Journal of Statistical Research *ISSN 0256 - 422 X*
2013, Vol. 47, No. 1, pp. 11-30

A PREDICTION-ORIENTED BAYESIAN SITE SELECTION APPROACH FOR LARGE SPATIAL DATA

JINCHEOL PARK
Mathematical Biosciences Institute (MBI), Department of Statistics,
The Ohio State University, Columbus, OH, USA
Email: park.1750@mbi.osu.edu

FAMING LIANG
Texas A&M University, College Station, TX, USA
Email: fliang@stat.tamu.edu

SUMMARY

The Gaussian geostatistical model has been widely used in spatial data modeling. In spite of its popularity, this model suffers from a severe implementation problem for Bayesian inference, for which a covariance matrix needs to be inverted at each iteration. This is infeasible when the number of observations is large. In this paper, we propose a prediction-oriented Bayesian site selection (BSS) approach to tackle this difficulty. By dividing the observations into two sets, response variables and explanatory variables, the BSS approach forms a regression model which relates the observations through a conditional likelihood derived from the original Gaussian geostatistical model, and then reduces the dimension of the data using a stochastic variable selection procedure. Our numerical results indicate that the BSS approach can produce very good parameter estimates and prediction for large spatial data, while significantly reducing the computational time required by conventional Bayesian approaches.

Keywords and phrases: Bayesian Variable Selection, Geostatistics, Markov Chain Monte Carlo, Spatial Data.

1 Introduction

Geostatistics is a branch of spatial statistics which deals with the data obtained by sampling from a spatially continuous process $\{X(s)\}$, $s \in \mathbb{R}^2$, at a discrete set of locations $\{s_i, i = 1, \ldots, n\}$ in a spatial region of interest $A \subset \mathbb{R}^2$. Consider a Gaussian geostatistical model,

$$Y(s_i) = \nu(s_i) + X(s_i) + \varepsilon_i, \tag{1.1}$$

where $\{Y(s_i)\}$ denotes the observations at locations $s_1, \ldots, s_n$, $\{\nu(s_i)\}$ denotes the mean of $\{Y(s_i)\}$, $\{X(s_i)\}$ denotes a spatial Gaussian process with mean zero, variance σ^2 and the correlation function Corr$\{X(s_i), X(s_j)\} = \rho(\|s_i - s_j\|)$ with Euclidean distance $\|\cdot\|$, and ε_i's are independent

Gaussian random errors with mean 0 and variance τ^2. The variance τ^2 is called the nugget variance in the literature of spatial statistics. The correlation function can be chosen from some parametric families, such as the Matérn, powered exponential or spherical (Cressie, 1993). Under model (1.1), $\{Y(s_1), \ldots, Y(s_n)\}$ follows a multivariate Gaussian distribution,

$$\{Y(s_1), \ldots, Y(s_n)\}^T \sim N(\boldsymbol{\nu}, \Sigma), \tag{1.2}$$

where $\boldsymbol{\nu} = (\nu(s_1), \ldots, \nu(s_n))^T$, $\Sigma = \sigma^2 R + \tau^2 I$, and I is the $n \times n$ identity matrix and R is an $n \times n$ correlation matrix with the $(i, j)^{th}$ element given by $\rho(\|s_i - s_j\|)$. In this paper, R is called the spatial correlation matrix of the locations $\{s_i, i = 1, \ldots, n\}$. Relevant covariates can be easily incorporated into the model by replacing the mean $\nu(s_i)$ by

$$\nu(s_i) = \xi_0 + \sum_{j=1}^{p} \xi_j c_j(s_i), \tag{1.3}$$

where $c_j(\cdot)$ denotes the j^{th} covariate, ξ_j denotes the corresponding regression coefficient, and ξ_0 is the intercept.

A problem of general interest in spatial statistics is to predict unobserved values of $\{Y(s_i^p)\}$ at a set of locations $\mathbf{s}^p = \{s_1^p, \ldots, s_{n_p}^p\}$. A core difficulty for this problem is at inverting the $n \times n$ covariance matrix Σ, which is involved in almost all standard statistical approaches to this problem, such as Kriging (see e.g., Stein, 1999) and Bayesian modeling (Diggle *et al.*, 1998). In Bayesian modeling, the covariance matrix needs to be inverted once at each iteration in order to evaluate the posterior for the updated parameters. It is known that the computational complexity of matrix inversion increases as $O(n^3)$. When n is large, this is infeasible due to the limit of the current computational power.

A simple strategy to deal with this difficulty is dependence truncation; that is, setting the long-range dependence among $Y(s_i)$'s to be zero. For example, the local Kriging method predicts the value of $Y(s_i^p)$ based only on the observations lying in a neighborhood of $Y(s_i^p)$, and the covariance tapering method (see e.g., Furrer *et al.*, 2006 and Kaufman *et al.*, 2008) sets the correlations to be zero for the pairs of observations with the distance exceeding a threshold value. Although these methods work well for many problems, how to make use of full data information in prediction is still a major concern to many researchers.

An alternative strategy to deal with the matrix inversion difficulty is to develop a new space process which approximates the process $\{X(s_i)\}$ in the fixed region $A \subset \mathbb{R}^2$ but with certain advantages in computation. A popular idea is to approximate the process $\{X(s_i)\}$ by a lower dimensional space process $\{\tilde{X}(s)\}$ with some smoothing techniques, such as kernel convolutions, moving averages, low rank splines, basis functions, or continuous global surfaces; see e.g., Wikle and Cressie (1999), Lin *et al.* (2000), Billings *et al.* (2002), Kammann and Wand (2003), Paciorek (2007), Banerjee *et al.* (2008), Cressie and Johannesson (2008), Stein (2008) and Finley *et al.* (2009). We note that for a large dataset, the dimension of the approximation process $\{\tilde{X}(s)\}$ can still be very high to the current computational power, and this may hinder the applicability of these methods. Another idea, which seems even more attractive in computation, is to approximate the

process $\{X(s_i)\}$ by a Markov process, for which the covariance matrix is sparse and thus manageable in computation even for a large dataset. Related work include Rue and Tjelmeland (2002), Rue and Held (2005), Besag and Mondal (2005), Lindgren *et al.* (2010), Park and Liang (2011), among others. Working on the approximation processes resolves the issue of matrix inversion, but leaves us little understanding to the underlying true Gaussian process. Recently, Rue et al. (2009) suggested the integrated nested Laplace approximation (INLA) method for approximate Bayesian inference of latent Gaussian models. Lindgren *et al.* (2010) applied INLA to the Gaussian field by representing it as a Gaussian Markov random field (GMRF) which incorporates a subclass of Matérn Covariance functions through solving a stochastic partial differential equation (SPDE). However, as pointed out in Lindgren *et al.* (2010), this method involves costs of solving stochastic partial differential equations and for irregularly spaced data, it needs additional costs for triangulation of locations of the observations.

In addition to lower-dimensional process approximations, some authors proposed to approximate the likelihood function of $\{Y(s_i)\}$ by a pseudo-likelihood that is more easily maximized, see e.g., Vecchia (1988), Jones and Zhang (1997) and Stein *et al.* (2004). The underlying idea of these methods is the high-order Markov process approximation. They work by partitioning the observations $\{Y(s_i)\}$ into some subvectors which have a certain kind of Markov structure, and thus the likelihood function can be approximated by the product of a series of lower-order conditional densities. How to partition the data appropriately is a major concern of the methods in applications.

