Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Hierarchical Low Rank Approximation of Likelihoods for Large Spatial Datasets

Huang Huang & Ying Sun

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2018, VOL. 27, NO. 1, 110–118
https://doi.org/10.1080/10618600.2017.1356324

Taylor & Francis
Taylor & Francis Group

Hierarchical Low Rank Approximation of Likelihoods for Large Spatial Datasets

Huang Huang and Ying Sun

CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ABSTRACT

Datasets in the fields of climate and environment are often very large and irregularly spaced. To model such datasets, the widely used Gaussian process models in spatial statistics face tremendous challenges due to the prohibitive computational burden. Various approximation methods have been introduced to reduce the computational cost. However, most of them rely on unrealistic assumptions for the underlying process and retaining statistical efficiency remains an issue. We develop a new approximation scheme for maximum likelihood estimation. We show how the composite likelihood method can be adapted to provide different types of hierarchical low rank approximations that are both computationally and statistically efficient. The improvement of the proposed method is explored theoretically; the performance is investigated by numerical and simulation studies; and the practicality is illustrated through applying our methods to two million measurements of soil moisture in the area of the Mississippi River basin, which facilitates a better understanding of the climate variability. Supplementary material for this article is available online.

ARTICLE HISTORY
Received July 2016
Revised June 2017

KEYWORDS
Gaussian process models; Likelihood approximation; Matérn covariance function; Soil moisture; Statistical efficiency

1. Introduction

Soil moisture is a key factor in climate systems, which has a significant impact on hydrological processes, runoff generations, and drought developments. To understand its spatial variability and predict values at unsampled locations, Gaussian process models are widely used (Stein 1999), where likelihood-based methods are appropriate for model fitting. However, it generally requires $O(n^3)$ computations and $O(n^2)$ memory for n irregularly spaced locations (Sun and Stein 2014). Similar to other climate variables, many satellite-based or numerical model generated soil moisture datasets have nearly a global coverage with high spatial resolutions, so that the exact computation of Gaussian likelihood becomes prohibitive. There are various existing methods, many of which were discussed by Sun, Li, and Genton (2012). For example, covariance tapering (Furrer, Genton, and Nychka 2006; Kaufman, Schervish, and Nychka 2008; Sang and Huang 2012) assumes a compactly supported covariance function, which leads to a sparse covariance matrix; low rank models, including space-time Kalman filtering (Wikle and Cressie 1999), low rank splines (Lin et al. 2000), moving averages (Ver Hoef, Cressie, and Barry 2004), predictive processes (Banerjee et al. 2008), and fixed rank kriging (Cressie and Johannesson 2008), make use of a latent process with a lower dimension where the resulting covariance matrix has a low rank representation; and Markov random field models (Cressie 1993; Rue and Tjelmeland 2002; Rue and Held 2005; Lindgren, Rue, and Lindström 2011) exploit fast-approximated conditional distributions assuming conditional independence with the precision matrix being sparse. These methods use models that may allow exact computations to reduce computations and/or storage, and each has its strength and weakness. For instance, Stein (2013) studied the properties of the covariance tapers and showed that covariance tapering sometimes performs even worse than assuming independent blocks in the covariance; Stein (2014) discussed the limitations on the low rank approximations; and Markov models depend on the observation locations, and realignment to a much finer grid with missing values is required for irregular locations (Sun and Stein 2014). Recently developed methods include the nearest-neighbor Gaussian process model (Datta et al. 2016), which is used as a sparsity-inducing prior within a Bayesian hierarchical modeling framework, the multiresolution Gaussian process model (Nychka et al. 2015), which constructs basis functions using compactly supported correlation function on different level of grids, equivalent kriging (Kleiber and Nychka 2015), which uses an equivalent kernel to approximate the kriging weight function when a nontrivial nugget exists, and multi-level restricted Gaussian maximum likelihood method (Castrillón-Candás, Genton, and Yokota 2016), for estimating the covariance function parameters using contrasts.

An alternative way to reduce computations is via likelihood and score equation approximations. Vecchia (1988) first proposed to approximate the likelihood using the composite likelihood method, where the conditional densities were calculated by choosing only a subset of the complete conditioning set. Stein, Chi, and Welty (2004) adapted this method for restricted maximum likelihoods approximation. Instead of approximating the likelihood itself, Sun and Stein (2016) proposed new unbiased estimating equations for score equation approximation, where the sparse precision matrix approximation is constructed by a similar method. In these approximation methods, the exact likelihood and the score equations can be obtained by using the complete conditioning set to calculate each conditional density. It was shown that the approximation quality or the statistical

CONTACT Ying Sun ✉ ying.sun@kaust.edu.sa CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

efficiency depends on the selected size of the subset. It is common that the subset is still inadequate by considering the largest possible number of nearest neighbors, which motivates this work.

In this article, we propose a generalized hierarchical low rank method for likelihood approximation. The proposed method uses low rank approximations hierarchically, which does not lead to a low rank covariance matrix approximation. Therefore, it is different from the predictive process method (Banerjee et al. 2008), where the covariance matrix is approximated by a low rank representation. Furthermore, the proposed method contains the independent blocks (Stein 2013) and nearest neighbors (Stein, Chi, and Welty 2004) approaches as special cases. The improvement of the proposed method is explored theoretically and the performance is investigated by numerical and simulation studies. We show that the hierarchical low rank approximation significantly improves the statistical efficiency of the most commonly used methods while retaining the computational efficiency, especially when the size of conditional subsets is restricted by the computational capacity, which is always the case for real datasets. For illustrations, our method is applied to a large real-world spatial dataset of soil moisture in the Mississippi River basin, U.S.A., to facilitate a better understanding of the hydrological process and climate variability. Our method is able to fit a Gaussian process model to two million irregularly spaced measurements with fast computations, making it practical and attractive for very large datasets.

2. Methodology

2.1. Approximating Likelihoods

Let $\{z(s) : s \in D \subset \mathbb{R}^d\}$ be a stationary isotropic Gaussian Process in a domain D in the d-dimensional Euclidean space, and typically $d = 2$. We assume the mean of the process is zero for simplicity and the covariance function has a parametric form $C(h; \theta) = \text{cov}\{z(s), z(s')\}$, where $h = \|s - s'\|$ and θ is the parameter vector of length p. Suppose that data are observed at n irregularly spaced locations $s_1, \ldots, s_n$, then,

$$Z = (z_1, \ldots, z_n)^{\mathrm{T}} \sim N(0, \Sigma(\theta)),$$

where $z_i = z(s_i), i = 1, \ldots, n$, and $\Sigma(\theta)$ is the variance-covariance matrix with the (i, j)th element $C(\|s_i - s_j\|; \theta)$. For simplicity, θ is omitted in notations hereinafter unless clarification is needed.

The maximum likelihood estimate can be obtained by maximizing the log-likelihood,

$$\ell(\theta \mid Z) = \log\{f(Z \mid \theta)\} = -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}Z^{\mathrm{T}}\Sigma^{-1}Z - \frac{n}{2}\log(2\pi),$$

where f is the multivariate normal density. In practice, if the mean of Z is a vector that depends linearly on unknown parameters, the restricted maximum likelihood estimate should be employed (Stein, Chi, and Welty 2004).

