse of matrices can be avoided, provide potential alleviation to the computational complexity and are appropriate topics for future research.

4. A practical example

This section applies the proposed SSBK to analyze a practical computer experiment. The ultimate goal of the study is to perform robust design optimization (Hoffman et al. 2003). To accomplish this goal, the availability of a good initial model is a necessity. Li and Sudjianto (2005) and Fang, Li, and Sudjianto (2006) adopted the OK model as the initial model. Joseph, Hung, and Sudjianto (2008) considered using the BK to build an initial model that is more accurate than the OK model. The objectives here are to illustrate the implementation of the SSBK and compare the performance of the SSBK with that of the BK, OK, and UK models. It turns out that the SSBK successfully identifies initial models that are more promising.

Piston slap is an unwanted engine noise caused by piston secondary motion. That is, the departure of a piston from the nominal motion prescribed by the slider crank mechanism. A computer experiment was performed by varying six factors to minimize the piston slap noise. The factors are set clearance between the piston and the cylinder liner (x_1), location of peak pressure (x_2), skirt length (x_3), skirt profile (x_4), skirt ovality (x_5), and pin offset (x_6). Because each computer experimental run requires intensive computational resources (because the power cylinder system is modeled using the multibody dynamics code ADAMS/Flex including a finite-element model and it takes 24 h for each run), a uniform design was employed to plan a computer experiment with 12 runs. The experimental design and the response data are displayed in Table 1. Note that this uniform design is a space-filling design where different factors may have different number of levels. For more details on the theory of uniform designs and their applications in computer experiments, refer to Fang et al. (2000, 2006).

Table 1. Piston slap noise data.

Run	x_1	x_2	x_3	x_4	x_5	x_6	y
1	71.00	16.80	21.00	2.00	1.00	0.98	56.75
2	15.00	15.60	21.80	1.00	2.00	1.30	57.65
3	29.00	14.40	25.00	2.00	1.00	1.14	53.97
4	85.00	14.40	21.80	2.00	3.00	0.66	58.77
5	29.00	12.00	21.00	3.00	2.00	0.82	56.34
6	57.00	12.00	23.40	1.00	3.00	0.98	56.85
7	85.00	13.20	24.20	3.00	2.00	1.30	56.68
8	71.00	18.00	25.00	1.00	2.00	0.82	58.45
9	43.00	18.00	22.60	3.00	3.00	1.14	55.50
10	15.00	16.80	24.20	2.00	3.00	0.50	52.77
11	43.00	13.20	22.60	1.00	1.00	0.50	57.36
12	57.00	15.60	23.40	3.00	1.00	0.66	59.64

4.1. Implementing the SSBK

We next illustrate the steps in Algorithm A.

[*Step 1*] Following Joseph, Hung, and Sudjianto (2008), the candidate set $\mathcal{C}$ consists of all the linear main effects, quadratic main effects, and all the two-factor interactions under the orthogonal polynomial coding (Wu and Hamada 2009) so that there are 72 variables in $\mathcal{C}$. Here the two-factor interactions include the linear-by-linear, linear-by-quadratic and quadratic-by-quadratic interactions. For $j = 1, \ldots, 6$, we use the notations x_{jl} and x_{jq} to denote the linear main-effect and quadratic main-effect of x_j, respectively. The two-factor interactions can then be defined as the products of these variables. For example, the linear-by-quadratic interaction between x_1 and x_3 is $x_{1l}x_{3q}$.

[*Step 2*] Set $\boldsymbol{\mu}^{(0)} = 0, \sigma^{2^{(0)}} = 3.7322, \boldsymbol{\delta}^{(0)} = 0, \boldsymbol{\rho}^{(0)} = (0.31, 0.99, 0.79, 0.99, 0.99, 0.49)$ and $j_{max} = 10,000$. The initial values for $\sigma^{2^{(0)}}$ and $\boldsymbol{\rho}^{(0)}$ follow the maximum-likelihood estimators derived in Joseph, Hung, and Sudjianto (2008).

[*Step 3–Step 7*] These steps can be conducted in a mechanistic fashion using the results in Proposition 1.

[*Step 8*] Those mean functions with the top five highest frequencies are selected as the promising ones (the remaining ones have small frequencies no more than 0.01 for this example, hence we discard them).

4.2. The results

The maximum ratio of the MCSE to the corresponding posterior sample mean among the 72 variables in $\mathcal{C}$ is 0.0362, which appears to be sufficiently small. So it is assurable that $j_{max} = 100,000$ (thinned by discarding the first 10,000 samples and every four samples and retaining the next one in the sequence) is long enough. The selection results of the SSBK are displayed in Table 2. In addition, the performances of the BK, OK, and UK (reported in Joseph, Hung, and Sudjianto 2008) are displayed as well. In Table 2, the column labeled "Model" represents the selected effects in the mean function; the column labeled "Freq" represents the frequency of the corresponding mean function in the posterior samples of $\boldsymbol{\delta}$; the column labeled "CVPE ($\hat{\rho}$)" represents the CVPE value of the corresponding model and the maximum-likelihood estimate of $\boldsymbol{\rho}$ in the parentheses; because an additional 100 testing points are available from the website http://www.personal.psu.edu/ril4/DMCE/MatlabCode/, the RMSPE values of the corresponding models can be calculated, which are listed in the last column. Three values in the "Freq" column are denoted as N/A because they cannot be calculated. The terms in the

Table 2. Summary of the selection and prediction results in the piston slap noise example.