In this paper, we propose a Bayesian site selection (BSS) method which, while reducing the dimension of data, attempts to avoid the shortcomings of the dependence truncation, lower-dimensional process approximation, and likelihood approximation methods. The BSS method first split the observations into two parts, the observations "near" the prediction sites (part I) and their remaining (part II). [How to select the observations "near" prediction sites will be discussed in Section 2.2.] Then, by treating the observations in part I as response variable and those in part II as explanatory variables, BSS forms a regression model which relates all observations $\{Y(s_i)\}$ through a conditional likelihood derived from the original model (1.1). The dimension of the data can then be reduced by applying a stochastic variable selection procedure to the regression model, which selects only a subset of the part II data as explanatory variables. The selected explanatory variables together with the response data thus form the basis of observations for inference of model (1.1) and prediction of unobserved values. Compared to the dependence truncation methods, BSS is able to catch the long range dependence through selection of appropriate explanatory variables. Compared to the lower-dimensional process and likelihood approximation methods, BSS can provide us more understanding to the underlying true Gaussian process, as it directly works on the original process without any approximations involved.

The remainder of this paper is organized as follows. In Section 2, we introduce the BSS method, describing how to form the regression model for a given dataset and discussing how the Metropolis-within-Gibbs sampler can be applied to BSS for parameter estimation and selection of appropriate explanatory variables. In Section 3, we study the performance of BSS using some simulation data. In Section 4, we test the performance of BSS on two real data sets. In Section 4, we conclude the paper with a brief discussion.

2 Bayesian Site Selection

2.1 The Regression Model Formulation

Let $D = \{y(s_i)\}$ denote the observations drawn from the model (1.1) at at n distinct locations $\mathbf{s} = \{s_1, \ldots, s_n\}$, and let $\mathbf{s}^p = \{s_1^p, \ldots, s_{n_p}^p\}$ denote n_p distinct locations of interest for prediction. Suppose that D has been partitioned two sets, $D_y = \{y(s_i); s_i \in \mathbf{s}^y, i = 1, \ldots, n^*\}$ and $D_{-y} = D \backslash D_y$, where $\mathbf{s}^y = \{s_1^y, \ldots, s_{n^*}^y\}$ is the set of locations of the observations contained in D_y. In addition, we assume that D_y has been selected to consist of all observations that are near the prediction sites $\mathbf{s}^p$. How to select D_y will be discussed in Section 2.2.

Let $Y(\mathbf{s}^y) = (Y(s_1^y), \ldots, Y(s_{n^*}^y))^T$ denote the vector of observations contained in D_y. Likewise, let $Z(\mathbf{s}^{-y})$ denote the vector of observations contained in D_{-y}. Following from model (1.1), the distribution of $Y(\mathbf{s}^y)$ conditioned on $Z(\mathbf{s}^{-y})$ follows a multivariate normal distribution; that is, a normal regression can then be formulated as

$$Y(\mathbf{s}^y) \sim Z(\mathbf{s}^{-y}),$$

where $Y(\mathbf{s}^y)$ works as the response variable and $Z(\mathbf{s}^{-y})$ works as the explanatory variable. Instead of using all $Z(\mathbf{s}^{-y})$ as explanatory variables, we would select a subset of $Z(\mathbf{s}^{-y})$ as the explanatory variables for $Y(\mathbf{s}^y)$, as the variables in $Z(\mathbf{s}^{-y})$ can be highly correlated given the nature of spatial model (1.1). With a little abuse of notations, we denote by $Z = (Z(s_1^z), \ldots Z(s_m^z))$ the set of variables used as the explanatory variables of $Y(\mathbf{s}^y)$, where $m = |Z|$ denotes the size of the set Z. Let $\boldsymbol{\nu}_y = E(Y(s^y))$, $\boldsymbol{\nu}_z = E(Z)$, $\Sigma_y = \text{Var}(Y(\mathbf{s}^y))$, $\Sigma_z = \text{Var}(Z)$, and $\Sigma_{yz} = \Sigma_{zy} = \text{Cov}(Y(\mathbf{s}^y), Z)$. Then the conditional distribution $Y(\mathbf{s}^y)|Z$ is given by

$$Y(\mathbf{s}^y)|Z \sim N\left(\boldsymbol{\nu}_{y|z}, \Sigma_{y|z}\right) \tag{2.1}$$

where

$$\begin{aligned} \boldsymbol{\nu}_{y|z} &= \boldsymbol{\nu}_y + \Sigma_{yz}\Sigma_z^{-1}\left(Z - \boldsymbol{\nu}_z\right), \\ \Sigma_{y|z} &= \Sigma_y - \Sigma_{yz}\Sigma_z^{-1}\Sigma_{zy}. \end{aligned} \tag{2.2}$$

Let R_y denote the spatial correlation matrix of the sites of $Y(\mathbf{s}^y)$, let R_z denote the spatial correlation matrix of the sites of Z, and let R_{yz} denote the spatial correlation matrix of the sites of $Y(\mathbf{s}^y)$ and Z. Note that R_y, R_z and R_{yz} are all submatrices of R as defined in (1.2). Then the covariance matrices in (2.2) can be expressed as

$$\Sigma_y = \sigma^2(R_y + \alpha I), \quad \Sigma_z = \sigma^2(R_z + \alpha I), \quad \Sigma_{yz} = \sigma^2 R_{yz}, \quad \Sigma_{yz} = \Sigma_{zy}^T,$$

where $\alpha = \tau^2/\sigma^2$.

In the case that covariates present in model (1.1), we have

$$\boldsymbol{\nu}_{y|z} = \boldsymbol{\nu}_y + \Sigma_{yz}\Sigma_z^{-1}\left(Z - \boldsymbol{\nu}_z\right) = \left(C_y - R_{yz}R_z^{-1}C_z\right)\boldsymbol{\xi} + R_{yz}R_z^{-1}Z, \tag{2.3}$$

where $\boldsymbol{\xi} = (\xi_0, \xi_1, \ldots, \xi_p)^T$ denotes the vector of regression coefficients as defined in (1.3), and C_y and C_z are the design matrices for the covariates and given by

$$C_y = \begin{bmatrix} 1 & c_{s_1^y,1} & \cdots & c_{s_1^y,p} \\ \vdots & & \vdots & \\ 1 & c_{s_{n^*}^y,1} & \cdots & c_{s_{n^*}^y,p} \end{bmatrix}, \quad C_z = \begin{bmatrix} 1 & c_{s_1^z,1} & \cdots & c_{s_1^z,p} \\ \vdots & & \vdots & \\ 1 & c_{s_m^z,1} & \cdots & c_{s_m^z,p} \end{bmatrix},$$

where $c_{s_i^y,j}$ and $c_{s_i^z,j}$ denote the observed values of the jth covariate at the locations s_i^y and s_i^z, respectively.

For the purpose of illustration, we consider the exponential correlation function

$$\rho(\|s_i - s_j\|) = \exp\{-\|s_i - s_j\|/\phi\}, \tag{2.4}$$

where $\phi > 0$ is the correlation length parameter. Then the model (2.1)-(2.3) can be parameterized by $\boldsymbol{\theta} = \left(\theta_1, \theta_2, \sigma^2, \boldsymbol{\xi}\right) = (\log(\phi), \log(\alpha), \sigma^2, \boldsymbol{\xi})$, where, for ease of sampling, ϕ and α has been reparameterized by their logarithms. To make Bayesian inference for the model (2.1)-(2.3), we specify the following priors for $\boldsymbol{\xi}$, σ^2 and ϕ:

$$\pi(\boldsymbol{\xi}|\sigma^2) \propto \epsilon_\xi^{1+p} \sigma^{-(1+p)} \exp\left(-\epsilon_\xi^2\, \boldsymbol{\xi}^T\boldsymbol{\xi}/(2\sigma^2)\right), \;\; \pi(\sigma^2) \propto IG(\epsilon, \epsilon), \;\; \pi(\phi) \propto IG(\epsilon, \epsilon), \tag{2.5}$$

where both ϵ_ξ and ϵ are small positive constants, and IG$(\cdot, \cdot)$ denotes an inverse Gamma distribution. For simplicity, the two hyperparameters of the prior inverse Gamma distribution are restricted to be the same in this paper. When $\epsilon \leq 2$, IG(ϵ, ϵ) leads to a vague prior, whose variance is infinite.