When computations become prohibitive, one way to approximate the likelihood is through log-conditional densities,

$$\ell(\theta \mid Z) = \log\{f(z_1 \mid \theta)\} + \sum_{j=1}^{n-1} \log\{f(z_{j+1} \mid Z_j, \theta)\},$$

where $Z_j = (z_1, \ldots, z_j)^{\mathrm{T}}$, for $1 \le j \le n-1$, indicating all the "past" observations of z_{j+1}. Since,

$$\text{cov}\begin{pmatrix} Z_j \\ z_{j+1} \end{pmatrix} = \begin{pmatrix} \Sigma_{jj} & \sigma_j \\ \sigma_j^{\mathrm{T}} & \sigma_{j+1,j+1} \end{pmatrix},$$

it is easy to show that for $j = 1, \ldots, n-1$,

$$\log\{f(z_{j+1}|Z_j)\} = -\frac{1}{2}\left\{\frac{\left(z_{j+1} - \sigma_j^{\mathrm{T}}\Sigma_{jj}^{-1}Z_j\right)^2}{\sigma_{j+1,j+1} - \sigma_j^{\mathrm{T}}\Sigma_{jj}^{-1}\sigma_j} + \log\left(\sigma_{j+1,j+1} - \sigma_j^{\mathrm{T}}\Sigma_{jj}^{-1}\sigma_j\right) + \log(2\pi)\right\}, \tag{1}$$

which is the log-density of $w_j = b_j^{\mathrm{T}}Z$, where $b_j = (-\sigma_j^{\mathrm{T}}\Sigma_{jj}^{-1}, 1, 0, \ldots, 0)^{\mathrm{T}}$. It can be shown that w_j's are independent and $w_j \sim N(0, v_j)$, where $v_j = b_j^{\mathrm{T}}\Sigma b_j$ (Stein, Chi, and Welty 2004). Sun and Stein (2016) further showed that the precision matrix is $\Sigma^{-1} = \sum_{j=0}^{n-1} b_j b_j^{\mathrm{T}}/v_j$, where $b_0 = (1, 0, \ldots, 0)^{\mathrm{T}}$ and $v_0 = b_0^{\mathrm{T}}\Sigma b_0$. More generally, z_{j+1} can be defined as a vector, which is usually more computationally efficient, and the corresponding $b_j = (-\sigma_j^{\mathrm{T}}\Sigma_{jj}^{-1}, I, 0, \ldots, 0)^{\mathrm{T}}$, where I is an identity matrix of size that equals the length of z_{j+1}.

However, for a large j, it is computationally expensive to evaluate $\Sigma_{jj}^{-1}\sigma_j$. Vecchia (1988) proposed approximating each conditional density by only conditioning on a subset z_{j+1} consisting of $r \ll j$ nearest neighbors. The same approach is used by Stein, Chi, and Welty (2004) for approximating the restricted maximum likelihood estimate. Sun and Stein (2016) also used the subset of nearest neighbors to approximate the precision matrix for score equation approximation.

In this article, we propose a generalized framework that allows to approximate these conditional densities hierarchically using a low rank representation. Although we implement our algorithm for application in Section 5 with z_{j+1} being a vector, we present and illustrate our methodology assuming z_{j+1} is scalar for simplicity.

2.2. Hierarchical Low Rank Representation

Motivated by the nearest neighbors method, where only $r \ll j$ nearest neighbors are selected to approximate $\Sigma_{jj}^{-1}\sigma_j$ for a large j in Equation (1), we propose a general approximation framework for $j > r$ using a low rank representation. We call each step approximating the conditional density $f(z_{j+1} \mid Z_j)$ a hierarchy, and the hierarchical low rank representation indicates that a low rank approximation is used in each hierarchy. Therefore, this proposed hierarchical approximation framework is different from Bayesian hierarchical models or fast algorithms developed for hierarchical matrices in numerical linear algebra, where a dense matrix is represented by low rank matrices at each level of a tree structure (Ambikasaran et al. 2016).

Denote $\Sigma_{jj}^{-1}\sigma_j$ by x_j, or $\Sigma_{jj}x_j = \sigma_j$. We propose to approximate x_j by a low rank representation $\hat{x}_j = A_{j,r}\tilde{x}_j$, where $\tilde{x}_j$ is a vector of length r and $A_{j,r}$ is a $j \times r$ matrix. Then, instead of solving $\Sigma_{jj}x_j = \sigma_j$, we minimize the norm $\|\Sigma_{jj}A_{j,r}\tilde{x}_j - \sigma_j\|_{\Sigma_{jj}^{-1}} = (\Sigma_{jj}A_{j,r}\tilde{x}_j - \sigma_j)^{\mathrm{T}}\Sigma_{jj}^{-1}(\Sigma_{jj}A_{j,r}\tilde{x}_j - \sigma_j)$ or equivalently solve

$A_{j,r}^{\mathrm{T}}\Sigma_{jj}A_{j,r}\tilde{x}_j = A_{j,r}^{\mathrm{T}}\sigma_j$. Therefore, x_j is approximated by

$$\hat{x}_j = A_{j,r}\tilde{x}_j = A_{j,r}\left(A_{j,r}^{\mathrm{T}}\Sigma_{jj}A_{j,r}\right)^{-1}A_{j,r}^{\mathrm{T}}\sigma_j, \quad (2)$$

which only involves a linear solve of dimension r. In this framework, we approximate x_j for each $j > r$ hierarchically by a low rank representation, which includes many commonly used strategies as special cases with different choices of $A_{j,r}$. The following are some examples:

Example 1. Independent blocks method (IND). In this method, no correlation between "past" points and the "current" point is considered. Namely, $A_{j,r}$ is a 0 matrix; however, z_{j+1} is a vector of length r here for fair comparison to other methods in terms of computation.

Example 2. Nearest neighbors method (NN). Choose r nearest neighbors of z_{j+1} from Z_j. The corresponding $A_{j,r}$ is of $j \times r$ dimensions, where each column consists of only one element 1 at the kth row if z_k is selected from Z_j and zero otherwise.

Example 3. Nearest neighboring sets method (SUM). Choose r nearest neighboring sets of z_{j+1}, where each set contains $m > 1$ neighbors and a total of $mr \ll j$ neighbors are selected from Z_j. The matrix $A_{j,r}$ is specified as a $j \times r$ matrix with each column having m elements of 1, indicating the sum of the m selected neighbors are considered. In this way, more neighbors are included while the computational cost remains the same.

Example 4. Nearest neighbors and nearest neighboring sets method (NNSUM). Combine Examples 2 and 3, where r_1 columns of $A_{j,r}$ are constructed as in Example 2, and $r - r_1$ are built as in Example 3. In this way, we use the exact information from the r_1 nearest neighbors and consider $r - r_1$ nearest neighboring sets with a total number of $r_1 + m(r - r_1)$ selected nearest neighbors.