Method	Model	Freq.	CVPE($\hat{\rho}$)	RMSPE
SSBK	$x_{1l}, x_{4q}, x_{1l}x_{5q}, x_{1l}x_{6l}$	0.0700	0.6506(0.21,0.99,0.90,0.99,0.98,0.12)	0.9744
	$x_{1l}, x_{2q}, x_{4q}, x_{6q}, x_{2q}x_{3l}, x_{2q}x_{3q}$	0.0315	0.7284(0.02,0.91,0.99,0.94,0.38,0.38)	2.0019
	$x_{1l}, x_{1l}x_{5q}, x_{1l}x_{6l}$	0.0305	1.3629(0.21,0.99,0.61,0.99,0.99,0.15)	1.4349
	$x_{1l}, x_{2q}, x_{4q}, x_{1l}x_{5q}, x_{1l}x_{6l}$	0.0250	0.9631(0.42,0.99,0.90,0.99,0.99,0.12)	0.9999
	$x_{1l}, x_{2q}, x_{4q}, x_{1l}x_{5q}$	0.0200	0.9882(0.42,0.99,0.90,0.99,0.99,0.12)	1.0805
BK	$x_{1l}, x_{1l}x_{6l}, x_{1q}x_{6l}$	N/A	1.1168(0.99,0.99,0.91,0.27,0.99,0.63)	1.0038
OK		N/A	1.4511(0.31,0.99,0.79,0.99,0.99,0.49)	1.3626
UK (linear)	$x_{1l}, x_{2l}, x_{3l}, x_{4l}, x_{5l}, x_{6l}$	N/A	1.1607(0.87,0.99,0.84,0.99,0.99,0.91)	1.5109

The CPU time of implementing the SSBK procedure for this example is about 3,409 s on a PC with 2.10 GHz triple-core AMD Phenom II N830 CPU and 8 GB memory.

"Model" column are selected by the competing methods, while the values in the "CVPE ($\hat{\rho}$)" and "RMSP" columns are calculated using *The DACE Toolbox* developed by Zhou, Qian, and Zhou (2011) in Matlab. It is interesting to observe that all the identified models in the "Model"column follow the three key variable selection principles, that is, effect hierarchy, effect sparsity, and effect heredity.

4.3. Comparisons

The SSBK identifies five promising models with the top five highest frequencies. The best model, in terms of the RMSPE value, is the model $x_{1l}, x_{4q}, x_{1l}x_{5q}, x_{1l}x_{6l}$, which has the highest frequency. Compared with the model $x_{1l}, x_{1l}x_{6l}, x_{1q}x_{6l}$ identified by the BK, the best model identified by the SSBK has smaller RMSPE and CVPE values. That is, the SSBK successfully identifies a more accurate model in terms of both the RMSPE and CVPE criteria for this case study. It is worth noting the model with the fourth highest frequency is also better than that identified by the BK in terms of both the RMSPE and CVPE criteria. This strongly recommends that a handful of models that have relatively high frequencies should be checked for further investigation. Neither the OK nor UK models perform well in this study.

5. Simulation studies

To further judge the value of the SSBK, some simulative experiments with known functions are conducted. Example 1 considers three common cases with different complexities so that the flexibility of the SSBK can be assessed. Example 2 revisits the simulation study in Hung (2011), where the PBK method was adopted. The simulation study in Marrel et al. (2008) is revisited in Example 3. The purposes of including Examples 2 and 3 are to compare the SSBK with the PBK and the methodology of Marrel et al (2008), respectively. As will be seen, the SSBK performs satisfactorily in terms of several important simulation measurements and compares favorably with the competitors. The performances of the frequently used OK and UK models are also evaluated as references. As expected, both the OK and UK models did not perform satisfactorily when the responses are believed to have large trends.

Example 1. We generate the simulative data from the following three models:

$$\text{Model I:} \quad y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + Z(\mathbf{x});$$

$$\text{Model II:} \quad y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1^2 + b_5x_2^2 + b_6x_3^2 + Z(\mathbf{x});$$

$$\text{Model III:} \quad y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1^2 + b_5x_2^2 + b_6x_3^2 + b_7x_1x_2 + b_8x_1x_3 + b_9x_2x_3 + Z(\mathbf{x}),$$

where $\mathbf{x} = (x_1, \ldots, x_{10})^T$. For each model, the regression coefficients are selected at random from [–20, –10] ∪ [10, 20] to represent large trends in the mean function, and $Z(\mathbf{x})$ is assumed to be a centered stationary GP whose correlation function is given by Eq. [2] with ρ_i generated from $U(0, 1)$ and $a_i = 2$ for $i = 1, \ldots, 10$. The process variance σ^2 is set to be one. For each model, three factors (x_1, x_2, and x_3), are assigned as active ones in the mean functions. However, the three models have different complexity: Model I considers the case where linear main effects are active; Model II considers the case where the linear main effects and quadratic main effects are active; while Model III considers the case where not only the main effects but also all of the two-factor interactions among x_1, x_2 and x_3 are active. Without loss of generality, the design region is taken to be $[0, 1]^{10}$ and the candidate set $\mathcal{C}$ consists of all the main effects plus all the two-factor interactions among $x_1, \ldots, x_{10}$. The experimental design is a 50-run uniform design with 50 levels for each factor, which can be conveniently generated by existing computer codes (e.g., Chen et al. 2016 and Huang et al. 2016). A 50-run maximin distance Latin hypercube design (Johnson, Moore, and Ylvisaker 1990) was also tested for this example and similar simulation results were obtained. Basically,

space-filling designs would be desirable for computer experiments. The training and testing data sets are generated as follows.

1. Let $\{\mathbf{x}_1, ..., \mathbf{x}_{50}\}$ be the points in the uniform design and $\{\mathbf{x}_{51}, ..., \mathbf{x}_{550}\}$ be the points of the 500-run maximin distance Latin hypercube design. The latter can be conveniently generated using *The Design of Experiments Toolbox* in Matlab.
2. Simulate a vector $\mathbf{Y} = (Y(\mathbf{x}_1), ..., Y(\mathbf{x}_{550}))^T$ as a realization of a 550-dimensional multivariate normal distribution with the mean vectors and covariance matrices constructed via Models I, *II* and I *II*. Denote the realization values by $\mathbf{y} = (y(\mathbf{x}_1), ..., y(\mathbf{x}_{550}))^T$.
3. Thus, $\{(\mathbf{x}_1, y(\mathbf{x}_1)), ..., (\mathbf{x}_{50}, y(\mathbf{x}_{50}))\}$ consists of the training data set and $\{(\mathbf{x}_{51}, y(\mathbf{x}_{51})), ..., (\mathbf{x}_{550}, y(\mathbf{x}_{550}))\}$ consists of the testing data set.