Since it is generally true that the nugget variance τ^2 is smaller than the variance σ^2, we set a uniform prior for $\alpha = \tau^2/\sigma^2$ on the interval $[0, 1]$; that is,

$$\pi(\alpha) = 1, \quad \alpha \in [0, 1]. \tag{2.6}$$

With a little abuse of notations, we denote the model (2.1) by Z and impose a truncated Poisson prior distribution on the space of models; that is,

$$\pi(Z) \propto \frac{\lambda^m}{m!} e^{-\lambda}, \quad m \in \{0, 1, \ldots, n - n^*\}, \tag{2.7}$$

where $m = |Z|$ denotes the number of sites included in Z and λ is a hyperparameter to be specified by the user. The rationale behind this prior can be explained as follows: To minimize the loss of data information caused by site selection, Z should be selected uniformly from the observation region of $\{Y(s_i)\}$ and thus, following the standard theory of Poisson process, the number of selected sites can be modeled as a Poisson random variable. To enhance this selection pattern, the prior (2.7) is used. Alternatively, one can specify a prior distribution that incorporates the spatial information of Z, but this will complicate the simulation of the posterior distribution.

Combining (2.2)–(2.3) and (2.5)–(2.7), we have the posterior of $\boldsymbol{\theta}$ given by

$$\begin{aligned} f(\boldsymbol{\theta}|Y(\mathbf{s}^y), Z) \propto \left|\Sigma_{y|z}\right|^{-1/2} \frac{1}{\sigma^{1+p}} \exp\left\{-\frac{1}{2\sigma^2} B^T (R_{y|z} + \epsilon_\xi^{-2} AA^T)^{-1} B\right\} \pi(\theta_1, \theta_2, \sigma^2) \\ \times \exp\left\{-\frac{1}{2\sigma^2} (\boldsymbol{\xi} - \Lambda^{-1}E)^T \Lambda (\boldsymbol{\xi} - \Lambda^{-1}E)^T\right\}, \end{aligned} \tag{2.8}$$

where $A = C_y - R_{yz}R_z^{-1}C_z$, $B = Y(\mathbf{s}^y) - R_{yz}R_z^{-1}\mathbf{z}$, $E = (A\boldsymbol{\xi} - B)^T\Sigma_{y|z}^{-1}$, $R_{y|z} = \sigma^{-2}(\Sigma_y - \Sigma_{yz}\Sigma_z^{-1}\Sigma_{zy}$ denotes the conditional correlation matrix of $Y(\mathbf{s}^y)$ given Z, and $\Lambda = A^T R_{y|z}^{-1}A + \epsilon_\xi^2 I$ is an $n^* \times n^*$ matrix. It is worth pointing out that both $\Sigma_{y|z}$ and $R_{y|z} + \epsilon_\xi^{-2}AA^T$ are also $n^* \times n^*$ matrix. Thus, BSS reduces the problem of inverting $n \times n$ matrices to that of inverting $n^* \times n^*$ matrices. How to determine the value of n^* will be discussed in Section 2.2.

Integrating out $\boldsymbol{\xi}$ and σ^2 from (2.8), we have

$$f(\theta_1, \theta_2 | Y(\mathbf{s}^y), Z) \propto \left|R_{y|z}\right|^{-1/2} |\Lambda|^{-1/2} \frac{\Gamma\left(\frac{n}{2} + \epsilon\right) \pi(\theta_1, \theta_2)}{\left\{B^T(R_{y|z} + \epsilon_\xi^{-2}AA^T)^{-1}B/2 + \epsilon\right\}^{\frac{n}{2}+\epsilon}}. \tag{2.9}$$

Following the standard theory of Bayesian model averaging, the predictive posterior distribution of $Y(\mathbf{s}^p)$ can be written as

$$f(Y(\mathbf{s}^p)|Y(\mathbf{s}^y), D_{-y}) = \sum_{Z \subset D_{-y}} \int f(Y(\mathbf{s}^p)|Y(\mathbf{s}^y), Z, \boldsymbol{\theta}) f(\boldsymbol{\theta}|Y(\mathbf{s}^y), Z)\pi(Z)d\boldsymbol{\theta}, \tag{2.10}$$

where Z denotes any subset of D_{-y} and also a particular model defined in (2.1)–(2.3). This implies that the expectation of $Y(\mathbf{s}^p)$ conditioned on the full observations D is given by

$$E[Y(\mathbf{s}^p)|Y(\mathbf{s}^y), D_{-y}] = \sum_{Z \subset D_{-y}} \int E[Y(\mathbf{s}^p)|Y(\mathbf{s}^y), Z, \boldsymbol{\theta}] f(\boldsymbol{\theta}|Y(\mathbf{s}^y), Z)\pi(Z)d\boldsymbol{\theta}. \tag{2.11}$$

Let $(\boldsymbol{\theta}^{(1)}, Z^{(1)}), \ldots, (\boldsymbol{\theta}^{(N)}, Z^{(N)})$ denote a sequence of samples drawn from the joint posterior of $(\boldsymbol{\theta}, Z)$, which is proportional to $f(\boldsymbol{\theta}|Y(\mathbf{s}^y), Z)\pi(Z)$. Then $Y(\mathbf{s}^p)$ can be predicted by

$$\widehat{Y}(\mathbf{s}^p) = \frac{1}{N}\sum_{i=1}^{N} E[Y(\mathbf{s}^p)|Y(\mathbf{s}^y), Z^{(i)}, \boldsymbol{\theta}^{(i)}], \tag{2.12}$$

where $E[Y(\mathbf{s}^p)|Y(\mathbf{s}^y), Z^{(i)}, \boldsymbol{\theta}^{(i)}]$ is the conditional mean of $Y(\mathbf{s}^p)$ given $Y(\mathbf{s}^y)$, the selected set of explanatory variables $Z^{(i)}$, and the parameter values $\boldsymbol{\theta}^{(i)}$. That is,

$$\widehat{Y}(\mathbf{s}^p) = \frac{1}{N}\sum_{i=1}^{N} \left\{\boldsymbol{\nu}_{\mathbf{s}^p}^{(i)} + \Sigma_{y(\mathbf{s}^p),w_i}\Sigma_{w_i}^{-1}\left(W_i - \boldsymbol{\nu}_{w_i}\right)\right\}, \tag{2.13}$$

where $\boldsymbol{\nu}_{\mathbf{s}^p}^{(i)}$ denotes the mean of $Y(\mathbf{s}^p)$ for the sample $(\boldsymbol{\theta}^{(i)}, Z_i)$, $W_i = (Y(\mathbf{s}^y), Z^{(i)})$ is the joint vector formed by $Y(\mathbf{s}^y)$ and $Z^{(i)}$, $\Sigma_{y(\mathbf{s}^p),w_i}$ is the covariance matrix of $Y(\mathbf{s}^p)$ and W_i, Σ_{w_i} is the covariance matrix of W_i, and $\boldsymbol{\nu}_{w_i}$ denotes the mean of W_i. Note that all the terms $\boldsymbol{\nu}_{\mathbf{s}^p}^{(i)}$, $\Sigma_{y(\mathbf{s}^p),w_i}$ and Σ_{w_i} in (2.13) depend on the sample $(\boldsymbol{\theta}^{(i)}, Z_i)$, and that the covariates $c_1(\mathbf{s}^p), \ldots, c_p(\mathbf{s}^p)$ are assumed to be available at the prediction sites $\mathbf{s}^p$. How to draw samples from the joint posterior of $(\boldsymbol{\theta}, Z)$ will be discussed in Section 2.3.

