A full-scale approximation of covariance functions for large spatial data sets

Huiyan Sang†

Department of Statistics, Texas A&M University, College Station, USA.

Jianhua Z. Huang

Department of Statistics, Texas A&M University, College Station, USA

Summary. Gaussian process models have been widely used in spatial statistics but face tremendous computational challenges for very large data sets. The model fitting and spatial prediction of such models typically require $O(n^3)$ operations for a data set of size n. Various approximations of the covariance functions have been introduced to reduce the computational cost. However, most existing approximations can not simultaneously capture both the large and small scale spatial dependence. A new approximation scheme is developed in this paper to provide a high quality approximation to the covariance function at both the large and small spatial scales. The new approximation is the summation of two parts: a reduced rank covariance and a compactly supported covariance obtained by tapering the covariance of the residual of the reduced rank approximation. While the former part mainly captures the large scale spatial variation, the latter part captures the small scale, local variation that is unexplained by the former part. By combining the reduced rank representation and sparse matrix techniques, our approach allows for efficient computation for maximum likelihood estimation, spatial prediction and Bayesian inference. We illustrate the new approach with simulated and real data sets.

Keywords: Covariance function; Gaussian processes; Geostatistics; Kriging; Large spatial dataset; Spatial processes

1. Introduction

Gaussian process models have been widely used in modeling spatial data (see, e.g., Diggle et al., 1998; Banerjee et al., 2004). However, large data sets pose tremendous computational challenges to the application of these models. In particular, spatial model fitting and spatial prediction (e.g., kriging) both involve inversion of an $n \times n$ covariance matrix for a data set of size n, which typically requires $O(n^3)$ operations and $O(n^2)$ memory, and is thus computationally intractable for very large n.

Various approximations of the spatial likelihood have been developed for efficient computation with large spatial data sets. Vecchia (1988) and Stein et al. (2004) used a product of conditional densities, where a careful choice of suitable conditional sets is required. The Gaussian Markov random field approximation by Rue and Tjelmeland (2002) and Rue and Held (2005) works best for gridded data and may have difficulty in prediction with massive

†*Address for correspondence:* Huiyan Sang, Department of Statistics, Texas A&M University, College Station, TX 77843, USA.
Email: huiyan@stat.tamu.edu

data sets. Fuentes (2007) worked in the spectral domain of spatial processes, a strategy suited mainly to stationary processes.

Two recently developed approaches have shown great appeal as general-purpose methodologies but each has its own drawbacks. The first approach is based on a reduced rank approximation of the underlying process. Methods following this path include kernel convolutions (see, e.g., Higdon, 2002), low rank splines or basis functions (e.g., Wikle and Cressie, 1999; Ver Hoef et al., 2004; Kammann and Wand, 2003; Cressie and Johannesson, 2008), and predictive process models (Banerjee et al., 2008; Finley et al., 2009). Reduced rank based methods have been proven successful in capturing large scale structure of spatial processes. However, they usually fail to accurately capture the local, small scale dependence structure (see, e.g., Stein, 2008; Finley et al., 2009).

The second approach seeks a sparse approximation of the covariance function and achieves computational efficiency through sparse matrix techniques. In particular, by setting to zero covariances of distant pairs of observations, covariance tapering has recently been introduced as a way for constructing sparse covariance matrix approximations and efficient algorithms have been developed for spatial prediction and parameter estimation (Furrer et al., 2006; Kaufman et al., 2008). The covariance tapering method works well in handling short range dependence but it may not be effective in accounting for spatial dependence with long range, because the tapered covariance function with a relatively small taper range fails to provide a good approximation to the original covariance function and may lead to bias in spatial prediction and parameter estimation.

We propose a full-scale approximation to the covariance function of a spatial process that facilitates efficient computation with very large spatial data sets, and in the meantime avoids the pitfalls of the two approaches mentioned above. The new approach uses a reduced rank process to capture the large scale spatial variation and a process with compactly supported covariance function to capture the small scale, local variation that is unexplained by the reduced rank process. The compactly supported covariance function is obtained by tapering the covariance function of the residual process from the reduced rank process approximation. By utilizing the reduced rank representation and sparse matrix techniques, the new approach significantly reduces the computational burden associated with very large data sets. The full-scale approximation works well with both the frequentist and the Bayesian approaches of spatial modeling. It can be conveniently applied to expedite computation for both the maximum likelihood and Bayesian inference of model parameters and to carry out spatial predictions through either the best linear unbiased prediction (i.e., kriging) or the Bayesian prediction.

The remainder of the paper is organized as follows. Section 2 reviews the Gaussian process models and two existing approximation methods for fast computation: the reduced rank and tapering methods. Section 3 presents the proposed new approximation method. Section 4 gives details of model fitting and spatial prediction using the new approximation. We then illustrate our method in Section 5 with a simulation study and a rainfall data analysis. Section 6 discusses some possible extensions and other applications.

2. Gaussian process models for spatial data sets

In this section, we present a summary of Gaussian process models for spatial data sets, and also review two existing approaches of approximating the covariance functions that allow rapid computation of the likelihood-based parameter estimation and spatial prediction. We

point out the drawbacks of these two approaches to motivate our new approach to be introduced in the next section. Our presentation of Gaussian process models is based on the standard treatment in Banerjee et al. (2004) and Schabenberger and Gotway (2005).

2.1. Gaussian process regression

Assume there is a response or dependent variable $Y(\mathbf{s})$ at a generic location $\mathbf{s} \in D \subset \mathbb{R}^2$ along with a $p \times 1$ vector of spatially referenced predictors $\mathbf{x}(\mathbf{s})$. A spatial regression model has the form

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{1}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients. The residual from the regression is decomposed into two independent parts: a spatial process, $w(\mathbf{s})$, modelling spatial association, and an independent process, $\epsilon(\mathbf{s})$, also known as the *nugget* effect, modelling measurement error (see, e.g., Chiles and Delfiner, 1999). The nugget effect $\epsilon(\mathbf{s})$ is often assumed to follow a normal distribution with variance τ^2 for every location $\mathbf{s}$. The spatial process $w(\mathbf{s})$ in (1) is often referred to as spatial random effects, capturing the effect of unmeasured or unobserved covariates with spatial pattern.

The most common specification for $w(\mathbf{s})$ is $w(\mathbf{s}) \sim GP(0, C(\cdot,\cdot))$, a zero-mean Gaussian process with a valid covariance function $C(\mathbf{s}, \mathbf{s}')$. It is often reasonable to assume a constant process variance and thus we specify $C(\mathbf{s}, \mathbf{s}') = \sigma^2\rho(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$, where $\rho(\cdot,\cdot;\boldsymbol{\theta})$ is a correlation function and $\boldsymbol{\theta}$ is a vector of correlation parameters.

Typically, the response variable $Y(\mathbf{s})$ is observed at a given collection of sites $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. The interest is in estimating the parameters $\Omega = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \tau^2)$ based on the observations and in predicting the responses at a set of new sites.