For the SSBK, we run 100,000 iterations for each repetition and identify the model yielding the smallest RMSPE value among those models whose mean functions have the top five highest frequencies in the posterior samples as the best one. The maximum ratio of the MCSE to the corresponding posterior sample mean across all repetitions is no larger than 0.0195 so it is assurable that 100,000 iterations (thinned by discarding the first 10,000 samples and every four samples and retaining the next one in the sequence) are adequate for the SSBK to achieve convergence. The simulation results are summarized in Table 3.

In Table 3, five measurements are employed to evaluate the performance: average of active effect identified rate (AEIR), average of inactive effect identified rate (IEIR), average size of the identified mean function (MEAN), the mean of the 1,000 RMSPE values (MRMSPE), and the standard deviation of the 1,000 RMSPE values (SRMSPE). Clearly, AEIR is the larger the better, IEIR, MRMSPE, and SRMSPE are the smaller the better, and MEAN is the target the better. The AEIR, IEIR, and MEAN values are calculated by the competing methods, while MRMSPE and SRMSPE are calculated using *The DACE Toolbox* in Matlab.

The performance of the SSBK, as expected, is getting worse as the complexity of the true model increases. But even for the worst case (Model III), the SSBK has an AEIR value larger than 80 percent, an IEIR value about 15 percent, and a MEAN value not significantly larger than the size of the true mean function. Although the SSBK does not perform the best in terms of all the measurements across all the three cases, it has the best MRMSPE and SRMSPE values. This indicates that the models identified by the SSBK are more accurate than the OK and UK models. On the contrary, the OK method, which completely misses the active effects, has the worst MRMSPE and SRMSPE values. On the other hand, the UK methods may perform well in terms of the MRMSPE values when they achieve perfect AEIR values, but their performance greatly deteriorates when they have low AEIR values. Besides the above measurements, the corrected Akaike information criterion (AICC) for variable selection is also examined for this example. The AICC value is computed as

$$AICC = -2\log L + 2n\frac{m_1 + m_2 + 1}{n - m_1 - m_2 - 2},$$

where L is the maximum value of the likelihood function for the model, m_1 is the number of variables in the regression function and m_2 in the covariance function. The AICC value is the smaller the better and such a criterion is suitable for evaluating the performance of kriging models because it not only takes the mean functions into account but also the spatial correlations (Hoeting et al. 2006). The means and standard deviations of the AICC values produced by the competing methods across the repetitions are visualized in Figure 1. To make the comparison results distinguishing, values that are larger than 100 are not displayed in Figure 1, so there may be no data points for some competitors (e.g., the OK under Model I). The data points in Figure 1 make it clear that the proposed SSBK performs the best with respect to the AICC criterion.

Table 3. Summary of simulation results in Example 1.

Model	Method	AEIR (%)	IEIR (%)	MEAN	MRMSPE
I	SSBK	98.50	6.37	6.904	0.6339(0.2098)
	OK	0.00	0.00	0.000	1.5370(3.6516)
	UK (linear)	100.00	11.29	10.000	0.7486(0.2441)
	UK (linear + quadratic)	100.00	27.42	20.000	0.7651(0.2723)
II	SSBK	92.25	9.43	11.099	0.6907(0.2935)
	OK	0.00	0.00	0.000	7.2544(9.4429)
	UK (linear)	50.00	11.86	10.000	7.0319(7.1919)
	UK (linear + quadratic)	100.00	23.73	20.000	0.7500(0.3675)
III	SSBK	83.13	15.04	15.904	0.9590(0.7763)
	OK	0.00	0.00	0.000	25.5810(19.8205)
	UK (linear)	33.33	12.50	10.000	20.3162(16.0509)
	UK (linear + quadratic)	66.67	25.00	20.000	19.0608(11.4181)

AEIR is the average rate at which all of the active effects in the true mean function are correctly identified.
IEIR is the average rate at which inactive effects in the true mean function are identified.
MEAN is the average number of the effects in the selected mean functions.
MRMSPE is the mean of the RMSPE values (the value in the parentheses is the standard deviation).
AEIR is larger the better; IEIR & MRMSPE are small the better; MEAN is target the better.
The CPU time of implementing the SSBK procedure for one repetition is about 1,020 s on a PC with 2.10 GHz triple-core AMD Phenom II N830 CPU and 8 GB memory.

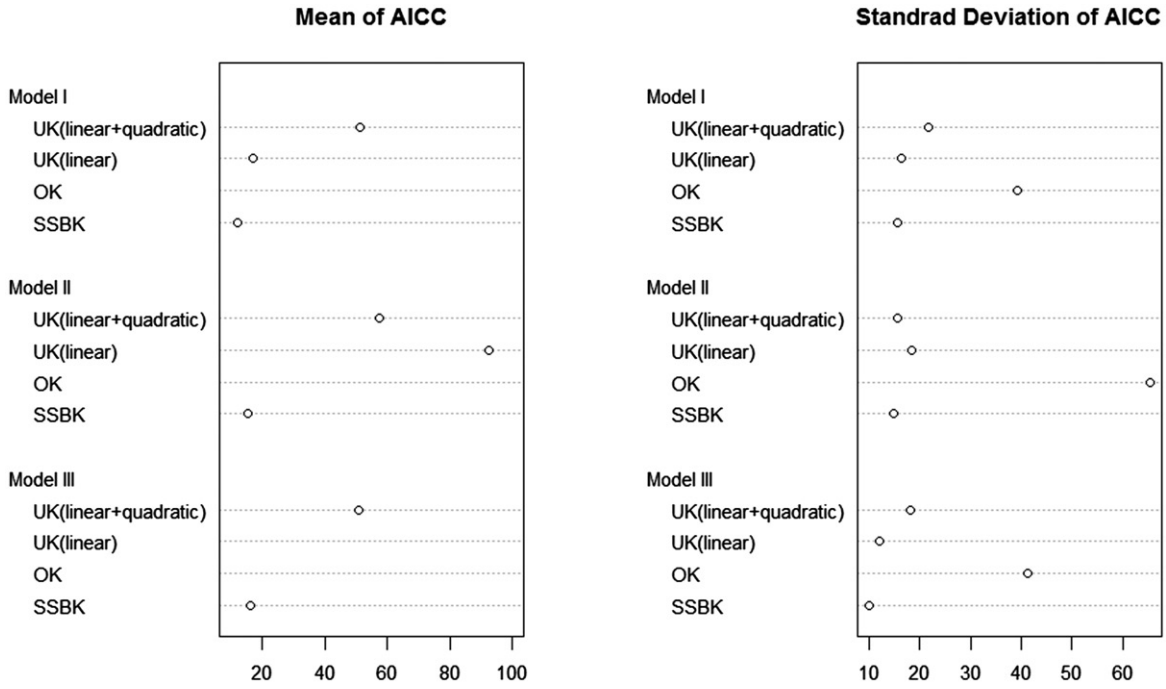

Figure 1. The left panel: the means of the AICC values of the competing methods (the smaller the better); the right panel: the standard deviations of the AICC values of the competing methods (the smaller the better). Values larger than 100 are not displayed.