2.2 Prediction-Oriented Response Variable Selection

In this section, we consider a prediction-oriented selection scheme for $Y(\mathbf{s}^y)$ with an expectation that $\{Y(\mathbf{s}^y)\}$ plays surrogates for $\{Y(\mathbf{s}^p)\}$. The scheme consists of the following steps:

1. Let $\mathbf{s} = \{s_1, \ldots, s_n\}$ denote the full set of observation sites, and let $\mathbf{s}^p = \{s_1^p, \ldots, s_{n_p}^p\}$ denote the set of prediction sites, where n_p is the total number of prediction sites.

   For $i = 1, \ldots n_p$, do the following sub-steps to identify the first tier of the nearest points to $\mathbf{s}^p$:

   (a) Draw a site s_i^p from the set $\mathbf{s}^p$ at random and without replacement.

   (b) Identify the nearest neighbor of s_i^p by setting

   $$s_{1,i}^y = \arg \min_{s \in \mathbf{s} \backslash \{s_{1,1}^y, \ldots, s_{1,i-1}^y\}} \|s - s_i^p\|.$$

   Set $\mathbf{s}_1^y = \{s_{1,1}^y, \ldots, s_{1,n_p}^y\}$.

2. Set $\mathbf{s} \leftarrow \mathbf{s} \backslash \mathbf{s}_1^y$ and repeat the substeps in step 1 to identify the second tier of the nearest points to $\mathbf{s}^p$. Denote the second tier neighboring set by $\mathbf{s}_2^y$.

   .....

k. Set $\mathbf{s} \leftarrow \mathbf{s} \backslash \mathbf{s}_{k-1}^y$ and repeat the substeps in step 1 to identify the k-th tier of the nearest points to $\mathbf{s}^p$. Denote the k-th tier neighboring set by $\mathbf{s}_k^y$.

The procedure outputs $\mathbf{s}^y = \cup_{j=1}^k \mathbf{s}_j^y$ as the set of response variables and $D_{-y} = \{s_1, \ldots, s_n\} \backslash \mathbf{s}^y$ as the set of explanatory variables. In practice, the value of k, which determines the size of $\mathbf{s}^y$ ($n^* = kn_p$), can be determined through an examination of the fitting to $\{Y(\mathbf{s}^y)\}$ or its subset. For example, we can choose the value of n^* such that the mean squared fitting errors (MSFE) for the first tier neighboring sites are minimized among a few values of n^* under consideration. Our numerical results indicate that MSFE can provide a good guideline for selection of n^*. In our experience, when $k \geq 3$, BSS often works very well provided $n \geq n^*$.

As shown in (2.8), BSS has reduced the problem of inverting $n \times n$ matrices to that of inverting $n^* \times n^*$ matrices. When n_p is large, we suggest to divide $\mathbf{s}^p$ into several small subsets and then run BSS for each of them separately. For example, the subsets can be constructed by drawing from $\mathbf{s}^p$ through a sampling-without-replacement procedure. This helps us to keep n^* in a reasonable range, and thus alleviate the heavy burden of computation caused by the cubic law of matrix inversion. In addition, the computation for different subsets can be done in parallel, which will significantly shorten our waiting time for the prediction results. Let n_p' denote the size of a subset of prediction sites. For the choice of n_p', we suggest to keep the relationship $n \geq 3n_p'$ hold, while keeping n^* in a reasonable range. In our experience, such a choice of n_p' often leads to good prediction results.

In practice, we can encounter a situation that there are no observations near some prediction sites. Since the prediction-oriented selection scheme is to select the observations nearest to the prediction sites, it still works under this situation. However, like any other approaches, BSS may produce prediction of high variability under this situation.

2.3 A Metropolis-within-Gibbs Sampling scheme

In this section, we consider a Metropolis-within-Gibbs sampler (Müller, 1991) for drawing samples from the posterior

$$f(\theta_1, \theta_2, Z|Y(\mathbf{s}^y)) \propto f(\theta_1, \theta_2|Y(\mathbf{s}^y), Z)\pi(Z),$$

where Z indexes a subset model and $f(\theta_1, \theta_2|Y(\mathbf{s}^y), Z)$ is given in (2.9).

Let $(\theta_1^{(t)}, \theta_2^{(t)}, Z^{(t)})$ denote the sample generated at iteration t of the Markov chain. Let $m = |Z^{(t)}|$ denote the number of sites included in $Z^{(t)}$. To update $Z^{(t)}$, we consider three types of moves, "birth", "death" and "exchange" with the respective proposal probabilities denoted by $q_{m,m+1}$, $q_{m,m-1}$ and $q_{m,m}$. In this paper, we set

$$\begin{aligned} q_{m_{\min},m_{\min}} &= (1/3), \quad q_{m_{\min},m_{\min}+1} = (2/3), \\ q_{m_{\max},m_{\max}} &= (1/3), \quad q_{m_{\max},m_{\max}-1} = (2/3), \\ q_{i,i+1} = q_{i-1,i} = q_{i,i} &= (1/3), \quad \text{for } m_{\min}+1 \le i \le m_{\max}-1\,, \end{aligned}$$

where $m_{\min} = 0$ and $m_{\max} = n - n^*$. One iteration of the Metropolis-within-Gibbs sampler consists of the following steps:

- Draw $\theta_1^{(t+1)}$ from the conditional distribution $f(\theta_1|\theta_2^{(t)}, Y(\mathbf{s}^y), Z)$ using the Metropolis algorithm with a random walk Gaussian proposal. The variance of this proposal is denoted by $\sigma^2_{\theta_1}$ and will be given in the context of numerical studies.

- Draw $\theta_2^{(t+1)}$ from the conditional distribution $f(\theta_2|\theta_1^{(t+1)}, Y(\mathbf{s}^y), Z)$ using the Metropolis algorithm with a random walk Gaussian proposal. The variance of this proposal is denoted by $\sigma^2_{\theta_2}$ and will be given in the context of numerical studies.

- Draw $Z^{(t+1)}$.

  – (*Birth*) Randomly select z^* out of $D_{-y}\backslash Z^{(t)}$ and set $Z^* = Z^{(t)} \cup z^*$. Set $Z^{(t+1)} = Z^*$ with probability

  $$\min\left\{1, \frac{f(\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\mathbf{s}^y), Z^*)\pi(Z^*)}{f(\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\mathbf{s}^y), Z^{(t)})\pi(Z^{(t)})} \frac{n-n^*-m}{m+1} \frac{q_{m+1,m}}{q_{m,m+1}}\right\}.$$

  Otherwise, set $Z^{(t+1)} = Z^{(t)}$.

  – (*Death*) Randomly select z^* out of $Z^{(t)}$ and set $Z^* = Z^{(t)}\backslash z^\star$. Accept $\mathbf{z}^\star_{m-1}$ with probability

  $$\min\left\{1, \frac{f(\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\mathbf{s}^y), Z^*)\pi(Z^*)}{f(\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\mathbf{s}^y), Z^{(t)})\pi(Z^{(t)})} \frac{m}{n-n^*-m+1} \frac{q_{m-1,m}}{q_{m,m-1}}\right\}.$$

  Otherwise, set $Z^{(t+1)} = Z^{(t)}$.