Both the maximum likelihood (or restricted maximum likelihood) and the Bayesian inference can be applied for parameter estimation. Denote $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$ and $\mathbb{X} = (\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_n))^T$. Let $\mathbf{C}_{n,n} = [C(\mathbf{s}_i, \mathbf{s}_j)]_{i=1:n, j=1:n}$ denote the covariance matrix of $w(\mathcal{S}) = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))$. The covariance matrix of $\mathbf{Y}$ is $\mathbf{C}_{n,n} + \tau^2\mathbf{I}$. The log likelihood function is

$$\begin{aligned}\ell_n(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \tau^2) = -\frac{n}{2}\log(2\pi) - \log\det\{\mathbf{C}_{n,n}(\boldsymbol{\theta}, \sigma^2) + \tau^2\mathbf{I}_n\} \\ - \frac{1}{2}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta})^T\{\mathbf{C}_{n,n}(\boldsymbol{\theta}, \sigma^2) + \tau^2\mathbf{I}_n\}^{-1}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}).\end{aligned} \tag{2}$$

Evaluation of the log likelihood involves calculation of the inverse and determinant of the $n \times n$ matrix $\mathbf{C}_{n,n}(\boldsymbol{\theta}, \sigma^2) + \tau^2\mathbf{I}_n$ which typically requires $O(n^3)$ operations. The computational burden makes it impractical to use the maximum likelihood for large data sets. The Bayesian approach assigns prior distributions to $\Omega = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \tau^2)$ and makes inference by sampling from the posterior distribution $p(\Omega|\mathbf{Y})$. Since the sampling procedure involves the evaluation of the likelihood function, the Bayesian approach faces the same kind of computational challenge as the maximum likelihood does. Actually, the computation is more demanding for the Bayesian method, because usually a large number of samples need to be drawn from the posterior distribution for a reliable inference.

Spatial prediction is customarily through the classical kriging method, i.e., the spatial best linear unbiased prediction (BLUP), or the Bayesian prediction. Conditional on the model parameters, the BLUP at a new location $\mathbf{s}_0$ is

$$\hat{Y}(\mathbf{s}_0) = \mathbf{x}^T(\mathbf{s}_0)\boldsymbol{\beta} + \mathbf{h}^T(\mathbf{s}_0)(\mathbf{C}_{n,n} + \tau^2\mathbf{I})^{-1}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}), \tag{3}$$

where $\mathbf{h}(\mathbf{s}_0) = (C(\mathbf{s}_0, \mathbf{s}_1), \ldots, C(\mathbf{s}_0, \mathbf{s}_n))^T$. The mean squared prediction error is

$$\text{MSE}[\hat{Y}(\mathbf{s}_0)] = C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{h}^T(\mathbf{s}_0)(\mathbf{C}_{n,n} + \tau^2\mathbf{I})^{-1}\mathbf{h}(\mathbf{s}_0) + \tau^2. \tag{4}$$

The computational bottleneck of applying the kriging equations (3) and (4) is the inversion of the $n \times n$ matrix $\mathbf{C}_{n,n} + \tau^2\mathbf{I}$ which typically has computational cost $O(n^3)$. On the other hand, the Bayesian prediction draws samples from the the predictive distribution $p(Y(\mathbf{s}_0)|\mathbf{Y}) = p(Y(\mathbf{s}_0)|\Omega, \mathbf{Y})p(\Omega|\mathbf{Y})$ at a new site $\mathbf{s}_0$ by composition. Again, the sampling from the posterior distribution $p(Y(\mathbf{s}_0)|\Omega, \mathbf{Y})$ involves the inversion of the $n \times n$ matrix $\mathbf{C}_{n,n} + \tau^2\mathbf{I}$, which is a computational burden for large data sets as shown in the previous paragraph.

2.2. The predictive process model: reduced rank approximation

Reduced rank methods approximate the spatial process $w(\mathbf{s})$ in (1) by a process $\tilde{w}(\mathbf{s})$ that lies in a fixed, finite-dimensional space. Since the resulting covariance matrix of the data has a fixed rank, great computational savings can be achieved for both likelihood inference and spatial prediction.

The reduced rank approximation can be motivated through the Karhunen-Loéve expansion of the spatial process (K-L expansion; Baker, 1977). Suppose the domain D of the process $w(\mathbf{s})$ is a compact set. Under certain conditions on the covariance function $C(\mathbf{s}, \mathbf{s}')$, the K-L expansion decomposes $w(\mathbf{s})$ into a countable orthogonal series,

$$w(\mathbf{s}) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\psi_i(\mathbf{s})Z_i, \tag{5}$$

where λ_i are the eigenvalues of the process, $\psi_i(\mathbf{s})$ are the corresponding orthonormal eigenfunctions, and $Z_i = \int_D w(\mathbf{s})\psi_i(\mathbf{s})ds/\sqrt{\lambda_i}$ are zero-mean unit-variance uncorrelated random variables. The eigenvalues λ_i in (5) are arranged in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots$. The eigenvalue-eigenfunction pairs are solutions to the integral equation

$$\int_D C(\mathbf{s}, \mathbf{t})\psi_i(\mathbf{t})d\mathbf{t} = \lambda_i\psi_i(\mathbf{s}) \tag{6}$$

with the constraint $\int_D \psi_i(\mathbf{s})\psi_j(\mathbf{s})d\mathbf{s} = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. The covariance function of the process has the representation $C(\mathbf{s}, \mathbf{s}') = \sum_{i=1}^{\infty} \lambda_i\psi_i(\mathbf{s})\psi_i(\mathbf{s}')$.

In general, the leading terms in the K-L expansion capture the main feature of the process and thus the remaining terms can be dropped from the expansion to yield a reasonable reduced rank approximation of the process. Keeping only the first m term in (5), we obtain a rank-m approximation

$$w(\mathbf{s}) \approx \bar{w}(\mathbf{s}) = \sum_{i=1}^{m} \sqrt{\lambda_i}\psi_i(\mathbf{s})Z_i. \tag{7}$$

The covariance function of $\bar{w}(\mathbf{s})$ is $C_l(\mathbf{s}, \mathbf{s}') = \boldsymbol{\psi}^T(\mathbf{s})\Lambda\boldsymbol{\psi}(\mathbf{s}')$, where $\boldsymbol{\psi}(\mathbf{s}) = (\psi_1(\mathbf{s}), \ldots, \psi_m(\mathbf{s}))^T$, and Λ is an $m \times m$ diagonal matrix with entries $\lambda_1, \ldots, \lambda_m$.

Application of the above reduced rank approximation relies on the ability to solve the integral equation, typically a hard task. Williams and Seeger (2001) proposed to solve the integral equation using the Nyström method. Consider a set of knots $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*\}$.