This example clearly demonstrates that selecting important variables for the mean function is critical to establishing an accurate kriging model.

Example 2. Consider the simulation study with a known function in Hung (2011). The known model is defined on a 12-dimensional ($p = 12$) input space $[0, 1]^{12}$, where the first six variables, $x_1, \ldots, x_6$, have decreasing effects on the computer experiment output and the remaining variables, $x_7, \ldots, x_{12}$, are irrelevant. The true model is

$$y(\mathbf{x}) = 0.4x_1 + 0.3x_2 + 0.2x_3 + 0.1x_4 + 0.05x_5 + 0.01x_6 + \epsilon, \quad [12]$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ with σ_ϵ=0.05. Response values are generated independently using Eq. [12], the experimental designs are Latin hypercube designs (McKay, Beckman, and Conover 1979) with 12 variables and sample sizes $n = 50$, 80, and 100, and the candidate set $\mathcal{C}$ consists of all the first-order main effects. For each fitted model, the RMSPE value is calculated according to 100 randomly generated testing points. For the SSBK, the identifying rule is the same as in the previous example; and for the PBK, the Lasso penalty (Tibshirani 1996) and adaptive Lasso penalty (Zou 2006) are employed.

Based on 500 repetitions, the simulation results are summarized in Table 4, where the results of PBK (Lasso), PBK (adaptive Lasso), OK, and UK were previously published by Hung (2011). The maximum ratio of the MCSE to the corresponding posterior sample mean across all repetitions is no larger than 0.0126, so it is assurable that 100,000 iterations

Table 4. Summary of simulation results in Example 2.

Sample size	Method	AEIR (%)	IEIR (%)	MEAN	MRMSPE
$n = 50$	SSBK	79.60	6.27	5.152	0.0541(0.0043)
	PBK (Lasso)	74.49	10.02	5.090	0.2133(0.0065)
	PBK (adaptive Lasso)	74.81	9.54	5.090	0.2115(0.0063)
	OK	0.00	0.00	0.000	0.2237(0.0065)
	UK	100.00	100.00	12.000	0.2282(0.0067)
$n = 80$	SSBK	81.60	6.03	5.258	0.0519(0.0039)
	PBK (Lasso)	75.00	3.72	4.720	0.2108(0.0062)
	PBK (adaptive Lasso)	75.28	4.45	4.790	0.2113(0.0063)
	OK	0.00	0.00	0.000	0.2220(0.0062)
	UK	100.00	100.00	12.000	0.2266(0.0061)
$n = 100$	SSBK	82.57	6.00	5.314	0.0514(0.0040)
	PBK (Lasso)	75.11	0.89	4.560	0.2084(0.0062)
	PBK (adaptive Lasso)	75.39	0.78	4.570	0.2092(0.0061)
	OK	0.00	0.00	0.000	0.2220(0.0061)
	UK	100.00	99.61	12.000	0.2238(0.0061)

AEIR is the average rate at which all of the active effects in the true mean function are correctly identified.
IEIR is the average rate at which inactive effects in the true mean function are identified.
MEAN is the average number of the effects in the selected mean function.
MRMSPE is the mean of the RMSPE values (the value in the parentheses is the standard deviation).
AEIR is larger the better; IEIR & MRMSPE are smaller the better; MEAN is target the better.
The CPU times of implementing the SSBK procedure for one repetition are about 2,224 s, 3,682 s and 5,529 s for $n = 50, 80$ and 100, respectively.

(thinned by discarding the first 10,000 samples and every four samples and retaining the next one in the sequence) are adequate for the SSBK to achieve convergence. Again, the five simulation measurements, AEIR, IEIR, MEAN, MRMSPE, and SRMSPE, are adopted to evaluate the performance.

It is shown that the SSBK is the best in terms of MRMSPE and SRMSPE. Compared with the PBKs, the SSBK results in better values in terms of AEIR and MEAN. When the sample size is large ($n = 80$ or 100), the IEIR value of the SSBK is larger than those of the PBKs. That is, the SSBK may tend to include more inactive effects into the mean function than the PBKs when the sample size is large. However, when the sample size is small ($n = 50$), the IEIR values of the PBKs greatly increase, while the SSBK is robust to IEIR values.

Example 3. Consider the simulation study with a known function in Marrel et al. (2008). The known function, called the *g*-function of Sobol, is defined for d input variables taking values on $[0,1]^d$:

$$g_{Sobol}(X_1,\ldots,X_d)=\prod_{k=1}^{d} g_k(X_k) \text{ where } g_k(X_k)=\frac{|4X_k-2|+b_k}{1+b_k} \text{ and } b_k\geq 0.$$