– (*Exchange*) Randomly select $z^\star$ out of $D_{-y}\backslash Z^{(t)}$ and $z_u^\star$ out of $Z^{(t)}$. Set $Z^* = Z^{(t)} \cup \{z^*\}\backslash\{z_u^\star\}$ by exchanging $z^\star$ and $z_u^\star$. Accept $\mathbf{z}_m^\star$ with probability

$$\min\left\{1, \frac{f(\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\mathbf{s}^y), Z^*)}{f(\theta_1^{(t+1)}, \theta_2^{(t+1)}|Y(\mathbf{s}^y), Z^{(t)})}\right\}$$

Otherwise, set $Z^{(t+1)} = Z^{(t)}$.

Given a MCMC sample $(\theta_1^{(t)}, \theta_2^{(t)}, Z^{(t)})$, $\boldsymbol{\xi}^{(t)}$ and $\sigma^{2(t)}$ can drawn from the following distributions:

$$\boldsymbol{\xi}^{(t)} \sim N\left(\Lambda^{-1}E, \Lambda^{-1}\right), \quad \sigma^{2(t)} \sim IG\left(n/2+\epsilon, B^T(R_{y|z}+\epsilon_\xi^{-2}AA^T)^{-1}B/2+\epsilon\right),$$

which can be simply derived from (2.8) with Λ, W, A, B and $R_{y|z}$ as previously defined. Given the samples $(\theta_1^{(t)}, \theta_2^{(t)}, Z^{(t)})$ and $(\boldsymbol{\xi}^{(t)}, \sigma^{2(t)})$, the prediction of $\{Y(\mathbf{s}^p)\}$ can then be simply done as in (2.13).

3 Simulation Studies

In this section, we assess the performance of BSS using two simulated examples along with some comparisons with the standard Bayesian method. For the simulated examples, we have the following common settings. In both data generation and posterior simulations, the spatial correlation function is as defined in (2.4). In posterior simulations, we set the hyperparameters $\epsilon_\xi = 0.01$ and $\epsilon = 1$. As previously explained, this leads to vague priors for $\boldsymbol{\xi}$, σ^2 and ϕ. For each dataset, BSS was run once with 10,000 iterations, where the first 5,000 iterations were discarded for the burn-in process and the remaining iterations were thinned by a factor of 5 to get 1000 samples.

3.1 An Illustrative Example

We simulated 30 independent data sets from the Gaussian geostatistical model (1.1). Each data set contains 1,100 observations with the sites uniformly distributed over the region $[0, 100] \times [0, 100]$. The data sets were generated using the function grf() in geoR (Ribeiro and Diggle, 2001) with the parameters $(\xi_0, \xi_1, \phi, \sigma^2, \tau^2) = (0.5, 1, 25, 1, 0.25)$ and the covariates generated from $N(0, 1)$. For each data set, a subset of size 1,000 was randomly selected and used for model training, and the remaining 100 samples were used for prediction.

BSS was first applied to this example with the hyperparameter $\lambda = 2$ and three different choices of $n^* = 200$, 300 and 500. In simulations, we set $\sigma_{\theta_1}^2 = 0.3$ and $\sigma_{\theta_2}^2 = 0.5$, which have been calibrated such that the Markov chain can mix well in each run. The resulting parameter estimates and mean squared prediction errors (MSPE) for the prediction set were summarized in Table 1. The numerical results indicate that as n^* increases, BSS produces better prediction. It is also interesting to point out that as n^* increases, m tends to decrease when the same value of λ is used. This is reasonable, as the response variables can explain each other in the regression model we formulated. It is known that for the model (1.1), when the correlation function is exponential or Matérn, the

parameters ϕ and σ^2 are non-estimable due to the existence of equivalent probability measures (Stein, 2004). However, in this case, the ratio ϕ/σ^2 is still estimable as shown in Zhang (2004). For this reason, we report in Table 1 the estimate of the ratio σ^2/ϕ, instead of the respective estimates of σ^2 and ϕ. Our numerical results indicate that BSS produced accurate estimates of ϕ/σ^2 for this example. As a possible tool for determining n^*, we also reported in Table 1 the mean squared fitting errors (MSFE_{t_1}) for the tier 1 neighboring observations. Apparently, MSFE_{t_1} provides a good ordering for MSPE.

Table 1: Comparison of BSS and BFD method for the illustrative example. The estimates were calculated by averaging over the results from 30 different datasets and the number in the parentheses denotes the standard deviation of the estimate. The CPU times were recorded for a single run of the algorithm on a desktop of Dual Core 3.0 GHz. BFD: Bayesian method for the full data; MSPE: mean squared prediction error; MSFE_{t_1}: mean squared fitting error for tier 1 neighbors. $\bar{m}$: average value of m obtained in simulations. Proportion: calculated in $(n^* + \bar{m})/n \times 100\%$.

	True	BSS(n^*, λ)			BFD
		$(200, 2)$	$(300, 2)$	$(500, 2)$	
$\bar{m}$	—	37(0.21)	34(0.23)	28.9(0.19)	—
Proportion	—	23.7%	33.4%	52.9%	100%
ξ_0	0.5	0.54(0.09)	0.52(0.09)	0.56(0.09)	0.42(0.00)
ξ_1	1.0	0.97(0.01)	0.99(0.02)	1.00(0.00)	0.99(0.06)
ϕ/σ^2	25	26.58(2.22)	25.67(1.94)	24.83(1.38)	23.85(0.93)
τ^2	0.25	0.23(0.01)	0.24(0.01)	0.24 (0.01)	0.25(0.01)
MSPE	—	0.413(0.01)	0.398(0.01)	0.384(0.01)	0.381(0.01)
MSFE_{t_1}	—	0.449(0.01)	0.416(0.01)	0.395(0.01)	—
CPU(h)	—	0.5	1.5	7.3	47.8

For comparison, we also applied the standard Bayesian approach to this example. This approach works on the full dataset. Letting the parameters be subject to the priors (2.5) and (2.6), and integrating out $\boldsymbol{\xi}$ and σ^2, we get the posterior

$$f(\theta_1, \theta_2|D) \propto |R + \alpha I|^{-\frac{1}{2}} |\tilde{\Lambda}|^{-\frac{1}{2}} \frac{\Gamma(\frac{n}{2} + \epsilon)}{\{\mathbf{y}^T (R + \alpha I + \epsilon_\xi^{-2} CC^T)^{-1} \mathbf{y}/2 + \epsilon\}^{\frac{n}{2}+\epsilon}} \pi(\theta_1, \theta_2), \quad (3.1)$$

where R is the correlation matrix as defined in (1.2), $\tilde{\Lambda} = C^T (R + \alpha I)^{-1} C + \epsilon_\xi^2 I$, $\mathbf{y}$ is an n-vector

which consists of all observations in D, and

$$C = \begin{bmatrix} 1 & c_{s_1,1} & \cdots & c_{s_1,p} \\ \vdots & & \vdots & \\ 1 & c_{s_n,1} & \cdots & c_{s_n,p} \end{bmatrix},$$

is the design matrix of covariates. The Metropolis-within-Gibbs sampler is also applied to simulate from the posterior (3.1), but with only two parameters θ_1 and θ_2 updated at each iteration. The algorithm was also run once for each dataset. Each run consisted of 10,000 iterations, where the first 5000 iterations were discarded for the burn-in process and 1000 samples were collected from the remaining iterations at equally-spaced time points. The resulting parameter estimates and the MSPE were reported in Table 1 in the column of BFD (Bayesian method for Full Data). The simulation is very time consuming, as it needs to invert an $n \times n$ matrix at each iteration.