Let $\mathbf{C}^*$ denote the $m \times m$ matrix whose (k,l) entry is $C(\mathbf{s}_k^*, \mathbf{s}_l^*)$. By discretizing the integral, the Nyström method transforms (6) into

$$\frac{1}{m}\sum_{k=1}^{m} C(\mathbf{s}, \mathbf{s}_k^*)\psi_i(\mathbf{s}_k^*) \approx \lambda_i \psi_i(\mathbf{s}). \tag{8}$$

By plugging in $\mathbf{s}$ with points in $\mathcal{S}^*$ into (8), we obtain a matrix eigenequation $\mathbf{C}^*\mathbf{u}_i^{(m)} = \lambda_i^{(m)}\mathbf{u}_i^{(m)}$, the solution of which is linked to the solution of (8) through

$$\boldsymbol{\psi}_i(\mathcal{S}^*) \approx \sqrt{m}\mathbf{u}_i^{(m)}, \quad \lambda_i \approx \frac{\lambda_i^{(m)}}{m},$$

where $\boldsymbol{\psi}_i(\mathcal{S}^*) = (\psi_i(\mathbf{s}_1^*), \ldots, \psi_i(\mathbf{s}_m^*))^T$, and the normalization $|\mathbf{u}_i^{(m)}|^2 = (1/m)|\boldsymbol{\psi}_i(\mathcal{S}^*)|^2 = 1$ is used. By (8), the Nyström approximation of the ith eigenfunction is

$$\psi_i(\mathbf{s}) \approx \frac{\sqrt{m}}{\lambda_i^{(m)}}\sum_{k=1}^{m}\mathbf{C}(\mathbf{s}, \mathbf{s}_k^*)\mathbf{u}_i^{(m)}.$$

Plugging this into (7) we obtain the following rank-m approximation

$$\bar{w}(\mathbf{s}) = \mathbf{C}(\mathbf{s}, \mathcal{S}^*)\sum_{i=1}^{m}\sqrt{\frac{1}{\lambda_i^{(m)}}}\mathbf{u}_i^{(m)}Z_i = \mathbf{C}(\mathbf{s}, \mathcal{S}^*)\mathbf{C}^{*-1}\sum_{i=1}^{m}\sqrt{\lambda_i^{(m)}}\mathbf{u}_i^{(m)}Z_i, \tag{9}$$

where $\mathbf{C}(\mathbf{s}, \mathcal{S}^*) = (C(\mathbf{s}, \mathbf{s}_1^*), \ldots, C(\mathbf{s}, \mathbf{s}_m^*))^T$. The covariance function of the rank-m process $\bar{w}(\mathbf{s})$ is $\bar{C}(\mathbf{s}, \mathbf{s}') = \mathbf{C}(\mathbf{s}, \mathcal{S}^*)^T\mathbf{C}^{*-1}\mathbf{C}(\mathbf{s}', \mathcal{S}^*)$, which is a finite rank approximation to the covariance function $C(\mathbf{s}, \mathbf{s}')$ of the original process $w(\mathbf{s})$. Applying this approximation, the matrix inversion in (2)–(4) can be computed efficiently using the Sherman–Woodbury–Morrison formula: $(\mathbf{A} + \mathbf{UBV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{B}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$.

The reduced rank approximation given above is based on truncating the K-L expansion of the process and the Nyström approximation of the eigensystem. Banerjee et al. (2008) proposed to construct an reduced rank approximation using spatial interpolation which, interestingly, yields the same approximation of the covariance function. Specifically, let $\mathbf{w}^* = [w(\mathbf{s}_i^*)]_{i=1}^m$ denote the realization of $w(\mathbf{s})$ at the m knots in $\mathcal{S}^*$. The BLUP of $w(\mathbf{s})$ at any fixed site $\mathbf{s}$ based on $\mathbf{w}^*$ is $\tilde{w}(\mathbf{s}) = \mathbf{C}(\mathbf{s}, \mathcal{S}^*)^T\mathbf{C}^{*-1}\mathbf{w}^*$. According to the classical theory of kriging, $\tilde{w}(\mathbf{s})$ minimizes the mean squared prediction error $E\{[w(\mathbf{s}) - f(\mathbf{w}^*)]^2\}$ over all linear functions $f(\mathbf{w}^*)$, and over all square integrable functions if the process is Gaussian. Because of the interpretation as the best prediction, $\tilde{w}(\mathbf{s})$ is called the predictive process. Banerjee et al. (2008) has demonstrated the utility of applying the predictive process approximation to achieve computational efficiency in Bayesian hierarchical modeling of large spatial data sets. Since the covariance matrix of $\mathbf{w}^*$ is $\mathbf{C}^*$, the process $\tilde{w}(\mathbf{s})$ has the same covariance function as the rank-m process $\bar{w}(\mathbf{s})$ defined in (9). Thus the approaches by Williams and Seeger (2001) and Banerjee et al. (2008) are equivalent.

Although methods based on the reduced rank approximation have proven successful in capturing large-scale variation of spatial processes, they share one common disadvantage: inaccuracy in representing local/small scale dependence (Stein, 2008; Finley et al., 2009). Top panel of Fig. 1 shows a typical example that the predictive process approximation is poor at short distances. For spatial processes with relatively fine scale spatial dependence,

the reduced rank approximation generally requires a relatively high rank m in order to preserve more complete information about the fine scale spatial pattern, and hence loses the computational advantage. Banerjee et al. (2008) pointed out that the performance of the predictive process approximation depends on the size of the spatial dependence range relative to the spacing of the knots. The quality of the predictive process approximation usually gets worse when the spatial dependence range gets shorter. In this regard, the predictive process with a limited number of knots will not be able to make reliable inference of dependence for pairs of sites that are very close to each other (relative to the spacing of the knots). Some numerical examples will be presented in Section 3 when the predictive process is compared with our new approximation scheme.

The predictive process model requires the selection of knot locations. Banerjee et al. (2010) briefly discussed several possible strategies for knot selection. For fairly evenly distributed data locations, they suggested to select knots on a uniform grid overlaid on the domain. For highly irregularly distributed locations, they suggested either to use the popular clustering algorithms such as k-means or the more robust median-based partitioning around medoids algorithms (e.g., Kaufman and Rousseeuw, 1990). One may also choose the knots following some formal design-based approaches based upon minimization of a spatially averaged predictive variance criterion (e.g., Diggle and Lophaven, 2006; Finley et al., 2009).

2.3. Sparse matrix approximation and covariance tapering

Another approximation approach is to approximate the data covariance matrix by a sparse matrix and then employ the sparse matrix algorithms (e.g., Pissanetzky, 2007) to achieve computational efficiency. If one believes that distant pairs of observations are uncorrelated, then one can use a compactly supported covariance function to model the spatial dependence (Gneiting, 2002). Since the commonly used covariance functions in spatial statistics are not compactly supported, Furrer et al. (2006) and Kaufman et al. (2008) proposed to approximate the model covariance function by a compactly supported one using tapering, or multiplying the data covariance matrix element-wise by a sparse correlation matrix.