2.3. *Hierarchical Low Rank Approximation Method*

In this section, we propose a generalized hierarchical low rank approximation method (HLR). In Equation (2), the matrix $A_{j,r}$ is a 0–1 matrix. The $r \times r$ matrix $A_{j,r}^{\mathrm{T}}\Sigma_{jj}A_{j,r}$ only extracts the corresponding rows or columns of Σ_{jj}. Now suppose we select mr nearest neighbors of z_{j+1}, and the corresponding $A_{j,mr}$ is of size $j \times mr$. To retain the same computational costs associated with rank r, we propose the following approximation:

$$A_{j,mr}^{\mathrm{T}}\Sigma_{jj}A_{j,mr} \approx P_jL_jP_j^{\mathrm{T}} + \epsilon_j^2 I_{mr}, \quad (3)$$

where L_j is a positive definite matrix of dimension $r \times r$, P_j is an $mr \times r$ matrix consisting of r basis functions, I_{mr} is the identity matrix of size mr, and ϵ_j^2 accounts for the fine-scale variability, or the reminders of the low rank approximation so that the approximated matrix is invertible. By the Sherman–Morrison–Woodbury formula,

$$\left(P_jL_jP_j^{\mathrm{T}} + \epsilon_j^2 I_{mr}\right)^{-1} = \epsilon_j^{-2}I_{mr} - \epsilon_j^{-4}P_j\left(L_j^{-1} + \epsilon_j^{-2}P_j^{\mathrm{T}}P_j\right)^{-1}P_j^{\mathrm{T}}, \quad (4)$$

then $(A_{j,mr}^{\mathrm{T}}\Sigma_{jj}A_{j,mr})^{-1}$ in Equation (2) can be approximated by inverting only an $r \times r$ matrix L_j.

This approach is similar to the predictive process (Banerjee et al. 2008) and fixed rank kriging (Cressie and Johannesson 2008). However, both methods approximate the covariance function by a low rank representation while the low rank approximation is done for each $j > r$ hierarchy in our method, and the resulting approximated covariance is no longer low rank. For the choice of P_j, Cressie and Johannesson (2008) discussed several popular options. In this article, we use P_j consisting of eigenvectors associated with the first r eigenvalues of $A_{j,mr}^{\mathrm{T}}\Sigma_{jj}A_{j,mr}$, and ϵ_j is then chosen to be the $(r+1)$th eigenvalue. Specifically, suppose $A_{j,mr}^{\mathrm{T}}\Sigma_{jj}A_{j,mr} = \sum_{i=1}^{mr}\lambda_i u_i u_i^{\mathrm{T}}$, where $\lambda_1 \geq \cdots \geq \lambda_{mr}$ are the eigenvalues and u_i's are the corresponding eigenvectors. Let $P_j = (u_1, \ldots, u_r)$ be the $mr \times r$ matrix with orthogonal columns. By choosing $\epsilon_j = \lambda_{r+1}$ and $L_j = \mathrm{diag}(\lambda_1 - \lambda_{r+1}, \ldots, \lambda_r - \lambda_{r+1})$, it is easy to derive that the low rank approximation error is $\sum_{i=r+2}^{mr}(\lambda_i - \lambda_{r+1})u_i u_i^{\mathrm{T}}$. Although it requires extra time to obtain these eigenvalues and eigenvectors, the computation is relatively cheap because the matrix size mr is generally small and we only need to compute $r+1$ leading eigenvalues and r leading eigenvectors without solving the entire eigen-decomposition.

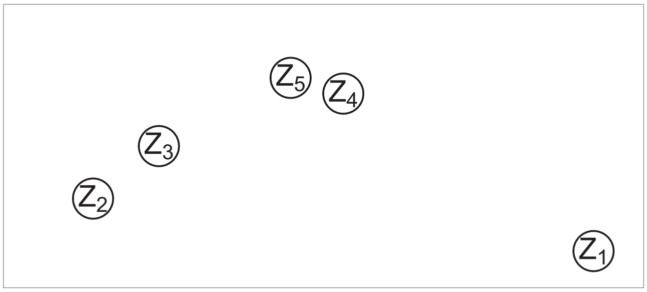

Figure 1. A random field where $n = 5$ locations have observations.

To help comprehend, Figure 1 illustrates the methods described in Sections 2.2 and 2.3 for $n = 5$ observations $Z = (z_1, \ldots, z_5)^{\mathrm{T}}$. Let $r = 2$, then IND considers three independent blocks, and $f(Z)$ is approximated by $f(z_5)f(z_4, z_3)f(z_2, z_1)$. For the other four methods, the conditional density is required to calculate in each hierarchy. For instance, in hierarchy $j = 5$, NN approximates the conditional density $f(z_5 \mid z_4, \ldots, z_1)$ by $f(z_5 \mid z_4, z_3)$; SUM by $f(z_5 \mid z_4 + z_3, z_2 + z_1)$; NNSUM by $f(z_5 \mid z_4, z_3 + z_2)$; and HLR by $f(z_5 \mid a_{14}z_4 + a_{13}z_3 + a_{12}z_2 + a_{11}z_1, a_{24}z_4 + a_{23}z_3 + a_{22}z_2 + a_{21}z_1)$, where a_{ij}'s are determined by the low rank approximation.

2.4. *Assessing Model Quality*

There are various ways to measure the performance of approximation methods, including the Kullback–Leibler divergence, the Godambe information matrix, and the Frobenius norm.

The Kullback–Leibler divergence computes the divergence of the approximated from the exact distributions. It measures the approximation accuracy through finding the expected difference in log-likelihood when assuming the approximated distribution is true. Stein (2014) used this measure for assessing the performance of the predictive process method (Banerjee et al. 2008), and pointed out that the Kullback–Leibler divergence provides a direct statistical interpretation for likelihood-based methods. In particular, if the Kullback–Leibler divergence is small, the predictive distributions yielded under the two laws are expected to be similar (Stein 1999, chap. 4). For the zero-mean Gaussian process, the Kullback–Leibler divergence has the closed form,

$$D_{\text{K-L}}(N_e \| N_a) = \frac{1}{2}\left\{\operatorname{tr}\left(\Sigma_a^{-1}\Sigma_e\right) + \log(|\Sigma_a|) - \log(|\Sigma_e|) - n\right\},$$

where N_e and N_a stand for the exact and the approximated distributions, respectively, Σ_e and Σ_a are the corresponding covariance matrices, and n is the dimension of the distribution.

The Godambe information matrix gives the asymptotic variances and covariances for the estimated parameters in the Gaussian process, as used by Kaufman, Schervish, and Nychka (2008) and Sun and Stein (2016). The Frobenius norm is another way to think about this problem. However, it is a matrix norm and does not penalize the positive definiteness of a covariance matrix (Stein 2014).