Due to its complexity (strongly nonlinear and nonmonotonic relationship) and the availability of analytical sensitivity indices, the *g*-function of Sobol is a well-known test example in the studies of global sensitivity analysis algorithms (Saltelli, Chan, and Scott 2000). Following Marrel et al. (2008), we choose $b_k = k$, the candidate set $\mathcal{C}$ consists of all the first-order main effects and different dimensions of input variables are considered, from 4 to 20: $d = 4, 6, \ldots, 20$. For each dimension d, we generate a training sample formed by $N_{LS} = d \times 10$ simulations of the *g*-functions of Sobol using the Latin hypercube design (McKay, Beckman, and Conover 1979). Using these training data, two kriging models are built: one following the proposed SSBK and one using the methodology of Marrel et al. (2008). In this example, the predicability coefficient Q_2 is computed to evaluate the performance of the established models. Q_2 corresponds to the classical coefficient of determination R^2 for a test sample, that is, for prediction residuals:

$$Q_2\left(Y,\hat{Y}\right)=1-\frac{\sum_{i=1}^{n_{test}}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n_{test}}\left(\bar{Y}-y_i\right)},$$

where $Y=(y_1,\ldots,y_{n_{test}})^T$ denotes the n_{test} observations of the test set and $\bar{Y}$ is their empirical mean, $\hat{Y}=(\hat{y}_1,\ldots,\hat{y}_{n_{test}})^T$ represents the kriging model predicted values at the test set (see Eq. [4]). For each method, the Q_2 coefficient is computed on a random test sample of $N_{TS} = 1{,}000$ points. For each dimension d, this procedure is repeated 50 times to obtain an average performance in terms of Q_2, which is the larger the better. The standard deviation of Q_2, which is the smaller the better, is also a good indicator of the robustness of each method.

Based on 50 repetitions, the means of Q_2 and the corresponding standard deviations of the competitors are visualized in Figure 2, where the results of Marrel et al.'s method were previously reported by Marrel et al. (2008). The maximum ratio of the MCSE to the corresponding posterior sample mean across all repetitions is no larger than 0.0446, so it is assurable that 100,000 iterations (thinned by discarding the first 10,000 samples and every four samples and retaining the next one in the sequence) are adequate for the SSBK to achieve convergence. From Figure 2, it is plain that the proposed SSBK compares very favorably with the methodology of Marrel et al. (2008) in terms of the Q_2 criterion. The performance of the OK, the UK with linear main effects and the UK with linear main effects plus quadratic effects are also evaluated. As expected, their Q_2 values are small (typically less than 0.6); hence, their simulation results are omitted to save space.

Remark 2. More insights and discussions regarding the simulation are given as follows:

1. In our simulation studies, the best models are identified among the top-posterior-frequency models using the RMSPE criterion. It turns out that the best models by the RMSPE criterion are always the ones with the highest posterior frequencies. This also occurs for the practical example in Section 4. However, it is recommended that a handful of relatively high-frequency models should be retained because sometimes the best models may not be identified with the highest posterior frequencies (George and McCulloch 1993).
2. Parameter tuning may further improve the efficiency of the SSBK. In this article, we set $c_i = 10$ and it works well in our examples. As did Beattie, Fong, and Lin (2002), we have also tested $c_i = 5, 100$ and 500. Unfortunately, $c_i = 5$ does not outperform $c_i = 10$ while $c_i = 100$ or 500 often causes singularity problems (see the matrix **A** in Proposition 1). A possible efficient tuning strategy is to tune c_i and τ_i simultaneously (see Chen et al. 2011, 2013; Beattie, Fong, and Lin 2002). A

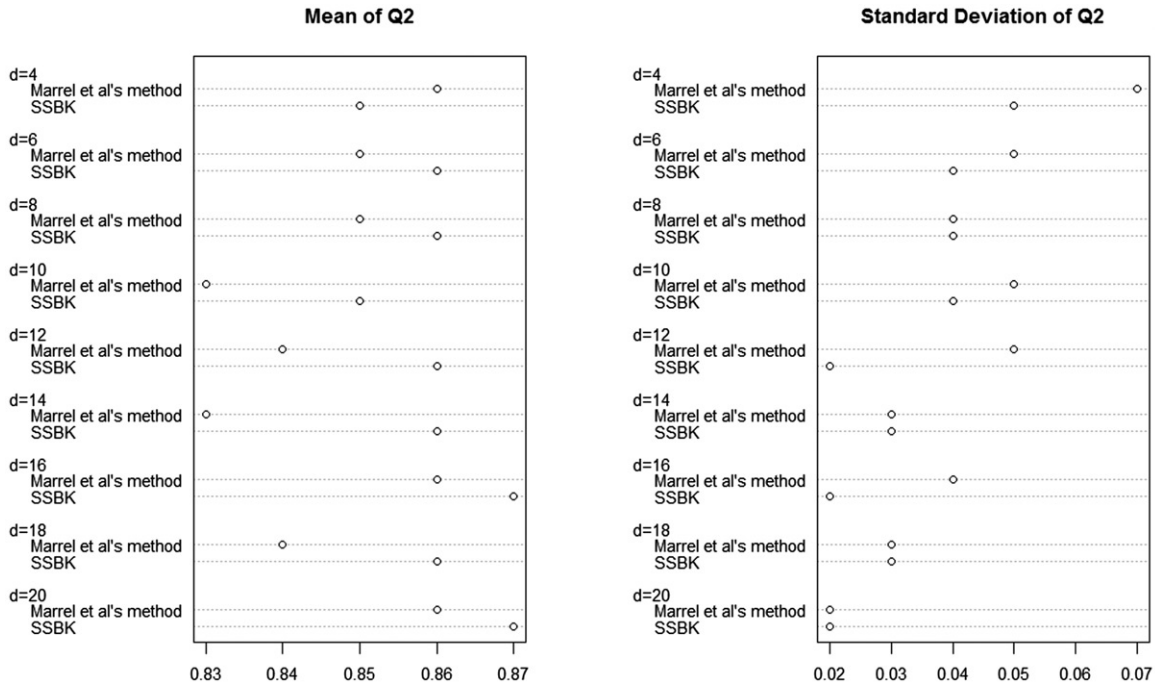

Figure 2. The left panel: the means of the Q_2 values of the competing methods (the larger the better); the right panel: the standard deviations of the AICC values of the competing methods (the smaller the better).

tuning strategy may add additional computational complexity. As noted in Remark 1, single-site Gibbs sampling schemes that diminish the expensive computation of the matrix inverse may deserve further study to alleviate the computational load.

3. Example 3 shows that the proposed SSBK compares favorably with the methodology of Marrel et al. (2008). However, it should be added that the methodology of Marrel et al (2008) is more general than the proposed SSBK in the sense that their idea aims to select important variables for both the mean function and the covariance function. Marrel et al. (2008) is based on a stepwise-type algorithm, which could be often stuck in local optimal mean functions. On the contrary, the proposed SSBK takes all possible mean functions into consideration and, therefore, has greater opportunities to obtain the optimal mean functions. In other words, general methodologies like Marrel et al. (2008) may require more challenging computational algorithms to deal with the high-dimensional problem although they are theoretically appealing.