Let $C(h)$, where $h = \|\mathbf{x} - \mathbf{x}^*\|$, denote the original covariance function for a stationary random field. Consider a tapering function $K_{taper}(h;\gamma)$, an isotropic correlation function which is identically zero whenever $h \geq \gamma$. The tapered covariance function is defined as

$$C_{taper}(h;\gamma) = C(h)K_{taper}(h;\gamma), \qquad h > 0$$

According to the Schur product theorem (Horn and Johnson, 1985, Section 7.5), the tapered covariance function is positive semi-definite and thus a valid covariance function. Note that the tapered covariance is exactly zero for data at any two locations whose distance is larger than γ. Thus the tapering range parameter γ can be viewed as the effective range for the spatial phenomenon being studied. By assigning γ a small value, we obtain a sparse covariance matrix and can use efficient sparse matrix algorithms for likelihood inference and spatial prediction. Many compactly supported correlation functions constructed in the literature can serve as a taper function (see, e.g., Wendland, 1995, 1998; Gneiting, 2002). Examples include the spherical covariance function defined as

$$K_{\text{spherical}}(h;\gamma) = (1 - \frac{h}{\gamma})_+^2(1 + \frac{h}{2\gamma}), \qquad h > 0, \tag{10}$$

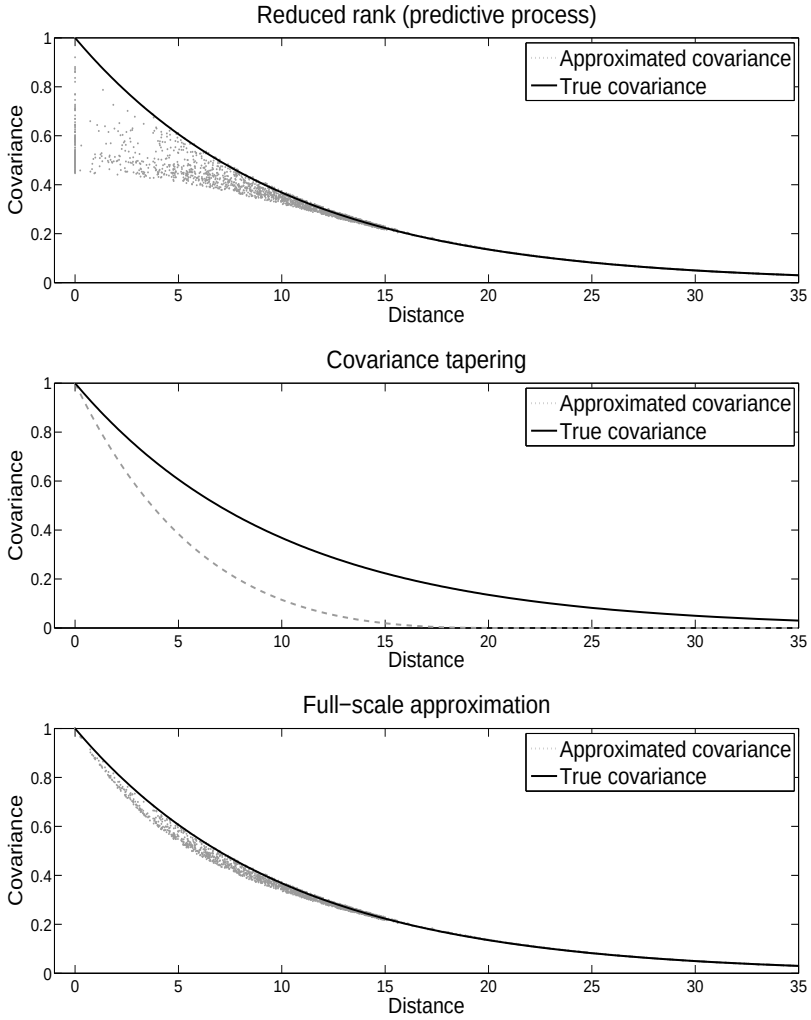

Fig. 1. The exponential covariance function $C(h) = \exp(-3h/50)$ and its approximations. The solid lines represent the true covariance function. The top panel plots the covariance matrix at 500 random locations from $[0, 100] \times [0, 100]$ using the predictive process approximation with $m = 100$ evenly placed knots. The middle panel displays the tapered covariance function using the spherical correlation taper with taper range $\gamma = 20$. The bottom panel plots the covariance matrix at the same 500 locations using the proposed full-scale approximation with $m = 100$ evenly placed knots and taper range $\gamma = 20$.

and the Wendland family with members such as

$$K_{\text{wendland},1}(h;\gamma) = (1 - \frac{h}{\gamma})_+^4(1 + 4\frac{h}{\gamma}), \qquad h > 0, \tag{11}$$

and

$$K_{\text{wendland},2}(h;\gamma) = (1-\frac{h}{\gamma})_+^6(1+6\frac{h}{\gamma}+\frac{35h^2}{2\gamma^2}), \qquad h>0. \tag{12}$$

See Furrer et al. (2006) for some suggestions on choosing the tapering function for the covariance tapering.

The covariance tapering does a good job in capturing the small scale spatial dependence, but may not be effective to account for the large scale dependence. The middle panel of Fig. 1 shows that the quality of approximation at long range can be rather poor. Because of the way how tapering works, the tapered covariance function with a relatively small tapering range fails to provide a good approximation to the original covariance function at long range, and hence may lead to serious bias in parameter estimation and inaccuracy in prediction. Using a larger tapering range may improve the quality of approximation but sacrifice the computational advantage of tapering. To adjust the bias in parameter estimation, Kaufman et al. (2008) proposed an alternative tapering method, referred to as the two-taper approximation, to approximate the log likelihood function by tapering both the model and sample covariance matrices. Although the two-taper estimates do not have large bias, Kaufman et al. (2008) indicated that they are not suitable for being plugged into the kriging procedure in prediction applications.

3. The full-scale covariance approximation

Our new approach combines the ideas of the reduced-rank process approximation and the sparse covariance approximation and has the advantages of both approaches while overcomes their individual shortcomings. We first decompose the spatial Gaussian process into two parts: a reduced rank process to characterize the large scale dependence and a residual process to capture the small scale spatial dependence that is unexplained by the reduced rank process. We then obtain sparse covariance approximation of the residual process using covariance tapering. Since the residual process mainly captures the small scale dependence and the tapering has little impact on such dependence other than introducing sparsity, the error of the new approximation is expected to be small. We refer to our new approach as the full-scale approximation because of its capability of providing high quality approximations at both the small and large spatial scales.

Specifically, for the spatial process $w(\mathbf{s})$ in (1), consider the decomposition

$$w(\mathbf{s}) = w_l(\mathbf{s}) + w_s(\mathbf{s}), \tag{13}$$

where $w_l(\mathbf{s})$ is a reduced rank approximation of $w(\mathbf{s})$ and $w_s(\mathbf{s}) = w(\mathbf{s}) - w_l(\mathbf{s})$ is the residual of the approximation. We specialize $w_l(\mathbf{s})$ to be the predictive process introduced in Section 2.2. The subscripts of $w_l(\mathbf{s})$ and $w_s(\mathbf{s})$ indicate respectively that they primarily capture the long range and short range dependence. For a fixed set $\mathcal{S}^*$ of m knots and the corresponding vector $\mathbf{w}^*$ of process realizations, the predictive process can be expressed as

$$w_l(\mathbf{s}) = \mathbf{C}^T(\mathbf{s},\mathcal{S}^*)\mathbf{C}^{*-1}\mathbf{w}^*, \tag{14}$$

whose finite-rank covariance function is

$$C_l(\mathbf{s},\mathbf{s}') = Cov\{w_l(\mathbf{s}), w_l(\mathbf{s}')\} = \mathbf{C}^T(\mathbf{s},\mathcal{S}^*)\mathbf{C}^{*-1}\mathbf{C}(\mathbf{s}',\mathcal{S}^*). \tag{15}$$

The predictive process is less variable than the original process in the sense that the marginal variance of the predictive process at a fixed location is equal to (when the location coincides

with a knot) or smaller than that of original covariance (Finley et al., 2009). It can capture reasonably well the long range but not the short range dependence.