For our numerical and simulation studies in Section 3, we choose the Kullback–Leibler divergence and the Godambe Information matrix to assess the quality of the approximation. Because the results in terms of showing the different performances are similar, we only present the results of Kullback–Leibler divergence. It will be shown numerically that the Kullback–Leibler divergence of the hierarchical low rank approximation method is always the smallest when the rank r is small. This is because for sufficiently large j, $j > r$, the hierarchical low rank approximation method provides a better approximation in Equation (2) by including more neighbors than the nearest neighbors method. In the following Theorem 1, we have proved that the hierarchical low rank approximation method always gives the smallest error in approximating the covariance matrix in Equation (2) in terms of matrix norm for each sufficiently large j. While our numerical study agrees with this result, it also provides some insights on why the hierarchical low rank approximation method outperforms other methods with the smallest Kullback–Leibler divergence for the setting we have considered. Let V_{jj}^{N} be the $r \times r$ matrix defined by $A_{j,r}^{\mathrm{T}}\Sigma_{jj}A_{j,r}$ in Equation (2) using the nearest neighbors method and let $V_{jj}^{\mathrm{H}} = P_jL_jP_j^{\mathrm{T}} + \epsilon_j^2 I_{mr}$ be the $mr \times mr$ matrix for approximating $A_{j,mr}^{\mathrm{T}}\Sigma_{jj}A_{j,mr}$ in Equation (3) by the hierarchical low rank approximation method, where P_j consists of eigenvectors. The following theorem shows the result that the approximation to Σ_{jj} induced by V_{jj}^{H} is better than that induced by V_{jj}^{N} in terms of the Frobenius norm.

Theorem 1. Let $\lambda_1 \geq \cdots \geq \lambda_{mr} > 0$ be the eigenvalues of $A_{j,mr}^{\mathrm{T}}\Sigma_{jj}A_{j,mr}$. If ϵ_j^2 in Equation (3) satisfies $\epsilon_j^2 < (\lambda_r + \lambda_{mr})/2$, we have

$$\left\| A_{j,mr}V_{jj}^{\mathrm{H}}A_{j,mr}^{\mathrm{T}} - \Sigma_{jj}\right\|_F \leq \left\| A_{j,r}V_{jj}^{\mathrm{N}}A_{j,r}^{\mathrm{T}} - \Sigma_{jj}\right\|_F,$$

where $\|\cdot\|_F$ means the Frobenius norm.

The proof is shown in the Appendix. Similar results hold for the comparison between hierarchical low rank approximation method and the nearest neighboring sets method, or the nearest neighbors and nearest neighboring sets method.

2.5. Computational Complexity and Parallelization

For our hierarchical low rank approximation method, we need to execute a linear solve of dimension r, which requires $O(\min(j,r)^3)$ computation in Equation (4) for each hierarchy $j = 1, \ldots, n-1$ assuming that the direct method is employed. Then the total computational cost is $O(r^3 n)$ for likelihood approximation per value. When $r \ll n^{2/3}$, the computational cost is much smaller than $O(n^3)$, which is required by the Cholesky decomposition.

Another advantage of the proposed hierarchical low rank approximation method is that it does not need to store any $n \times n$ covariance matrix explicitly, which generally requires $O(n^2)$ memory. In each hierarchy, our proposed method only requires the covariance matrix of the chosen neighboring points to be stored, so that it uses a much smaller amount of memory than storing the entire $n \times n$ matrix.

In practice, the computation time can be reduced further by choosing z_{j+1} as a vector of an appropriate size because it leads to a smaller number of hierarchies that need to be evaluated while the increased computation in each hierarchy is comparable small. It is also worth noting that our approach can be parallelized easily because the computation of each hierarchy is independent of each other.

3. Numerical Study

3.1. Design Setup

In the numerical study in this section and the following simulation study in Section 4, we focus on irregularly spaced data with an unstructured covariance matrix (Sun and Stein 2016). The observations are generated at the locations $n^{-1/2}(r - 0.5 + X_{r\ell}, \ell - 0.5 + Y_{r\ell})$ for $r, \ell \in \{1, \ldots, n^{1/2}\}$, where n is the number of locations, and $X_{r\ell}$'s and $Y_{r\ell}$'s are independent and identically distributed, uniform on $(-0.4, 0.4)$. The advantage of this design is that it is irregular, and we can guarantee that no two locations are too close.

Here, we study the performances of different approximation methods proposed in Sections 2.2 and 2.3 in different settings. We consider a zero-mean Gaussian process model with Matérn covariance function possibly with a nugget,

$$C(h; \alpha, \beta, \nu, \tau^2) = \alpha\{(2\nu)^{1/2}h/\beta\}^{\nu}K_\nu\{(2\nu)^{1/2}h/\beta\}/\{\Gamma(\nu)2^{\nu-1}\} + \tau^2\mathbb{1}(h = 0), \quad (5)$$

where $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν, $\Gamma(\cdot)$ is the gamma function, $\mathbb{1}(\cdot)$ is the indicator function, $h \geq 0$ is the distance between two locations, $\alpha > 0$ is the sill parameter, $\beta > 0$ is the range parameter, $\nu > 0$ is the smoothness parameter, and τ^2 is the nugget effect.

For n irregularly spaced locations, the description of the five methods considered is shown in Table 1.

In Sections 3.2–3.4, we present the Kullback–Leibler divergence calculated from different settings for the five methods with α fixed at 1 and $n = 900$. In Section 3.5, we discuss the effect of sample size n and the rank r.

3.2. Dependence Level

In the Matérn model in Equation (5), the range parameter β controls the dependence of the process. In this section, we consider different β. Given $\nu = 0.5$, which corresponds to an exponential covariance function and $\tau^2 = 0.15$, the top-left and

Table 1. Description of the five methods used in the numerical study. IND, independent blocks method; NN, nearest neighbors method; SUM, nearest neighboring sets method; NNSUM, nearest neighbors and nearest neighboring sets method; HLR, the hierarchical low rank approximation method.

Method	Description
IND	Divide the locations into $\lceil n/r \rceil$ blocks, each of which contains at most r points. $\lceil n/r \rceil$ means the least integer that is greater or equal to n/r.
NN	A number of r nearest neighbors are selected to construct $A_{j,r}$.
SUM	A number of r nearest neighboring sets are selected and each set has two locations. Then a total number of $2r$ nearest neighbors are used to construct $A_{j,r}$.
NNSUM	A number of $\lceil r/2 \rceil$ nearest neighbors are first selected, then the following $2(r - \lceil r/2 \rceil)$ nearest neighbors are divided into $r - \lceil r/2 \rceil$ sets of size 2.
HLR	A number of $2r$ nearest neighbors are considered, where L_j is an $r \times r$ diagonal matrix with elements corresponding to the r leading eigenvalues. P consists of the r corresponding eigenvectors.

top-right panels of Figure 2 show the Kullback–Leibler divergence for $\beta = 0.1$, which means a weaker dependence, and $\beta = 0.5$, which indicates a stronger dependence, as the rank r increases from 2 to 8. We can see that the HLR approximation is always the best with the smallest Kullback–Leibler divergence, and SUM and NNSUM win against NN for $r = 2$ when $\beta = 0.1$, while when $\beta = 0.5$, the improvement of SUM and NNSUM exists up to $r = 6$. It implies that when a strong correlation is present, a small number of nearest neighbors is not adequate to provide a good approximation of the conditional density. It is also worth noting that the range of r/n in this study is from 0.22% to 0.89%. For very large n, and $r \ll n$, the improvement from HLR, SUM, or NNSUM approaches can be substantial.