6. Concluding remarks and further discussions

Computer experiments have received a great deal of attention in the literature since the pioneering work of McKay, Beckman, and Conover (1979) and Sacks et al. (1989). Variable selection in computer experiments has become a promising research topic. In most cases, kriging is used as a metamodeling technique in computer experiments and variable selection based on kriging can be usually conducted in two stages: (1) selecting variables with significant impact on the GP and then (2) selecting variables (terms) for the mean function. Stage (2) is important due to the strong trends potentially existing in the response. In this article, we propose a Bayesian variable selection method, called the SSBK, for the mean function of kriging. A novel property of the SSBK is that, within the postulated Bayesian framework, the posterior probabilities of the "most likely" models are larger than the posterior probabilities of the "less likely" models in a relative sense. This is guaranteed by the well-known Markov convergence property of the Gibbs sampler. A practical example and some simulation studies show that the SSBK compares favorably with the BK method of Joseph, Hung, and Sudjianto (2008) and the PBK method of Hung (2011). Also, the SSBK may improve the efficacy of the existing computer experiment software (in this work we use *The DACE Toolbox* in Matlab) by providing reasonable mean functions for them.

More insights should be noted on the advantages of using the SSBK over the BK of Joseph, Hung, and Sudjianto (2008) and the PBK of Hung (2011). The

theoretical properties of the BK method are difficult to derive because of the forward selection manner. For the PBK method, its appealing oracle property only holds when the used experimental design forms a complete lattice (Hung 2011). When the experimental resources are limited, however, a complete lattice design is typically infeasible. In contrast, the proposed SSBK possesses the Markov convergence property under a wide class of experimental designs. As shown in Sections 4 and 5, the superiority of the SSBK is demonstrated over the BK and PBK. The advantages of using the SSBK over the "universal" approach, like the methodology of Marrel et al. (2008), are also discussed. However, the relationship between the Markov convergence property and the selection consistency remains to be determined, which deserves further investigation.

About the authors

Dr. Huang is Assistant Professor in the College of Mathematics and Statistics at Guangxi Normal University. His email address is hhzgxnu@126.com.

Dr. Lin is University Distinguished Professor of Statistics at Penn State University. He is a fellow of IMS, RSS, ASQ and ASA. His email address is DKL5@psu.edu.

Dr. Liu is Professor in the School of Statistics and Data Science at Nankai University. His email address is mqliu@nankai.edu.cn.

Dr. Zhang is Associate Professor in the School of Statistics and Data Science at Nankai University. Her email address is zhangqz@nankai.edu.cn.

Acknowledgments

The authors thank the Editor and two referees for their valuable comments.

Funding

This work was supported by the National Natural Science Foundation of China (grant nos. 11771220, 11701109, 11431006, and 11601244), National Ten Thousand Talents Program, Natural Science Foundation of Guangxi Province (grant no. 2018JJB110027), Tianjin Development Program for Innovation and Entrepreneurship, Tianjin "131" Talents Program, and Project 61331903. This work was also supported by the National Security Agency (grant no. H98230-15-1-0253) and National Science Foundation (grant no. DMS-18102925). The authorship is listed in alphabetic order.

References

Beattie, S. D., D. K. H. Fong, and D. K. J. Lin. 2002. A two-stage Bayesian model selection for supersaturated designs. *Technometrics* 44 (1):55–63. doi: 10.1198/004017002753398326.

Box, G. E. P., and R. D. Meyer. 1986. An analysis for unreplicated fractional factorials. *Technometrics* 28 (1):11–8. doi: 10.1080/00401706.1986.10488093.

Chang, S. M., R. B. Chen, and Y. C. Chi. 2016. Bayesian variable selections for probit models with componentwise Gibbs samplers. *Communications in Statistics-Simulation and Computation* 45 (8):2752–66. doi: 10.1080/03610918.2014.922983.

Chen, H., H. Z. Huang, D. K. J. Lin, and M. Q. Liu. 2016. Uniform sliced Latin hypercube designs. *Applied Stochastic Models in Business and Industry* 32 (5):574–84. doi: 10.1002/asmb.2192.

Chen, R. B., C. H. Chu, T. Y. Lai, and Y. N. Wu. 2011. Stochastic matching pursuit for Bayesian variable selection. *Statistics and Computing* 21 (2):247–59. doi: 10.1007/s11222-009-9165-4.

Chen, R. B., J. Z. Weng, and C. H. Chu. 2013. Screening procedure for supersaturated designs using a Bayesian variable selection method. *Quality and Reliability Engineering International* 29 (1):89–101. doi: 10.1002/qre.1299.

Fang, K. T., R. Li, and A. Sudjianto. 2006. *Design and modeling for computer experiments*. Boca Raton, FL: Chapman & Hall/CRC.

Fang, K. T., D. K. J. Lin, P. Winker, and Y. Zhang. 2000. Uniform design: Theory and application. *Technometrics* 42 (3):237–48. doi: 10.1080/00401706.2000.10486045.

George, E. I., and R. E. McCulloch. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88 (423):881–9. doi: 10.1080/01621459.1993.10476353.

Haaland, B., and P. Z. G. Qian. 2011. Accurate emulators for large-scale computer experiments. The *Annals of Statistics* 39 (6):2974–3002. doi: 10.1214/11-AOS929.

Han, G., T. J. Santner, and J. J. Rawlinson. 2009. Simultaneous determination of tuning and calibration parameters for computer experiments. *Technometrics* 51 (4):464–74. doi: 10.1198/TECH.2009.08126.

Han, Z. H., and S. Görtz. 2012. Hierarchical kriging model for variable-fidelity surrogate modeling. *AIAA Journal* 50 (9):1885–96. doi: 10.2514/1.J051354.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57 (1):97–109. doi: 10.1093/biomet/57.1.97.

Hoeting, J. A., R. A. Davis, A. A. Merton, and S. E. Thompson. 2006. Model selection for geostatistical models. *Ecological Applications* 16 (1):87–98. doi: 10.1890/04-0576.