The comparison indicates that although BSS costs much less CPU times than BFD, it can produce parameter estimates and prediction which both are as good as those produced by BFD. We note that the parameter estimates resultant from BSS may be biased due to the selection of $Y(\mathbf{s}^y)$ and inclusion of explanatory variables. For this example, this bias is ignorable because the prediction sites are randomly selected from the full dataset and the number of explanatory variables included in each model is relatively small. How to use BSS for parameter estimation will be discussed in the Discussion section.

To understand why BSS works so well in both prediction and estimation, we conduct the following experiment to test if BSS can catch the long range dependence of the data. The experiment was done in the following procedure:

- For each sample in D_{-y} find its minimum distance to $\mathbf{s}^y$; that is, set

  $$d(s) = \min_{s_i^y \in \mathbf{s}^y} \|s - s_i^y\|,$$

  for each site $s \in D_{-y}$.

- Divide the samples in D_{-y} into 10 groups according to $d(s)$. Group 1 contains the one-tenth samples with the smallest values of $d(s)$, ..., and Group 10 contains one-tenth samples with the largest values of $d(s)$.

- Run BSS with $n^* = 500$ and $\lambda = 2$ for one dataset.

- Count the sampling frequency of the explanatory variables Z from each group.

Figure 1 shows the relative sampling frequency of the explanatory variables Z from each group. All 10 groups have more or less same relative frequencies and the highest is obtained for group 1. This indicates that BSS is indeed able to catch the long range dependence of the data. Therefore, it is understandable why BSS performs like BFD in estimation and prediction even with only a subset of the data being used. It is also reasonable that group 1 has the highest relative frequency, as the samples in group 1 have higher correlations with the response samples than those in other groups.

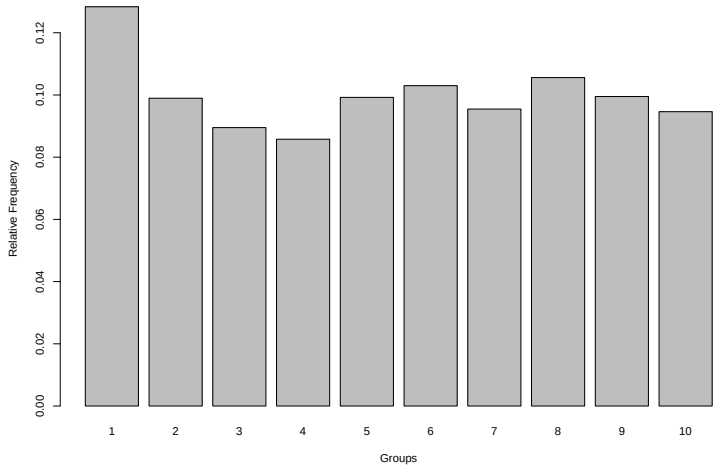

Figure 1: Sampling frequency of the explanatory variables Z drawn by BSS for one dataset with $n^* = 500$ and $\lambda = 2$.

To assess the sensitivity of BSS to the choice of λ, we tried different values of $\lambda = 1$, 2, 3, 5, and 10 for the case $n^* = 200$. The results were summarized in Table 2. The results indicate that as λ increases, the number of explanatory variables included in the model tends to increase, the resulting regression model tends to be overfitted (the estimate of τ^2 tends to decrease slightly) and the contribution of covariates to the regression model tends to decrease (the estimate of ξ_1 tends to decrease). This experiment suggests that a small value of λ may be used, which will lead to a parsimony regression model in general.

In summary, the numerical results of this example suggests us to choose a reasonably large value of n^* within the limit of our computer power, as a large value of n^* can generally work better in both parameter estimation and prediction. However, an excessively large value of n^* is not necessary, especially when one aims at prediction, as the prediction accuracy depends mainly on the neighbors of the prediction site. In practice, the value of n^* can be determined according to the value of MSFE_{t_1}. When n^* is reasonably large, say, the tier 3 neighboring points have been included in the response, a small value of λ, say, 1 or 2, may be used.

3.2 A Large Data Example

To assess the performance of BSS for large spatial data, we simulated 30 independent datasets from the model (1.1) with the same parameters as for the last example. Each dataset contains 20,100 samples, where 100 randomly selected samples were used for prediction and the remaining 20,000 samples were used for model building.

BSS was applied to this example with $\sigma^2_{\theta_1} = \sigma^2_{\theta_2} = 0.3$, $\lambda = 1$, and $n^* = 300$, 500 and 700.

Table 2: Sensitivity analysis of BSS for the value of λ. Refer to Table 1 for the notation.

	BSS(n^*, λ)				
	$(200, 1)$	$(200, 2)$	$(200, 3)$	$(200, 5)$	$(200, 10)$
$\bar{m}$	26.5(0.17)	37(0.21)	45.8(0.27)	58(0.34)	82 (0.43)
Proportion	22.7%	23.7%	24.6%	25.8%	28.2%
ξ_0	0.52(0.08)	0.54(0.09)	0.53(0.09)	0.52(0.09)	0.50 (0.09)
ξ_1	0.98(0.01)	0.97(0.01)	0.96(0.01)	0.95(0.01)	0.94 (0.01)
ϕ/σ^2	26.06(2.16)	26.58(2.22)	25.11(2.39)	25.67(2.28)	24.52 (2.41)
τ^2	0.25(0.01)	0.23(0.01)	0.24(0.01)	0.23(0.01)	0.22 (0.10)
MSPE	0.414(0.01)	0.413(0.01)	0.414(0.01)	0.414(0.01)	0.415(0.01)

The results were summarized in Table 3. The performance of BSS for this example is similar to that of the last example. It produced very reasonable parameter estimates and MSPE values. For this example, we also calculated MSFE_{t_1}. The results indicate again that MSFE_{t_1} is highly correlated with MSPE and can be used as a tool for choosing appropriate settings for BSS. It is worth pointing out that for this example, even with only less than 5% (on average) of samples being used at each iteration, BSS still performs reasonably well in both parameter estimation and prediction. BSS can have many applications. Recently, it has been applied to Gaussian process regression by the authors.

4 Real Data Study

4.1 Precipitation Anomaly Data

To demonstrate the performance of *BSS* for real problems, we considered a precipitation dataset from the National Climatic Data Center (NCDC) for the years 1895 to 1997. This data has been studied by many authors including Johns *et al.* (2003), Furrer *et al.* (2006), and Kaufman *et al.* (2008), among others. In this study, following Kaufman *et al.* (2008), we use the precipitation anomalies of 1962, which are available at `http://www.image.ucar.edu/Data/precip_tapering/`. This dataset consists of 7,352 samples (sites) and, as mentioned by Kaufman *et al.* (2008), there is no noticeable evidence for nonstationarity.

For this example, we randomly choose a subset of 250 out of 7,352 samples for model testing, and use the remaining samples for model building. We tried different values of $n^* = 250$, 500 and 750. Since our results reported in the previous section indicate that BSS is not sensitive to the value of λ, we set $\lambda = 1$ for this example. For each value of n^*, BSS was run 5 times independently with $\sigma^2_{\theta_1} = \sigma^2_{\theta_2} = 0.3$. Each run consisted of 10,000 iterations, with the first 5,000 iterations being discarded for the burn-in process and 1000 samples being collected from remaining 5,000 iterations

Table 3: Performance of BSS for the large data example. Refer to Table 1 for the notation.

	BSS(n^*, λ)		
	$(300, 1)$	$(500, 1)$	$(700, 1)$
$\bar{m}$	136(0.68)	134(0.98)	133(0.78)
Proportion	2.18%	3.17%	4.17%
ξ_0	0.665(0.100)	0.687(0.098)	0.705(0.096)
ξ_1	0.967(0.010)	0.985(0.006)	0.990(0.004)
ϕ/σ^2	23.05(1.86)	22.96(1.52)	23.39(1.58)
τ^2	0.228(0.007)	0.232(0.006)	0.237(0.004)
MSPE	0.345(0.00)	0.326(0.00)	0.316(0.00)
MSE_{t_1}	0.343(0.00)	0.320(0.00)	0.305(0.00)
Time(hr)	2.6	11.0	21.9

at equally spaced time points. The results were summarized in Table 4.