3.3. Smoothness Level

In the Matérn covariance function, a larger ν indicates a smoother process. In this section, we fix $\beta = 0.1$ and $\tau^2 = 0.15$. We consider two smoothness levels with $\nu = 0.5$ and $\nu = 1$, which correspond to the exponential and Whittle covariance functions, respectively. The top-left and bottom-left panels of Figure 2 show the Kullback–Leibler divergence. Similarly, the HLR approach outperforms the other methods. For the smoother process when $\nu = 1$, the improvement of NNSUM and SUM over NN for small ranks decreases, and all the methods need a large r to achieve similar performances as $\nu = 0.5$.

3.4. Noise Level

The nugget effect can be viewed as measurement errors or the micro-structure in the underlying process. In this section, we consider different τ^2. Given $\beta = 0.1$ and $\nu = 0.5$, the bottom-right and top-left panels of Figure 2 show the Kullback–Leibler divergence for $\tau^2 = 0$ and $\tau^2 = 0.15$. In both cases, the HLR approach still provides the best approximation. We can see that SUM, NNSUM, and HLR give better approximations when the

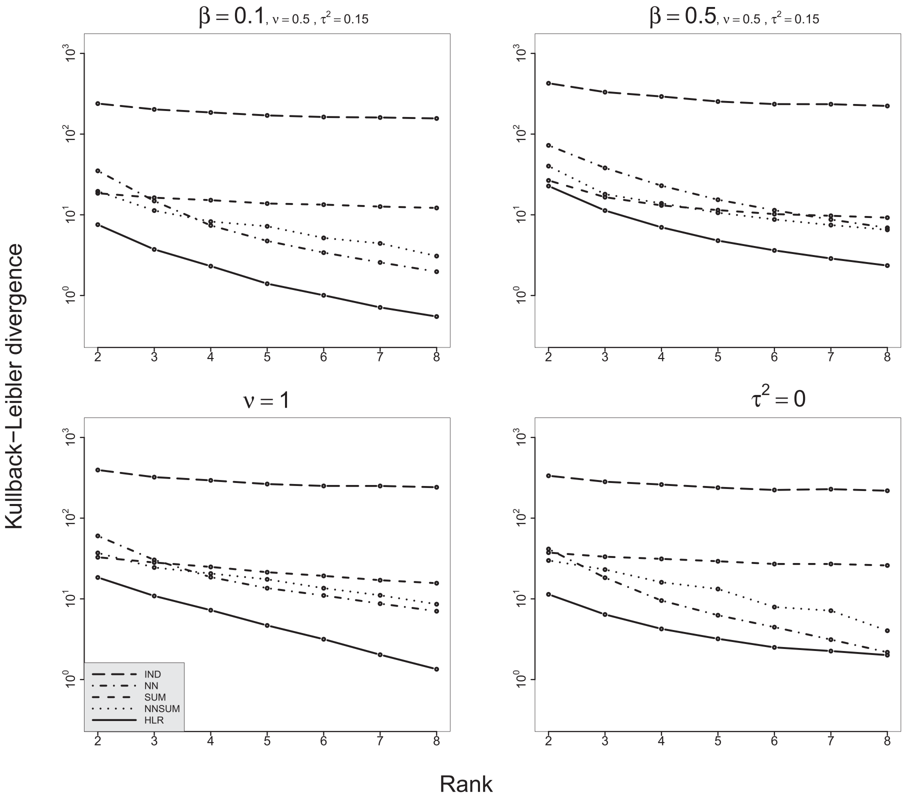

Figure 2. Four panels showing the Kullback–Leibler divergence against rank with 900 locations in IND (long dashes, — — —), NN (dot dashes, · - ·), SUM (dashes, - - -), NNSUM (dots, · · ·), and HLR (solid lines, —) methods. The corresponding parameters are indicated in the titles.

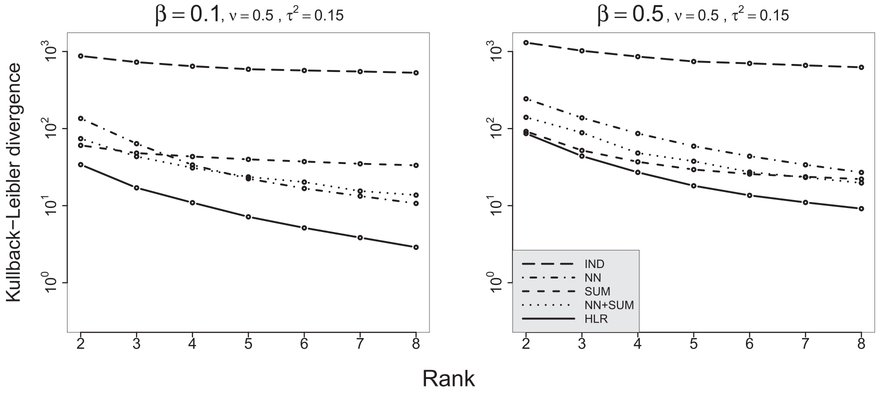

Figure 3. Two panels showing the Kullback–Leibler divergence against rank with 2500 locations in IND (long dashes, — — —), NN (dot dashes, · - ·), SUM (dashes, - - -), NNSUM (dots, · · ·), and HLR (solid lines, —) methods. The corresponding parameters are indicated in the titles.

process is noisy or with a larger τ^2. And if the rank r is limited to a small number, SUM or NNSUM can improve NN for noisy processes.

3.5. Sample Size and Rank

In this section, we explore the effect of sample size given the rank r or the ratio of r/n. Figure 3 shows the results for a similar design as in the first row of Figure 2 but with $n = 2500$. Comparing Figure 3 to the first row of Figure 2, we can see that for a given process, a larger number of locations does require larger ranks to achieve a similar approximation quality. When r is fixed, NN is often not adequate, especially for large n, and SUM, NNSUM, and HLR can improve the approximation by including more neighbors.

Although it is not realistic for a large dataset, we also investigate a situation where NN is adequate to provide a good approximation at rank r, and then compare the Kullback–Leibler divergence for NN at $r + 1$ and NNSUM with the same first r nearest neighbors and one additional set containing the next two nearest neighbors. We find that for $\alpha = 1$, $\beta = 0.5$, $\nu = 0.5$, $\tau^2 = 0$, and $n = 900$, NN with rank $r + 1 = 51$ gives a Kullback–Leibler divergence as 0.19 and NNSUM reduces Kullback–Leibler divergence by 1%.

4. Simulation Study

In Section 3, we calculated the Kullback–Leibler divergence at the true parameter values. In this section, we aim to test our estimation procedures with replications. To save computational time, only a relatively small number of locations is considered. We generate $n = 900$ observations with parameters $\alpha = 1$, $\beta = 0.25$, $\nu = 0.5$, and $\tau^2 = 0.15$. We run the optimization for α, β while fixing τ^2 and ν at the true value and obtain the estimates of α, β by maximizing the approximated likelihoods with $r = 2$. We repeat the estimates procedure 100 times and the boxplots of α and β are shown in Figure 4. Compared to the exact maximum likelihood estimates, the boxplots show that all the approximation methods perform reasonably well. Nevertheless, we can see that the estimates obtained by the hierarchical low rank approximation method have the smallest trimmed root mean squared errors (TRMSE) among all the approximation methods. The TRMSE for a given parameter θ is defined as follows by only keeping the 50 central estimates out of the 100 simulations:

$$\mathrm{TRMSE}(\theta) = \left(\sum_{i=26}^{75} \frac{(\theta - \hat{\theta}_{[i]})^2}{100} \right)^{0.5},$$

where θ is the true parameter value, and $\hat{\theta}_{[i]}$ denotes the ith largest estimate in the 100 simulations.