The reason that the predictive process approach can not capture the short range dependence is because it discards entirely the residual process $w_s(\mathbf{s})$. Our novelty here is a more careful treatment of the covariance function that can both preserve most information present in the residual process and also achieve computational efficiency. The covariance function of the residual process is

$$C(\mathbf{s},\mathbf{s}') - \mathbf{C}^T(\mathbf{s},\mathcal{S}^*)\mathbf{C}^{*-1}\mathbf{C}(\mathbf{s}',\mathcal{S}^*).$$

We propose a sparse matrix approximation of this function using tapering. The tapering function, denoted as $K_{taper}(\mathbf{s},\mathbf{s}';\gamma)$, is chosen to be a compactly supported correlation function that is identically zero whenever $|\mathbf{s}-\mathbf{s}'| \geq \gamma$ for a positive taper range γ (Gneiting, 2002). We approximate the covariance function of the residual process $w_s(\mathbf{s})$ by the following tapered function

$$C_s(\mathbf{s},\mathbf{s}') = \{C(\mathbf{s},\mathbf{s}') - \mathbf{C}^T(\mathbf{s},\mathcal{S}^*)\mathbf{C}^{*-1}\mathbf{C}(\mathbf{s},\mathcal{S}^*)\}K_{taper}(\mathbf{s},\mathbf{s}';\gamma), \tag{16}$$

which is a valid covariance function with compact support. By assigning the taper range parameter γ a small value, one obtains sparse covariance matrices, which can be manipulated using efficient sparse matrix algorithms.

Putting things together, our approximation of the original covariance function has the form

$$C^\dagger(\mathbf{s},\mathbf{s}') = C_l(\mathbf{s},\mathbf{s}') + C_s(\mathbf{s},\mathbf{s}'), \tag{17}$$

where $C_l(\mathbf{s},\mathbf{s}')$ and $C_s(\mathbf{s},\mathbf{s}')$ are given in (15) and (16) respectively. Our formulation includes existing approaches as special cases. If the tapered part is void, we get the predictive process of Banerjee et al. (2008); if the set $\mathcal{S}^*$ of locations is empty, then the predictive process part is void and we get the tapered version of the original covariance function as used in Furrer et al. (2006) and Kaufman et al. (2008).

The approximated covariance function given in (17) is indeed a valid covariance function provided both $C(\cdot,\cdot)$ and $K_{taper}(\cdot,\cdot)$ are valid covariance functions. To give a more precise statement, we introduce some notations. Let $k(\cdot,\cdot)$ be a function on $D \times D$. Given n points $\mathbf{s}_1,\ldots,\mathbf{s}_n$ in D, the $n \times n$ matrix K with elements $K_{ij} = k(\mathbf{s}_i,\mathbf{s}_j)$ is called the Gram matrix of k with respect to $\mathbf{s}_1,\ldots,\mathbf{s}_n$; the Gram matrix is called positive semi-definite, if $\mathbf{c}^T K\mathbf{c} \geq 0$ for all vectors $\mathbf{c} \in \mathbb{R}^n$, and positive definite, if in addition, $\mathbf{c}^T K\mathbf{c} = 0$ only when $\mathbf{c}$ is a vector of 0's. The function $k(\cdot,\cdot)$ is positive (semi-)definite if the corresponding Gram matrix is positive (semi-)definite for all possible choices of $\mathbf{s}_1,\ldots,\mathbf{s}_n$.

Proposition 1.
(i) If $K_{taper}(\cdot,\cdot)$ is positive semi-definite, then $C^\dagger(\cdot,\cdot)$ is positive semi-definite if and only if $C(\cdot,\cdot)$ is positive semi-definite.
(ii) If $K_{taper}(\cdot,\cdot)$ is positive definite, then $C^\dagger(\cdot,\cdot)$ is positive definite if and only if $C(\cdot,\cdot)$ is positive definite.

The proof is given in the Appendix.

A positive semi-definite function of two arguments is a valid covariance function. Part (i) of Proposition 1 suggests that the proposed full-scale approximation provides a valid covariance function. Part (ii) of the proposition indicates that the full-scale approximation is not of reduced rank, which is in contrast to the fact that the predictive process approximation is reduced rank. If $C(\cdot,\cdot)$ is positive definite, then the Gram matrix produced by the

full-scale approximation is always of full rank provided that the taper function is positive definite, but the Gram matrix by the predictive process approximation is singular.

To compare the approximation property of the new approximation approach and the reduced rank and the covariance tapering approaches, we employ the Matérn family of stationary correlation functions (Stein, 1999):

$$\rho(\mathbf{s},\mathbf{s}';\nu,\phi)=\frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\nu^{1/2}\|\mathbf{s}-\mathbf{s}'\|}{\phi}\right)^{\nu}J_{\nu}\left(\frac{2\nu^{1/2}\|\mathbf{s}-\mathbf{s}'\|}{\phi}\right),\qquad \nu>0,\phi>0, \tag{18}$$

where $\Gamma(\cdot)$ is the gamma function, J_ν is the modified Bessel function of the second kind with order ν, and $\|\cdot\|$ denotes the Euclidean distance. The parameter ν is called the smoothness parameter, controlling the degree of smoothness (i.e., the degree of differentiability of the sample paths) of the random field, and ϕ is a spatial range parameter. Matérn family is perhaps the most widely used covariance function in spatial statistics because it flexibly encompasses several class of valid covariance functions, including the exponential ($\nu=.5$) and the Gaussian ($\nu\to\infty$). Furrer et al. (2006) gave some suggestions for choosing different tapering functions for the members of the Matérn family according to their smoothness.

A special case of the Matérn family has been used in generating Fig. 1 when we illustrated the approximation properties of the reduced rank and tapering approximations. The true covariance function presented in Fig. 1 corresponds to model (1) where $\boldsymbol{\beta}=\mathbf{0}$, $\sigma^2=1$, $\tau^2=0$, and the correlation function of the spatial random effects belongs to the Matérn family with $\nu=.5$ and $\phi=50/3$. We considered the square domain $[0,100]\times[0,100]$ in $\mathbb{R}^2$ and depicted the approximated covariance functions by evaluating 3000 of the covariance values from the 125,250 possible pairs generated from 500 random locations from the square. We used a 10×10 equally spaced grid as the set of knots for the predictive process approximation and chose the spherical taper with $\gamma=20$ for the covariance tapering. The same set of knots and choice of taper range were used for the full-scale approximation. The bottom panel of Fig. 1 shows the result of using the full-scale approximation. It is clear that the full-scale approximation offers good approximations to the original covariance function at both short distances and long distances, and does not have the serious biases appeared in either the reduced rank or the covariance tapering approximations as shown on the other two panels of Fig. 1.