Table 4 shows an interesting pattern: The estimate of ϕ/σ^2 tends to decrease as n^* increases. This is reasonable. When $n^* = 250$, D_y consists of only the tier 1 sites, which are far from each other. To establish the dependence among these sites, a large value of ϕ/σ^2 is needed. When n^* increases, the estimate of ϕ/σ^2 will converge to its true value. However, as long as n^* is reasonably large, say, $n^* \geq 3n_p$, BSS will perform very well in prediction. The reason is that the sparsity of neighboring information can be partially compensated by the updated parameter estimates. Table 4 shows that BSS produced similar prediction results with $n^* = 500$ and $n^* = 750$ in terms of MSPE. Based on this observation, we conclude that BSS is a useful approach for prediction.

To show that BSS can produce reasonable parameter estimates for model (1.1), we compare the predicted anomalies on a regular grid of 500×400 with the unit grid size (longitude $\times$ latitude) 0.065×0.12, where the anomalies were predicted using the covariance tapering method (Furrer *et al.*, 2006) with the BSS estimates given in Table 4. In this study, we tapered the estimated covariance matrices by a spherical family with a range of 50 miles. The results were shown in Figure 2. The BSS prediction matches with observations very well, even for the case with $n^* = 250$. This indicates that the estimates produced by BSS are reasonable for this data. It needs to emphasize that BSS uses only a small proportion of the data at its each iteration.

4.2 Gold Mine Data

The Gold mine data, available at `http://www.kriging.com/datasets/`, is constructed based on a Wits type gold mine. The samples are chipped from the face of the reef in a working section of the mine (stope). As the face advances, new chip samples are taken. Values within a stope

Table 4: BSS results for the anomalies of 1962. The estimates were calculated by averaging over the results of 5 independent runs, with their standard errors given in the parenthesis. The CPU times were recorded for a single run on a Desktop of Dual Core 3.0 GHz. Proportion was calculated in $(n^* + \bar{m})/n \times 100\%$.

	BSS(n^*, λ)		
	$(250, 1)$	$(500, 1)$	$(750, 1)$
$\bar{m}$	89(0.65)	90(0.59)	89(0.67)
Proportion	4.61%	8.03%	11.41%
ξ_0	-0.046(0.013)	-0.076(0.005)	-0.08(0.00)
ϕ/σ^2	206.16(11.10)	196.59(1.10)	172.76(1.93)
τ^2	0.096(0.011)	0.123(0.001)	0.112(0.001)
$MSPE$	0.320(0.003)	0.272(0.001)	0.272(0.000)
Time(hr)	1.3	7.8	24.8

are traditionally estimated using the sample values from the face. The data set was used in Clark and Harper (2000). To ensure the data normality holds for model (1.1), we work on the logarithm of the observations.

The data set consists of 21,577 observations. We randomly select 250 observations for model testing and use the remaining observations for model building. BSS was run for 5 times independently with $\sigma^2_{\theta_1} = 0.2$ and $\sigma^2_{\theta_2} = 0.3$. Each run consists of 10,000 iterations, where the first 5,000 iterations were discarded for the burn-in and 1,000 samples were collected from the remaining iterations at equally-spaced time points. The numerical results were summarized in Table 5.

Table 5 shows a similar pattern to Table 4: As n^* increases, the estimate of ϕ/σ^2 tends to decrease. Figure 3 shows the images of the observations and prediction surfaces. It indicates again that BSS can produce reasonable parameter estimates for model (1.1), even with only a small proportion (less than 5%) of the data being used at each iteration.

5 Discussion

In this paper, we have proposed a prediction-oriented BSS approach for dealing with the large matrix inverse problem encountered in geostatistics. The BSS approach works by performing a regression analysis based on the prediction request, with the data dimension being reduced through a stochastic variable selection procedure. Like other dimension reduction approaches, such as those proposed by Banerjee *et al.* (2008), Cressie and Johannesson (2008), Finley *et al.* (2009) and Stein (2008), BSS tries to make use of all data available. In BSS, this is done through Bayesian model averaging. By averaging over the outputs of the models built with different sets of explanatory vari-

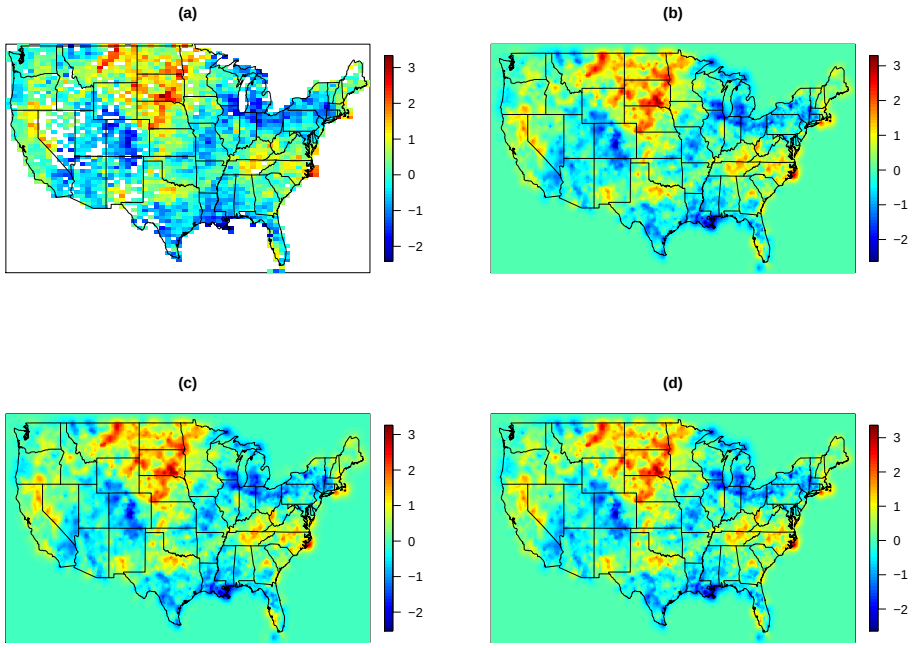

Figure 2: Images of observed and predicted anomalies of 1962 on a regular grid of size 500×400. (a) Observed anomalies; (b) prediction surface for $n^* = 250$; (c) prediction surface for $n^* = 500$; (d) prediction surface for $n^* = 750$.

ables, BSS can essentially incorporate all data information into the resulting parameter estimates and future value prediction. Our simulated examples show that with an appropriate choice of response variables and an appropriate choice of λ, BSS can produce parameter estimates and prediction which both are nearly as good as those produced by the Bayesian method with the full data, although BSS uses only a small proportion of the data at each iteration. For a really large data set, say, the number of observations is over 20,000, our numerical results (of the 2nd simulated example and the 2nd real example) indicate that BSS can produce very reasonable parameter estimates and predictions with only less than 5% of the data used at each iteration.