5. Application

In this section, we apply our method to modeling soil moisture, a key factor in evaluating the state of the hydrological process, including runoff generation and drought development. Soil moisture information is also valuable to a wide range of

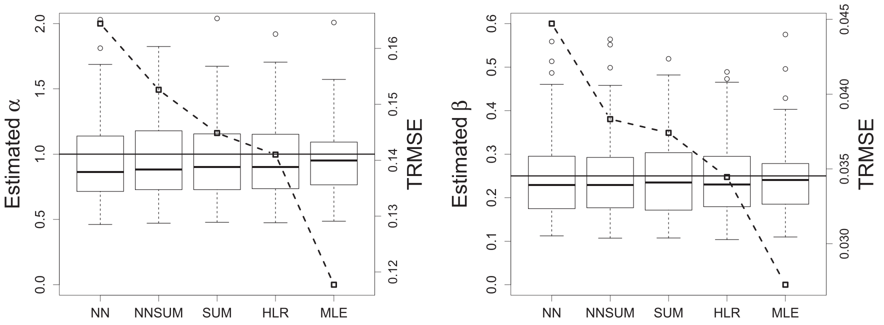

Figure 4. Two panels showing the boxplot of parameter estimates and the trimmed root mean squared error. The solid line is a reference for the true parameter value, and the dash line with squares (- - □ - -) is the corresponding trimmed root mean squared error of the 100 times of estimates in each method. Left: illustration for estimated α; Right: illustration for estimated β.

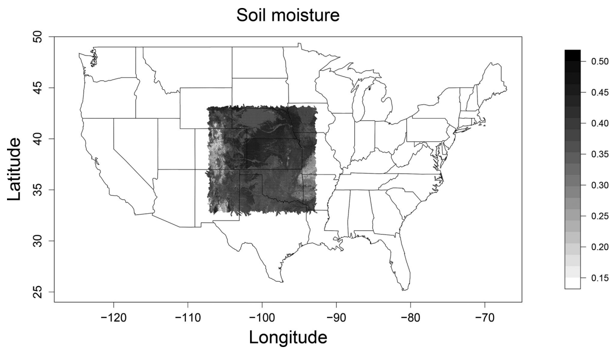

Figure 5. Soil moisture (unit: percentage) at the top layer of the Mississippi basin, U.S.A. on January 1, 2004.

applications, such as early warning of flood and drought, irrigation management, crop yield prediction, and weather pattern forecasting. Because soil moisture controls the energy exchange between the land and the atmosphere through evaporation and plant transpiration, it has been shown that better characterization of soil moisture in weather prediction models can lead to significant improvements on temperature and precipitation forecasting. As a result, the development of better statistical models for soil moisture plays an important role in understanding its spatial variability. Furthermore, many other environmental data, such as temperature, pressure, and humidity, often share similar data structure as the soil moisture that we have analyzed, and their spatial variability is also of great interest. Therefore, our analyses can be used as an illustration to model and make inferences on such high-resolution irregularly spaced spatial datasets.

5.1. *Dataset Description*

We consider high-resolution daily soil moisture data at the top layer of the Mississippi basin, U.S.A., on January 1, 2004 (Chaney, Metcalfe, and Wood 2016). The spatial resolution is of 0.0083° and the range is from 92.4749°W to 107.7166°W in longitude and from 32.3711°N to 43.4377°N in latitude. The grid consists of $1830 \times 1329 = 2,432,070$ locations with 2,153,888 observations and 278,182 missing values. The illustration of the data is shown in Figure 5.

We know that a 1° difference in latitude along any longitude line is equivalent to 111 km; however, the distance of 1° difference in longitude depends on the corresponding latitude. As the range of the latitude in this region is relatively small, for simplicity, we use the distance of 1° difference in longitude at the center location of the region to represent all others, which is 87.5 km, namely, in this region, 1° in latitude is 111 km and 1° in longitude is 87.5 km.

To understand the structure of the day's soil moisture, we fit a Gaussian process model with a Matérn covariance function. From all the locations, we randomly pick $n = 2{,}000{,}000$ points, which are irregularly spaced, to train our model. To assess the quality of our model, the fitted models can be used to predict part of the left out observations.

5.2. *Estimation and Prediction*

To use a Gaussian process model, we first fit a linear model to the longitude and latitude as the covariates to the soil moisture. After fitting, we find the negatively skewed residuals, hence we apply a logarithm transformation with some shift. The histogram of the transformed residual is shown in the left panel of Figure 6, which does not show strong departure from Gaussianity. To examine the isotropy of this process, we calculate the directional empirical variograms as illustrated in the right panel of Figure 6. We see the variograms on the circle with the same radius to the origin have similar values, suggesting that it is reasonable to assume an isotropic model.

Let $z(s)$ denote the transformed residual and the region D be the set of the selected locations, then the proposed Gaussian process model here is $\{z(s) : s \in D \subset \mathbb{R}^d\} \sim \mathrm{GP}(0, C(h; \theta))$. We choose three different covariance functions: the exponential, which has the smoothness parameter $\nu = 0.5$; the Whittle, which has $\nu = 1$; and the Matérn covariance function, which has an unknown ν. The formula is given in Equation (5). Given that the 2,000,000 observations follow $Z \sim N(0, \Sigma(\theta))$, $\Sigma(\theta)$ is the two million by two million variance-covariance matrix, obtained from the chosen covariance function. We use nearest neighbors and hierarchical low rank approximation methods with rank $r = 60$ to get the approximated likelihood and then obtain the parameter estimates. The results are shown in Table 2. The Matérn covariance model is more flexible by allowing to estimate ν. The estimated ν in the Matérn covariance model by both methods is smaller than 0.5, and the estimated β has the largest value. It suggests a rougher process with a larger dependence range compared to the estimated exponential covariance model. The last row of Table 2 shows the values of log-likelihood

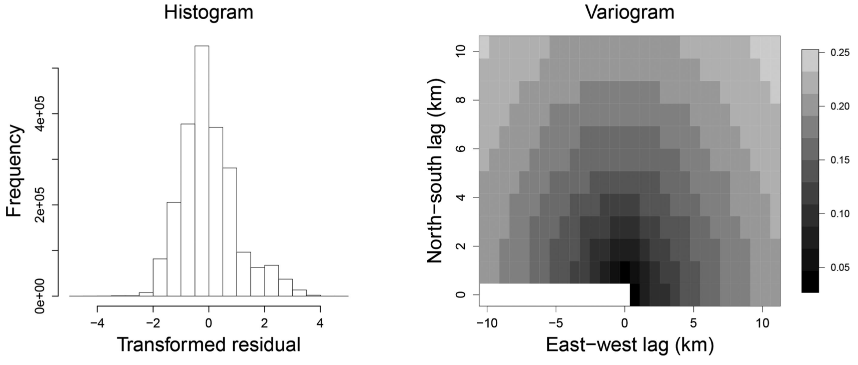

Figure 6. Left: the histogram of the transformed residuals; Right: the image plot of the empirical variogram at different distances and along different directions.