Hoffman, R. M., A. Sudjianto, X. Du, and J. Stout. 2003. Robust piston design and optimization using piston secondary motion analysis. SAE Paper 2003–01–0148, SAE Transactions.

Huang, H. Z., H. B. Fang, and M. Tan. 2018. Experimental design for multi-drug combination studies using signaling networks. *Biometrics* 74 (2):538–47. doi: 10.1111/biom.12777.

Huang, H. Z., D. K. J. Lin, M. Q. Liu, and J. F. Yang. 2016. Computer experiments with both qualitative and quantitative variables. *Technometrics* 58 (4):495–507. doi: 10.1080/00401706.2015.1094416.

Huang, H. Z., S. S. Zhou, M. Q. Liu, and Z. F. Qi. 2017. Acceleration of the stochastic search variable selection via componentwise Gibbs sampler. *Metrika* 80 (3):289–308. doi: 10.1007/s00184-016-0604-x.

Hung, Y. 2011. Penalized blind kriging in computer experiments. *Statistica Sinica* 21 (3):1171–90. doi: 10.5705/ss.2009.226.

Johnson, M. E., L. M. Moore, and D. Ylvisaker. 1990. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference* 26 (2):131–48. doi: 10.1016/0378-3758(90)90122-B.

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):1537–47. doi: 10.1198/016214506000000492.

Joseph, V. R., Y. Hung, and A. Sudjianto. 2008. Blind kriging: A new method for developing metamodels. *Journal of Mechanical Design* 130 (3):031102. doi: 10.1115/1.2829873.

Kennedy, M. C., and A. O'Hagan. 2000. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* 87 (1):1–13. doi: 10.1093/biomet/87.1.1.

Le, N. D., and J. V. Zidek. 2006. *Statistical analysis of environmental space-time processes*. New York, NY: Springer.

Li, R., and A. Sudjianto. 2005. Analysis of computer experiments using penalized likelihood in Gaussian kriging models. *Technometrics* 47 (2):111–20. doi: 10.1198/004017004000000671.

Linkletter, C. D., D. Bingham, N. Hengartner, D. Higdon, and K. Q. Ye. 2006. Variable selection for Gaussian process models in computer experiments. *Technometrics* 48 (4):478–90. doi: 10.1198/004017006000000228.

Marrel, A., B. Iooss, F. Van Dorpe, and E. Volkova. 2008. A efficient methodology for modeling complex computer codes with Gaussian processes. *Computational Statistics and Data Analysis* 52 (10):4731–44. doi: 10.1016/j.csda.2008.03.026.

Martin, J. D., and T. W. Simpson. 2005. On the use of kriging models to approximate deterministic computer models. *AIAA Journal* 43 (4):853–63. doi: 10.2514/1.8650.

McKay, M. D., R. J. Beckman, and W. J. Conover. 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21 :239–45. doi: 10.2307/1268522.

Morris, M. D. 1991. Factorial sampling plans for preliminary computational experiments. *Technometrics* 33 (2): 161–74. doi: 10.1080/00401706.1991.10484804.

Neal, R. M. 2003. Slice sampling. *The Annals of Statistics* 31 (3):705–67. doi: 10.1214/aos/1056562461.

Qian, P. Z. G., and C. F. J. Wu. 2008. Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics* 50 (2):192–204. doi: 10.1198/004017008000000082.

Ranjan, P., R. Haynes, and R. Karsten. 2011. A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data. *Technometrics* 53 (4):366–78. doi: 10.1198/TECH.2011.09141.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn. 1989. Design and analysis of computer experiments. *Statistical Science* 4 (4):409–23. doi: 10.1214/ss/1177012413.

Saltelli, A., K. Chan, and E. M. Scott. 2000. *Sensitivity analysis*. Hoboken, NJ: John Wiley and Sons.

Santner, T. J., B. J. Williams, and W. I. Notz. 2003. *The design and analysis of computer experiments*. New York, NY: Springer.

Shao, J. 2003. *Mathematical statistics*. New York, NY: Springer.

Tan, M. H. Y., and C. F. J. Wu. 2012. Robust design optimization with quadratic loss derived from Gaussian process models. *Technometrics* 54 (1):51–63. doi: 10.1080/00401706.2012.648866.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B* 58:267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x.

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. doi: 10.2307/1269548.

Wu, C. F. J., and M. Hamada. 2009. *Experiments: Planning, analysis, and optimization*. 2nd ed. New York, NY: Wiley.

Zhou, Q., P. Z. G. Qian, and S. Zhou. 2011. A simple approach to emulation for computer models with qualitative and quantitative factors. *Technometrics* 53 (3): 266–73. doi: 10.1198/TECH.2011.10025.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101 (476): 1418–29. doi: 10.1198/016214506000000735.

Appendix

A.1. Proof of Proposition 1

The joint distribution of all variables, denoted by $[\boldsymbol{\mu}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{\rho}, \mathbf{Y}]$, can be expressed as

$$[\boldsymbol{\mu}, \sigma^2, \boldsymbol{\delta}, \boldsymbol{\rho}, \mathbf{Y}] = [\mathbf{Y}|\boldsymbol{\mu}, \sigma^2, \boldsymbol{\rho}]\,[\boldsymbol{\mu}|\boldsymbol{\delta}, \sigma^2]\,[\boldsymbol{\delta}][\sigma^2][\boldsymbol{\rho}], \tag{13}$$

where

$$[\mathbf{Y}|\boldsymbol{\mu}, \sigma^2, \boldsymbol{\rho}] \propto \sigma^{-n}|\mathbf{R}|^{-\frac{1}{2}} \exp\left\{-(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu})/(2\sigma^2)\right\}, \tag{14}$$

$$[\boldsymbol{\mu}|\boldsymbol{\delta}, \sigma^2] \propto \sigma^{-k}|\mathbf{D}|^{-\frac{1}{2}} \exp\left\{-\boldsymbol{\mu}^T\mathbf{D}^{-1}\boldsymbol{\mu}/(2\sigma^2)\right\}, \tag{15}$$

$$[\boldsymbol{\delta}] = 2^{-k}, [\sigma^2] \propto 1/\sigma^2, [\boldsymbol{\rho}] = I_{(0,1)^k}(\boldsymbol{\rho}). \tag{16}$$

Eq. [14] results from Eq. [5], Eq. [15] results from Eq. [7] and the independence assumption for each component of $\boldsymbol{\mu}$, Eq. [16 results from the assumptions presented in the last paragraph of Section 2.2.