As previously mentioned, the parameter estimates produced by BSS can be biased due to the choice of the response variables and inclusion of explanatory variables. For example, when the response variables are not uniformly selected from the set of observations and the number of explanatory variables included in the regression is too large, the resulting parameter estimates may be biased. To address this issue, we propose an ensemble BSS approach, which works in a style of bootstrap sampling as follows:

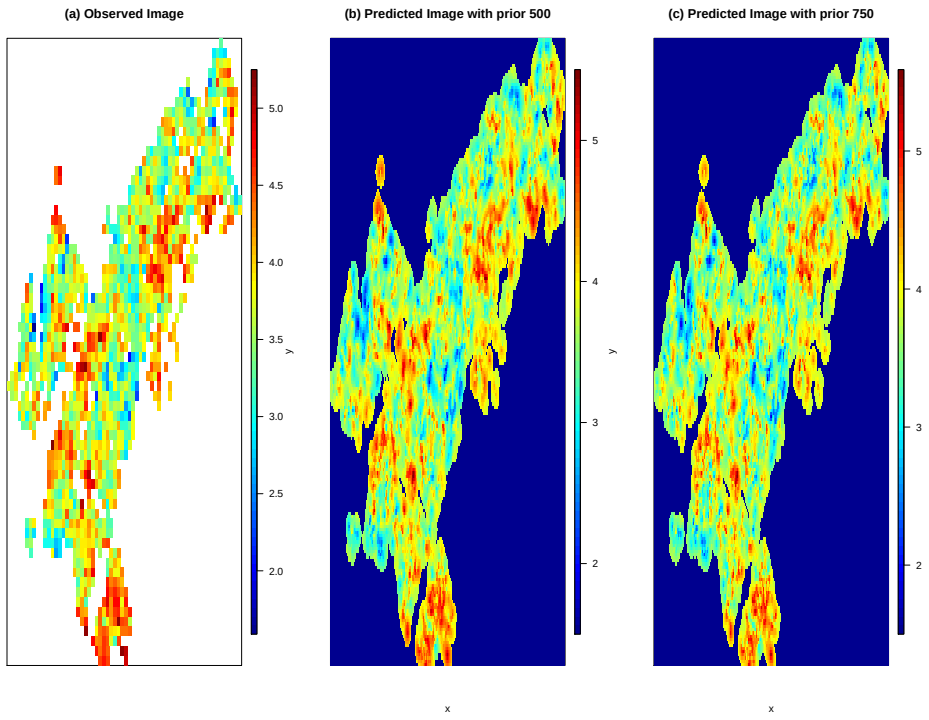

Figure 3: Images of observations and predicted surfaces on a regular grid of size 300×200 for the goldmine data. The prediction surfaces were produced by local Kriging for which each grid point is predicted based on the nearest 100 points. (a) Images of observations; (b) prediction surface by the BSS estimate with $n^* = 500$; and (c) prediction surface by the BSS estimate with $n^* = 750$.

Table 5: BSS results for the gold mine data. The estimates were calculated by averaging over the results of 5 independent runs, with their standard errors given in the parenthesis. The CPU times were recorded for a single run on a Desktop of Dual Core 3.0 GHz. Proportion was calculated in $(n^* + \bar{m})/n \times 100\%$.

	BSS(n^*, λ)	
	$(500, 1)$	$(750, 1)$
$\bar{m}$	151(0.76)	152(1.38)
Proportion	3.02%	4.18%
ξ_0	3.76(0.003)	3.77(0.002)
$\phi/\sigma 2$	99.45(1.08)	71.21(1.19)
$\tau 2$	0.098(0.001)	0.058(0.001)
$MSPE$	0.154(0.000)	0.139(0.000)
Time(hr)	9.4	28.0

- Select multiple response sets, with each being drawn randomly from the set of observations.
- Run BSS for each response set.
- Average the parameter estimates resultant from each response set.

In this case, the hyperparameter λ may be set to a small number or even zero, as one aims at parameter estimation instead of prediction. Following from the standard theory of bootstrap (Efron and Tibshirani, 1993), the parameter estimates resultant from the ensemble BSS approach is unbiased.

Acknowledgment

The authors thank the editor, associate editor and two referees for their comments which led to significant improvement of this paper. Liang's research was partially supported by grants from the National Science Foundation (DMS-1007457 and DMS-1106494) and the award (KUS-C1-016-04) made by King Abdullah University of Science and Technology (KAUST).

References

[1] Banerjee, S., Gelfand, A.E., Finley, A.O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc. B,* **70**, *825-848.*

[2] Besag, J., Modal, D. (2005). First-order intrinsic autoregressions and the de Wijs process. *Biometrika,* **92**, *909-920.*

[3] Billings, S.D., Newsam, G.N., Beatson, R.K. (2002). Gaussian predictive process models for large spatial data sets. *Geophysics,* **67**, *1823-1834.*

[4] Clark, I. and Harper, V.W. (2000). *Practical Geostatistics. Ecosse North America Llc.*

[5] Cressie, N.A.C. (1993). *Statistics for Spatial Data, 2nd edition, Wiley, New York.*

[6] Cressie, N, Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc. B,* **70**, *209-226.*

[7] Diggle, P.J., Tawn, J.A., and Moyeed, R.A. (1998). Model based geostatistics (with discussion). *Appl. Statist.,* **47**, *299-350.*

[8] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap. Chapman & Hall/CRC.*

[9] Finley, A.O., Sang, H., Banerjee, S., and Gelfand, A.E. (2009). Improving performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis,* **53**, *2873-2884.*

[10] Furrer, R., Genton, M. G. and Nychka, D. (2006). Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics,* **15(3)**, *502-523.*

[11] Jones, C.J., Nychka, D., Kittel, T.G.T. and Daly, C. (2003). Infilling Sparse Records of Spatial Fields, *J. Am. Statist. Ass.,* **98**, *796-806.*

[12] Jones, R.H. and Zhang, Y. (1997). Models for continuous stationary space-time processes. In *Modeling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions (eds P.J. Diggle, W.G. Warren and R.D. Wolfinger). New York: Springer*.

[13] Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. *Appl. Statist.,* **52**, *1-18*

[14] Kaufman, C., Schervish, M., and Nychka, D. (2008). Covariance Tapering for Likelihood-Based Estimation in Large Spatial Datasets. *J. Am. Statist. Ass.,* **103**, *1156-1569.*

[15] Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (2000). Smoothing spline ANOVA models for large datasets with Bernoulli observations and the randomized GACV. *Ann. Statist.,* **28**, *1570-1600.*

[16] Lindgren, F., Rue, H., Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: SPDE approach. *J. R. Statist. Soc. B,* **73**, *423-498.*

[17] Müller, P. (1991). A generic approach to posterior integration and Gibbs sampling. Technical report, Purdue University, West Lafayette, Indiana (1991).

[18] Park, J. and Liang, F. (2012). Bayesian Analysis of Geostatistical Models with an Auxiliary Lattice. *Journal of Computational and Graphical Statistics.*, **21**, *453-475*.

[19] Paciorek, C.J. (2007). Computational techniques for spatial logistic regression with large datasets. *Computational Statistics and Data Analysis*, **51**, *3631-3653.*

[20] Ribeiro Jr., P.J. and Diggle, P.J.(2001) *geoR: A package for geostatistical analysis*. R-NEWS Vol1, No 2. ISSN 1609-3631.

[21] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC.*

[22] Rue H., Martino S. and Chopin N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *J. R. Statist. Soc. B*, **71**, *319392*.

[23] Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian field. *Scan. J. Statist.*, **29**, *31-49.*

[24] Stein, M.L. (2008). A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society*, **37**, *3-10.*

[25] Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory of Kriging. New York: Springer.*

[26] Stein, M.L. Chi, Z. and Welty, L.J. (2004). Approximating likelihoods for large spatial data sets. *J. R. Statist. Soc. B*, **66**, *275-296.*

[27] Vecchia, A.V. (1988). Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc. B*, **50**, *297-312.*

[28] Wikle, C. and Cressie, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**, *815-829.*

[29] Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Am. Statist. Ass.*, **99**, *250-261*.