Table 2. Parameter estimation results.

	Nearest neighbors			Hierarchical low rank approximation		
	Exponential	Whittle	Matérn	Exponential	Whittle	Matérn
Estimated α	1.0073	0.9787	1.0597	1.0065	0.9789	1.0539
Estimated β (km)	21.6115	5.9316	222.6545	21.2944	5.8216	178.2051
Estimated τ^2	0.0107	0.0013	0.0000	0.0096	0.0012	0.0001
Estimated ν	0.5000	1.0000	0.2079	0.5000	1.0000	0.2214
log-likelihood/n	– 0.1042	– 0.1417	– 0.0852	– 0.0941	– 0.1308	– 0.0761

per observation. For each given covariance model, the likelihood with parameters estimated by the hierarchical low rank approximation method is always larger than that by the nearest neighbors method. Among different covariance models, the likelihood with Matérn covariance is the largest.

The size of the problem in this application is in the millions, a dataset that is far beyond the ability of classic analysis methods. However, nearest neighbors and hierarchical low rank approximation methods can evaluate the approximated likelihood at each iteration in the optimization procedure within 5 and 14 minutes, respectively. The fast computation makes it highly practical for applying the proposed methods to a large real-world spatial dataset problem. The experiment is performed with the Intel Xeon E5-2680 v3@2.50GHz processor. Next, we use the fitted Matérn model by the hierarchical low rank approximation method to predict soil moisture at 1000 among the left out locations by kriging, which is known to provide the best linear unbiased prediction as well as the prediction standard errors (Cressie 1993). However, the problem here is of size $n = 2,000,000$, hence kriging cannot be employed directly, because it involves a linear solve of size n (Furrer, Genton, and Nychka 2006). In fact, the proposed methods in this article can be adopted for approximating kriging equations as well. But for the purpose of validating the fitted model, we explore the exact computation method by treating the irregularly spaced data as observations on a finer regular grid with missing values. The resulting covariance matrix has a block Toeplitz Toeplitz block structure, which can be embedded in a block circulant circulant block matrix (Kozintsev 1999). Then kriging can be done by fast Fourier transformation. More details can be found in Chan and Ng (1996). The mean squared prediction errors (MSPE) over the 1000 validation locations are 4.23×10^{-5} and 5.07×10^{-5} for the hierarchical low rank approximation and the nearest neighbors method, respectively. The boxplots of the 1000 prediction errors are shown in Figure 7. We see that the hierarchical low rank approximation method leads to a better prediction accuracy and the MSPE is 17% smaller than that of the nearest neighbors method.

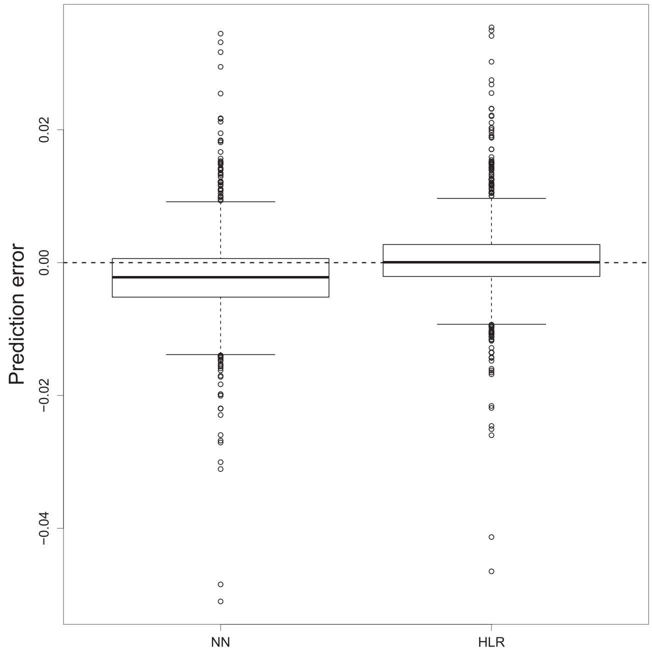

Figure 7. The boxplots of the prediction errors by the nearest neighbors and the hierarchical low rank approximation. The dash line is a constant zero for a reference.

6. Discussion

The computer code for the simulation and numerical studies is provided in the supplementary materials. The implementation in this article was done with a single-thread program, however as aforementioned in Section 2.5, computation in each hierarchy can be parallelized, which would reduce the computation time dramatically and make applications even more practical. The proposed method can be also extended to more complicated settings. For example, although the rank was fixed to the same in each hierarchy, it can be chosen flexibly in accordance with the number of "past" observations that are involved in the hierarchy, which, we believe, would give a better approximation. Moreover, for prediction problems, the proposed method can be further investigated to approximate kriging equations for large irregularly spaced spatial datasets.

Appendix: Proof of Theorem 1

Recall that the dimension of V_{jj}^{H} is $mr \times mr$, V_{jj}^{N} is $r \times r$, $A_{j,r}$ is $j \times r$, $A_{j,mr}$ is $j \times mr$, and Σ_{jj} is $j \times j$. Define B to be the $mr \times r$ matrix by keeping the mr selected rows from $A_{j,r}$, or $B = A_{j,mr}^{\mathrm{T}} A_{j,r}$. Let M denote $A_{j,mr}^{\mathrm{T}} \Sigma_{jj} A_{j,mr}$. The proof of Theorem 1 is as follows.

Proof of Theorem 1. With the equation,

$$\begin{aligned}
&\left\| A_{j,mr} V_{jj}^{\mathrm{H}} A_{j,mr}^{\mathrm{T}} - \Sigma_{jj} \right\|_F^2 - \left\| A_{j,r} V_{jj}^{\mathrm{N}} A_{j,r}^{\mathrm{T}} - \Sigma_{jj} \right\|_F^2 \\
&= \left\| A_{j,mr}^{\mathrm{T}} \left(A_{j,mr} V_{jj}^{\mathrm{H}} A_{j,mr}^{\mathrm{T}} - \Sigma_{jj} \right) A_{j,mr} \right\|_F^2 - \left\| A_{j,mr}^{\mathrm{T}} \right. \\
&\quad \left(A_{j,r} V_{jj}^{\mathrm{N}} A_{j,r}^{\mathrm{T}} - \Sigma_{jj} \right) A_{j,mr} \|_F^2 \\
&= \left\| V_{jj}^{\mathrm{H}} - A_{j,mr}^{\mathrm{T}} \Sigma_{jj} A_{j,mr} \right\|_F^2 - \left\| B V_{jj}^{\mathrm{N}} B^{\mathrm{T}} - A_{j,mr}^{\mathrm{T}} \Sigma_{jj} A_{j,mr} \right\|_F^2 \\
&= \left\| V_{jj}^{\mathrm{H}} - M \right\|_F^2 - \left\| B V_{jj}^{\mathrm{N}} B^{\mathrm{T}} - M \right\|_F^2,
\end{aligned}$$

it suffices to show,

$$\left\| V_{jj}^{\mathrm{H}} - M \right\|_F \leq \left\| B V_{jj}^{\mathrm{N}} B^{\mathrm{T}} - M \right\|_F .$$