Next, we compare the approximation properties of different approaches of covariance approximation under various choices of the smoothness and spatial range parameters of the Matérn family. To make our presentation concise, we assess covariance approximation by means of the Kullback-Leibler (K-L) divergence between distributions and the Frobenius distance between covariance matrices. In specific, consider two multivariate normal distributions $\mathcal{L}_1=\mathrm{MVN}_1(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ and $\mathcal{L}_2=\mathrm{MVN}_2(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ on $\mathbb{R}^n$. The K-L divergence from $\mathcal{L}_1$ to $\mathcal{L}_2$ is defined as

$$\mathrm{KL}(\mathcal{L}_1,\mathcal{L}_2)=\frac{1}{2}\{\log\det(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)+\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)+(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)-n\}.$$

The Frobenius distance between the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ is defined as $F(\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2)=\{\sum_{i,j}(\boldsymbol{\Sigma}_{1,ij}-\boldsymbol{\Sigma}_{2,ij})^2\}^{1/2}$.

We considered Model (1) with $\boldsymbol{\beta}=\mathbf{0}$, $\sigma^2=1$, $\tau^2=0.01$, and the Matérn correlation function for the spatial random effects for a set of varying spatial range parameters and smoothness parameters. We adopted the same settings of the knot intensity, the taper range, and the sampling locations as in Fig. 1. Fig. 2 shows the K-L divergence from the

approximated distribution to the true distribution under each approach, and the relative Frobenius distance between the approximated and the true covariance matrices, defined as $F(\mathbf{\Sigma}_{\text{approx}}, \mathbf{\Sigma}_{\text{true}})/F(\mathbf{\Sigma}_{\text{true}})$. It is not surprising to see that, when the smoothness parameter is fixed and the range parameter increases, the K-L divergence and the relative Frobenius distance increase under the covariance tapering approximation, while the same measures decrease under the predictive process approximation. Given the same spatial range parameter and the smoothness parameter, it is evident that the full-scale approximation produces substantially smaller K-L divergence and relative Frobenius distance than the other two existing approaches, indicating that the new approximation is more accurate.

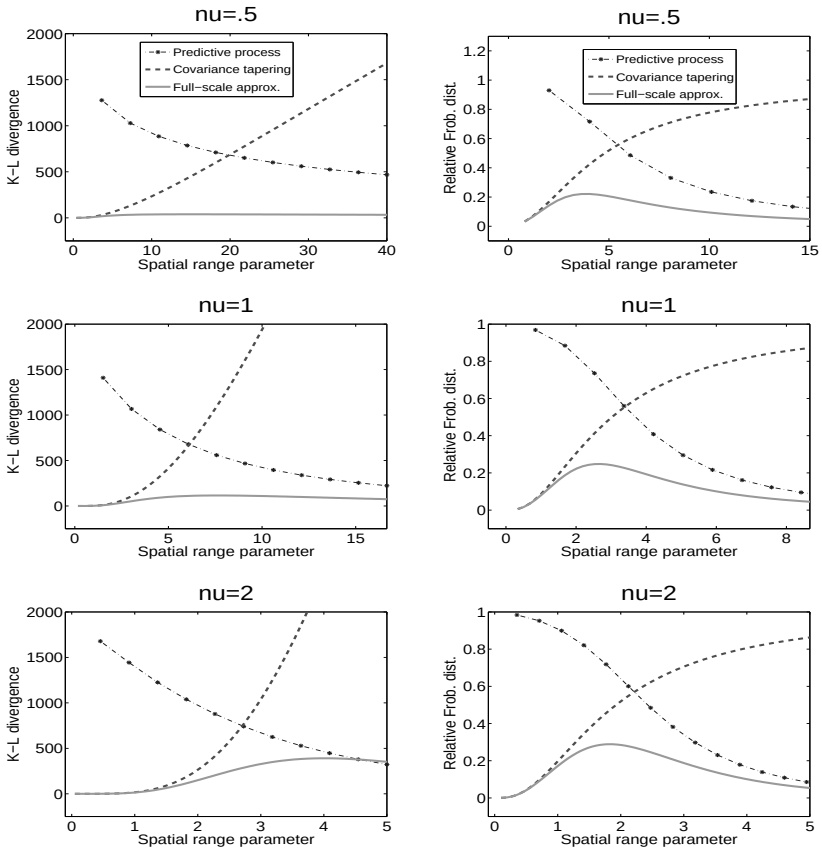

Fig. 2. The K-L distance and the relative Frobenius distance from the approximated covariance matrix to the true covariance matrix for the Matérn family with different smoothness parameters and spatial range parameters. The relative Frobenius distance is defined as $F(\mathbf{\Sigma}_{\text{approx}}, \mathbf{\Sigma}_{\text{true}})/F(\mathbf{\Sigma}_{\text{true}})$. We adopt the same settings of the knot intensity, the taper range, and the sampling locations as in Fig. 1. The spherical function in (10) is used for $\nu = .5$; the Wendland function in (11) is used for $\nu = 1$, and the Wendland function in (12) is used for $\nu = 2$.

Next, we examine how the approximation quality changes as we vary the knot intensity and tapering range for the full-scale approximation. Fix the smoothness parameter at $\nu = .5$ and the range parameter at $\phi = 10$. Fig. 3 displays the K-L divergence and the relative Frobenius distance for varying knot intensities and taper range. We considered four knot intensities, $m = 64$, $m = 100$, $m = 225$, $m = 400$, and a dense grid of taper range. We observed that higher knot intensity and larger taper range are associated with better approximation quality. More precisely, when the knot intensity is fixed and the taper range increases, both the K-L divergence and the relative Frobenius distance decrease; when the taper range is fixed and the knot intensity increases, the K-L divergence and the relative Frobenius distance decrease. Usually various combinations of knot intensity and taper range can be used to achieve similar approximation quality. For example, for the K-L distance to be 50, one may choose a high knot intensity ($m = 400$) combined with a low taper range ($\gamma \approx 2$), a median knot intensity ($m = 225$) with a median taper range ($\gamma \approx 4$) or a low knot intensity ($m = 12$) with a large taper range ($\gamma \approx 15$).

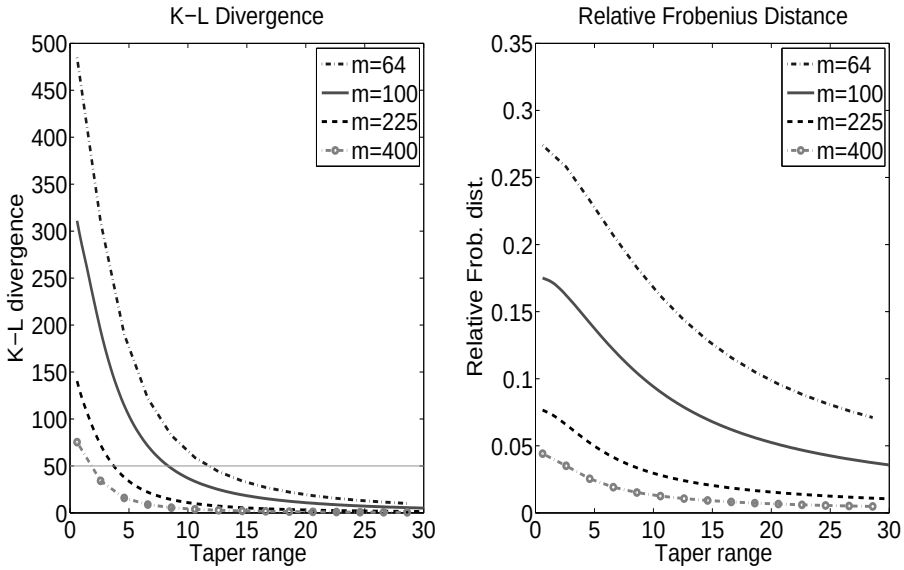

Fig. 3. The K-L distance and the relative Frobenius distance from the approximated covariance matrix to the true covariance matrix for the Matérn family with four different knot intensities ($m = 64, 100, 225, 400$) and a dense grid of taper range. The true covariance is an exponential correlation function with $\phi = 10$.