The full conditional distribution of $\boldsymbol{\mu}$ up to a normalization constant can be expressed by omitting the irrelevant terms on the right-hand side of Eq. [13], that is,

$$\begin{aligned}[\boldsymbol{\mu}|\sigma^2, \boldsymbol{\delta}, \boldsymbol{\rho}, \mathbf{Y}] &\propto [\mathbf{Y}|\boldsymbol{\mu}, \sigma^2, \boldsymbol{\rho}]\,[\boldsymbol{\mu}|\boldsymbol{\delta}, \sigma^2] \\ &\propto \exp\left\{-\left[(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu}) + \boldsymbol{\mu}^T\mathbf{D}^{-1}\boldsymbol{\mu}\right]/(2\sigma^2)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\sigma^{-2}\left(\tilde{\mathbf{F}}^T\mathbf{R}^{-1}\tilde{\mathbf{F}} + \mathbf{D}^{-1}\right)\boldsymbol{\mu} + \sigma^{-2}\mathbf{Y}^T\mathbf{R}^{-1}\tilde{\mathbf{F}}\boldsymbol{\mu}\right\}.\end{aligned}$$

Then Eq. [8] follows from Lemma B.1.1 of Santner, Williams, and Notz (2003). Similarly,

$$\begin{aligned}[\sigma^2|\boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\rho}, \mathbf{Y}] &\propto [\mathbf{Y}|\boldsymbol{\mu}, \sigma^2, \boldsymbol{\rho}]\,[\boldsymbol{\mu}|\boldsymbol{\delta}, \sigma^2][\sigma^2] \\ &\propto (\sigma^2)^{-\left[(n+k)/2+1\right]} \exp\left\{-\left[(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu})^T\mathbf{R}^{-1}(\mathbf{Y}-\tilde{\mathbf{F}}\boldsymbol{\mu}) + \boldsymbol{\mu}^T\mathbf{D}^{-1}\boldsymbol{\mu}\right]/(2\sigma^2)\right\}.\end{aligned}$$

By Wu and Hamada (2009), the last term of the above expression is the density kernel of an inverted gamma distribution, hence Eq. [9] follows.

Next, by omitting the terms on the right-hand side of Eq. [13] that are irrelevant to $\boldsymbol{\delta}$, we have

$$\left[\delta_i|\boldsymbol{\delta}_{(-i)},\boldsymbol{\mu},\sigma^2,\boldsymbol{\rho},\mathbf{Y}\right]\propto\left[\boldsymbol{\mu}|\delta_i,\boldsymbol{\delta}_{(-i)},\sigma^2\right]\left[\delta_i,\boldsymbol{\delta}_{(-i)}\right].$$

Eq. [10] follows from the fact that the distribution of δ_i is Bernoulli and $[\delta_i,\boldsymbol{\delta}_{(-i)}]=2^{-k}$.

Finally, by omitting the terms on the right-hand side of Eq. [13] that are irrelevant to $\boldsymbol{\rho}$, we have

$$\left[\boldsymbol{\rho}|\boldsymbol{\mu},\sigma^2,\boldsymbol{\delta},\mathbf{Y}\right]\propto\left[\mathbf{Y}|\boldsymbol{\mu},\sigma^2,\boldsymbol{\rho}\right][\boldsymbol{\rho}].$$

Then Eq. [11] is straight forward.

A.2. Computation issues of MCSE

Without loss of generality, let $\mathbf{X}=\{X_1,X_2,\ldots\}$ be a Gibbs sequence with a target distribution π having support $\mathcal{X}$, and g be a real-valued, π-integrable function. Under regularity conditions, the Gibbs sequence $\mathbf{X}$ and the function g will admit a central limit theorem, that is,

$$\sqrt{n}(\bar{g}_n-E_\pi g)\xrightarrow{d}N\left(0,\sigma_g^2\right),$$

where $\bar{g}_n=n^{-1}\sum_{i=1}^n g(X_i)$, "$\xrightarrow{d}$" denotes convergence in distribution, $E_\pi g=\int_{\mathcal{X}}g(x)\pi(dx)$ and $\sigma_g^2=\mathrm{var}\{g(X_1)\}+2\sum_{i=2}^{\infty}\mathrm{cov}_\pi\{g(X_1),g(X_i)\}$. Suppose that the Gibbs sampler algorithm is run for a total of $n=a_nb_n$ iterations and define

$$\bar{Y}_j=\frac{1}{b_n}\sum_{i=(j-1)b_n+1}^{jb_n}g(X_i)\ \text{ for } j=1,\ldots,a_n.$$

That is, the sequence is broken into a_n blocks of equal size b_n. Then the batch means estimate of σ_g^2 is

$$\hat{\sigma}_g^2=\frac{b_n}{a_n-1}\sum_{j=1}^{a_n}\left(\bar{Y}_j-\bar{g}_n\right)^2.$$

Jones et al. (2006) showed that, if the batch size and the number of batches are allowed to increase as the overall length of the simulation increases by setting $b_n=\lfloor n^\theta\rfloor$ and $a_n=\lfloor n/b_n\rfloor$, then $\hat{\sigma}_g^2$ is a consistent estimator. Often $\theta=1/2$ (i.e., $b_n=\lfloor\sqrt{n}\rfloor$ and $a_n=\lfloor n/b_n\rfloor$) is a convenient choice. The Monte Carlo standard error (MCSE) of $\bar{g}_n$ is calculated by

$$\mathrm{MCSE}(\bar{g}_n)=\frac{\hat{\sigma}_g}{\sqrt{n}}.$$

It is easy to see that $\mathrm{MCSE}(\bar{g}_n)\xrightarrow{p}0$ as $n\to\infty$. Hence, if the MCSE is sufficiently small, for example, $|\mathrm{MCSE}(\bar{g}_n)/\bar{g}_n|\le 0.05$, one may conclude that the sequence is convergent.