Noting that $V_{jj}^{\mathrm{H}} = P_j L_j P_j^{\mathrm{T}} + \epsilon_j^2 I_{mr}$, we have $\|V_{jj}^{\mathrm{H}} - M\|_F = \|P_j L_j P_j^{\mathrm{T}} - (M - \epsilon_j^2 I_{mr})\|_F$. Since $\epsilon_j^2 < (\lambda_r + \lambda_{mr})/2$, we know that the eigenvalues of $M - \epsilon_j^2 I_{mr}$ satisfy $\lambda_1 - \epsilon_j^2 \geq \cdots \geq \lambda_r - \epsilon_j^2$ and $|\lambda_r - \epsilon_j^2| \geq \max_{k=r+1}^{mr}(|\lambda_k - \epsilon_j^2|)$. Thus, $|\lambda_1 - \epsilon_j^2| \geq \cdots \geq |\lambda_r - \epsilon_j^2| \geq \max_{k=r+1}^{mr}(|\lambda_k - \epsilon_j^2|)$. By the construction of P_j and L_j, and Eckart–Young–Mirsky theorem (Eckart and Young 1936; Mirsky 1960), we know

$$\|P_j L_j P_j^{\mathrm{T}} - (M - \epsilon_j^2 I_{mr})\|_F = \inf_{rank(X) \leq r} \|X - (M - \epsilon_j^2 I_{mr})\|_F .$$

Noting that the rank of $B V_{jj}^{\mathrm{N}} B^{\mathrm{T}}$ is r, we have $\|V_{jj}^{\mathrm{H}} - M\|_F = \|P_j L_j P_j^{\mathrm{T}} - (M - \epsilon_j^2 I_{mr})\|_F \leq \|B V_{jj}^{\mathrm{N}} B^{\mathrm{T}} - (M - \epsilon_j^2 I_{mr})\|_F = \|(B V_{jj}^{\mathrm{N}} B^{\mathrm{T}} - M) - \epsilon_j^2 I_{mr}\|_F$. It is easy to observe that the diagonal elements of $B V_{jj}^{\mathrm{N}} B^{\mathrm{T}} - M$ is nonpositive, thus $\|(B V_{jj}^{\mathrm{N}} B^{\mathrm{T}} - M) - \epsilon_j^2 I_{mr}\|_F \leq \|B V_{jj}^{\mathrm{N}} B^{\mathrm{T}} - M\|_F$. Then $\|V_{jj}^{\mathrm{H}} - M\|_F \leq \|B V_{jj}^{\mathrm{N}} B^{\mathrm{T}} - M\|_F$. This completes the proof. □

Supplementary Materials

Code folders: method folder contains R code for all the methods developed in this publication, and function folder includes all the necessary R functions for numerical studies.

Acknowledgments

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST). The authors thank the anonymous reviewers for their valuable comments.

References

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., and O'Neil, M. (2016), "Fast Direct Methods for Gaussian Processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38, 252–265. [111]

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 825–848. [110,111,112]

Castrillón-Candás, J. E., Genton, M. G., and Yokota, R. (2016), "Multi-Level Restricted Maximum Likelihood Covariance Estimation and Kriging for Large Non-Gridded Spatial Datasets," *Spatial Statistics*, 18, 105–124. [110]

Chan, R. H., and Ng, M. K. (1996), "Conjugate Gradient Methods for Toeplitz Systems," *SIAM Review*, 38, 427–482. [117]

Chaney, N. W., Metcalfe, P., and Wood, E. F. (2016), "HydroBlocks: A Field-Scale Resolving Land Surface Model for Application Over Continental Extents," *Hydrological Processes*, 30, 3543–3559. [116]

Cressie, N. (1993), *Statistics for Spatial Data* (2nd ed.), New York: Wiley. [110,117]

Cressie, N., and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Spatial Datasets," *Journal of the Royal Statistical Society*, Series B, 70, 209–226. [110,112]

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016), "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets," *Journal of the American Statistical Association*, 111, 800–812. [110]

Eckart, G., and Young, G. (1936), "The Approximation of One Matrix by Another of Lower Rank," *Psychometrika*, 1, 211–218. [118]

Furrer, R., Genton, M. G., and Nychka, D. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15, 502–523. [110,117]

Kaufman, C. G., Schervish, M. J., and Nychka, D. (2008), "Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets," *Journal of the American Statistical Association*, 103, 1545–1555. [110,113]

Kleiber, W., and Nychka, D. W. (2015), "Equivalent Kriging," *Spatial Statistics*, 12, 31–49. [110]

Kozintsev, B. (1999), "Computations With Gaussian Random Fields," Ph.D. dissertation, University of Maryland, College Park, MD. [117]

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R., and Klein, B. (2000), "Smoothing Spline ANOVA Models for Large Data Sets With Bernoulli Observations and the Randomized GACV," *Annals of Statistics*, 28, 1570–1600. [110]

Lindgren, F., Rue, H., and Lindström, J. (2011), "An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach," *Journal of the Royal Statistical Society*, Series B, 73, 423–498. [110]

Mirsky, L. (1960), "Symmetric Gauge Functions and Unitarily Invariant Norms," *Quarterly Journal of Mathematics*, 11, 50–59. [118]

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015), "A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 24, 579–599. [110]

Rue, H., and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Boca Raton, FL: Chapman and Hall/CRC. [110]

Rue, H., and Tjelmeland, H. (2002), "Fitting Gaussian Markov Random Fields to Gaussian Fields," *Scandinavian Journal of Statistics*, 29, 31–49. [110]

Sang, H., and Huang, J. Z. (2012), "A Full Scale Approximation of Covariance Functions for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 74, 111–132. [110]

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer. [110,113]

——— (2013), "Statistical Properties of Covariance Tapers," *Journal of Computational and Graphical Statistics*, 22, 866–885. [110,111]

——— (2014), "Limitations on Low Rank Approximations for Covariance Matrices of Spatial Data," *Spatial Statistics*, 8, 1–19. [110,112,113]

Stein, M. L., Chi, Z., and Welty, L. J. (2004), "Approximating Likelihoods for Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 66, 275–296. [110,111]

Sun, Y., Li, B., and Genton, M. G. (2012), "Geostatistics for Large Datasets," in *Advances And Challenges In Space-time Modelling of Natural Events* (Vol. 207), eds. J. M. Montero, E. Porcu, and M. Schlather, New York: Springer, pp. 55–77 [110]

Sun, Y., and Stein, M. L. (2016), "Statistically and Computationally Efficient Estimating Equations for Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 25, 187–208. [110,111,113]

Vecchia, A. V. (1988), "Estimation and Model Identification for Continuous Spatial Processes," *Journal of the Royal Statistical Society*, Series B, 50, 297–312. [110,111]

Ver Hoef, J. M., Cressie, N., and Barry, R. P. (2004), "Flexible Spatial Models for Kriging and Cokriging Using Moving Averages and the Fast Fourier Transform (FFT)," *Journal of Computational and Graphical Statistics*, 13, 265–282. [110]

Wikle, C. K., and Cressie, N. (1999), "A Dimension-Reduced Approach to Space-Time Kalman Filtering," *Biometrika*, 86, 815–829. [110]