Remark. It is easy to see that (17) is the covariance function of the following process

$$w^{\dagger}(\mathbf{s}) = \mathbf{C}^{T}(\mathbf{s}, \mathcal{S}^{*})\mathbf{C}^{*-1}\mathbf{w}^{*} + \{w(\mathbf{s}) - \mathbf{C}^{T}(\mathbf{s}, \mathcal{S}^{*})\mathbf{C}^{*-1}\mathbf{w}^{*}\}\xi(\mathbf{s}), \tag{19}$$

where $\xi(\mathbf{s})$ is a spatial random field independent of $w(\mathbf{s})$ and with $K_{taper}(\mathbf{s}, \mathbf{s}'; \gamma)$ as the covariance function. However, the process $w^{\dagger}(\mathbf{s})$ is not Gaussian and so cannot be used for likelihood approximation of the spatial random effects $w(\mathbf{s})$ of the original model. Likelihood approximation using (17) based on the original model will be discussed in detail in the next section.

4. Parameter estimation and spatial prediction

We discuss in this section how the covariance approximation proposed in the previous section can be used to develop efficient computational algorithms for parameter estimation and spatial prediction. Both the maximum likelihood and Bayesian inference of model parameters are considered (Cressie, 1993; Banerjee et al., 2004).

4.1. Maximum likelihood

To derive an approximate likelihood that facilitates efficient computation, we replace the covariance function of the spatial process $w(\mathbf{s})$ by the approximation given in (17). Recall that $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_n^*\}$ is the set of knots used for the reduced rank approximation (15). Denote $\mathbf{C}_{n,m} = [C(\mathbf{s}_i, \mathbf{s}_j^*)]_{i=1:n,j=1:m}$ and $\mathbf{C}_{m,m}^* = [C(\mathbf{s}_i^*, \mathbf{s}_j^*)]_{i=1:m,j=1:m}$. According to (13), $w(\mathbf{s})$ is decomposed as the sum of a reduced rank process $w_l(\mathbf{s})$ and a residual process $w_s(\mathbf{s})$. It follows from (15) and (16) that the covariance matrix of $\{w_l(\mathbf{s}), \mathbf{s} \in \mathcal{S}\}$ is given by

$$\mathbf{C}_l = \mathbf{C}_{n,m}\mathbf{C}_{m,m}^{*-1}\mathbf{C}_{n,m}^T \tag{20}$$

and the tapered covariance matrix of $\{w_s(\mathbf{s}), \mathbf{s} \in \mathcal{S}\}$ is given by

$$\mathbf{C}_s = (\mathbf{C}_{n,n} - \mathbf{C}_{n,m}\mathbf{C}_{m,m}^{*-1}\mathbf{C}_{n,m}^T) \circ \mathbf{T}(\gamma), \tag{21}$$

where $\mathbf{T}(\gamma) = [K_{\text{taper}}(\mathbf{s}_i, \mathbf{s}_j; \gamma)]_{i=1:n,j=1:n}$, and the "$\circ$" notation refers to the element-wise matrix product, also called Schur or Hadamard product. It follows from (17) that the covariance matrix of $\{w_l(\mathbf{s}), \mathbf{s} \in \mathcal{S}\}$ is approximated by $\mathbf{C}^\dagger = \mathbf{C}_l + \mathbf{C}_s$, and the approximate covariance matrix of $\mathbf{Y}$ is $\mathbf{C}^\dagger + \tau^2\mathbf{I}$, where $\mathbf{C}^\dagger$ depends on the parameters $\boldsymbol{\theta}$ and σ^2. The approximate log likelihood function is

$$\begin{aligned}
&\ell_n(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \tau^2) \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\{\mathbf{C}_l(\boldsymbol{\theta}, \sigma^2) + \mathbf{C}_s(\boldsymbol{\theta}, \sigma^2) + \tau^2\mathbf{I}_n\} \\
&\qquad - \frac{1}{2}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta})^T\{\mathbf{C}_l(\boldsymbol{\theta}, \sigma^2) + \mathbf{C}_s(\boldsymbol{\theta}, \sigma^2) + \tau^2\mathbf{I}_n\}^{-1}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}).
\end{aligned} \tag{22}$$

Evaluation of the log likelihood requires calculation of the inverse and determinant of the $n \times n$ matrix

$$\mathbf{C}_l + \mathbf{C}_s + \tau^2\mathbf{I}_n = \mathbf{C}_{n,m}\mathbf{C}_{m,m}^{-1}\mathbf{C}_{n,m}^T + \mathbf{C}_s + \tau^2\mathbf{I}_n.$$

Applying the Sherman-Woodbury-Morrison formula for inverse matrices, we obtain

$$\begin{aligned}
&(\mathbf{C}_{n,m}\mathbf{C}_{m,m}^{-1}\mathbf{C}_{n,m}^T + \mathbf{C}_s + \tau^2\mathbf{I}_n)^{-1} \\
&\qquad = (\mathbf{C}_s + \tau^2\mathbf{I}_n)^{-1} - (\mathbf{C}_s + \tau^2\mathbf{I}_n)^{-1}\mathbf{C}_{n,m}\times \\
&\qquad\qquad \{\mathbf{C}_{m,m} + \mathbf{C}_{n,m}^T(\mathbf{C}_s + \tau^2\mathbf{I}_n)^{-1}\mathbf{C}_{n,m}\}^{-1}\mathbf{C}_{n,m}^T(\mathbf{C}_s + \tau^2\mathbf{I}_n)^{-1}.
\end{aligned} \tag{23}$$

On the other hand, the determinant can be computed using

$$\begin{aligned}
&\det(\mathbf{C}_s + \tau^2\mathbf{I}_n + \mathbf{C}_{n,m}\mathbf{C}_{m,m}^{-1}\mathbf{C}_{n,m}^T) \\
&\qquad = \det\{\mathbf{C}_{m,m} + \mathbf{C}_{n,m}^T(\mathbf{C}_s + \tau^2\mathbf{I}_n)^{-1}\mathbf{C}_{n,m}\}\{\det(\mathbf{C}_{m,m})\}^{-1}\det(\mathbf{C}_s + \tau^2\mathbf{I}_n).
\end{aligned} \tag{24}$$

According to (21), the $n \times n$ matrix $\mathbf{C}_s + \tau^2 \mathbf{I}_n$ is a sparse matrix. The right hand sides of (23) and (24) only involve inversion and determinant of and multiplication with sparse $n \times n$ matrices as well as the inversion and determinant of $m \times m$ matrices. Thus the computational complexity of the log likelihood calculation is of the order $O(nm^2 + nk^2)$, where m is the number of knots and k is the average number of nonzero entries per row in $\mathbf{C}_s$. By using a small number m and a short taper range γ (which results in a small k), the computational cost in fitting the spatial model can be greatly reduced relative to the expensive computational cost of using the original covariance function, where the computational complexity is typically of the order $O(n^3)$.

By the likelihood theory under the increasing domain asymptotic framework, the MLE is asymptotically normal with the covariance matrix being the inverse of the information matrix (Mardia and Marshall, 1984). In particular, the standard errors of the MLE's are estimated based on the inverse of the Fisher information matrix or the observed information matrix. Detailed derivation of the Fisher information is similar to (Mardia and Marshall, 1984) and omitted. Our simulation study shows that this information based variance estimation works well (results not shown to save space).

4.2. Bayesian inference of model parameters

The Bayesian inference for the model parameters begins with assigning prior (hyperprior) distributions to the model parameters (hyperparameters). We will follow the standard method for prior specifications. A vague multivariate normal prior is assigned to the regression coefficient vector $\boldsymbol{\beta}$. Inverse gamma priors could be assigned to the variance parameters σ^2 and τ^2, usually with a reasonable guess of mean and variance. Prior specifications for $\boldsymbol{\theta}$ will depend upon the choice of correlation function. For example, consider the Matérn correlation function in (18). Customarily, a uniform prior at $[0, 2]$ is assigned for the smoothness parameter ν since higher orders of smoothness are typically difficult to identify from real data. We assign a reasonably informative prior for the spatial range parameter ϕ, for example, a uniform prior with its support specified to reflect one's prior belief about the practical spatial range of the data.

Let $\Omega = (\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \tau^2)$ denote collectively the model parameters. The MCMC method is used to draw samples of the model parameters from the the posterior

$$p(\Omega \,|\, \mathbf{Y}) \propto p(\boldsymbol{\beta})\, p(\boldsymbol{\theta})\, p(\sigma^2)\, p(\tau^2)\, p(\mathbf{Y} \,|\, \Omega). \tag{25}$$

Sampling proceeds by first updating $\boldsymbol{\beta}$ from an MVN$(\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot}, \Sigma_{\boldsymbol{\beta}|\cdot})$ distribution with the covariance matrix

$$\Sigma_{\boldsymbol{\beta}|\cdot} = [\Sigma_{\boldsymbol{\beta}}^{-1} + \mathbf{X}^T \{\mathbf{C}_l(\boldsymbol{\theta}, \sigma^2) + \mathbf{C}_s(\boldsymbol{\theta}, \sigma^2) + \tau^2 \mathbf{I}\}^{-1} \mathbf{X}]^{-1}$$

and the mean

$$\boldsymbol{\mu}_{\boldsymbol{\beta}|\cdot} = \Sigma_{\boldsymbol{\beta}|\cdot} [\Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} + \mathbf{X}^T \{\mathbf{C}_l(\boldsymbol{\theta}, \sigma^2) + \mathbf{C}_s(\boldsymbol{\theta}, \sigma^2) + \tau^2 \mathbf{I}\}^{-1} \mathbf{Y}],$$

where $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\Sigma_{\boldsymbol{\beta}}$ are the mean and covariance matrix of the prior distribution of $\boldsymbol{\beta}$. For the parameters which do not have closed form posterior conditional distributions, we will need to draw samples using Metropolis–Hasting steps (Gelman et al., 2004). For example, for the Matérn correlation function, we may use a truncated normal proposal centered at the current value for updating the smoothness parameter ν, and a log-normal proposal centered

at the current log value for the range parameter ϕ. Log-normal proposals can also be used for the variance parameters σ^2 and τ^2. Following the MCMC sampling, posterior inferences such as posterior means and credible intervals are then made by computing summaries of the posterior samples.

Efficient computation can be achieved by using (23) and (24) for likelihood evaluation, Fisher information matrix calculation, and sampling from the posterior distribution.

4.3. Spatial prediction

Assuming the model parameters are known, the approximate BLUP at a new location $\mathbf{s}_0$ is

$$\hat{Y}(\mathbf{s}_0) = \mathbf{x}^T(\mathbf{s}_0)\boldsymbol{\beta} + \mathbf{h}^T(\mathbf{s}_0)(\mathbf{C}_l + \mathbf{C}_s + \tau^2\mathbf{I})^{-1}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}), \tag{26}$$

where $\mathbf{h}(\mathbf{s}_0) = [C_l(\mathbf{s}_0, \mathbf{s}_i) + C_s(\mathbf{s}_0, \mathbf{s}_i)]_{i=1}^n$, and the associated approximate mean squared prediction error is

$$\mathrm{MSE}[\hat{Y}(\mathbf{s}_0)] = \sigma^2 - \mathbf{h}^T(\mathbf{s}_0)(\mathbf{C}_l + \mathbf{C}_s + \tau^2\mathbf{I})^{-1}\mathbf{h}(\mathbf{s}_0) + \tau^2. \tag{27}$$

In practice, one needs to substitute estimates of the unknown model parameters in the above expressions.

The Bayesian inference seeks to find the predictive distribution $p[Y(\mathbf{s}_0) \,|\, \mathbf{Y}]$ at a new site $\mathbf{s}_0$. Generically denoting the set of all model parameters by Ω, we first obtain a set of posterior samples $\{\Omega^{(l)}, l = 1, \ldots, L\}$ from the posterior distribution $p[\Omega \,|\, \mathbf{Y}]$. For a given Ω value, $p[Y(\mathbf{s}_0) \,|\, \Omega, \mathbf{Y}]$ is a Gaussian distribution with the mean and the variance given by

$$E[Y(\mathbf{s}_0)|\Omega, \mathbf{Y}] = \mathbf{x}^T(\mathbf{s}_0)\boldsymbol{\beta} + \mathbf{h}^T(\mathbf{s}_0)(\mathbf{C}_l + \mathbf{C}_s + \tau^2\mathbf{I}_n)^{-1}(\mathbf{Y} - \mathbb{X}\boldsymbol{\beta}) \tag{28}$$

and

$$Var[Y(\mathbf{s}_0)|\Omega, \mathbf{Y}] = \sigma^2 - \mathbf{h}^T(\mathbf{s}_0)(\mathbf{C}_l + \mathbf{C}_s + \tau^2\mathbf{I}_n)^{-1}\mathbf{h}(\mathbf{s}_0) + \tau^2. \tag{29}$$

The predictive distribution is sampled by composition, drawing $Y^{(l)}(\mathbf{s}_0) \sim p[Y(\mathbf{s}_0) \,|\, \Omega^{(l)}, \mathbf{Y}]$ for each $\Omega^{(l)}, l = 1, \ldots, L$, where $\Omega^{(l)}$ is the lth sample from the posterior distribution $p[\Omega \,|\, \mathbf{Y}]$. Again, inversion of the matrix $\mathbf{C}_l + \mathbf{C}_s + \tau^2\mathbf{I}$ appeared in (26)–(29) can be efficiently computed using (23).

5. Illustrations

In this section, we use one simulation example and one real data example to illustrate the full-scale approximation and compare it with the predictive process and the covariance tapering approaches. The implementation of all methods was written in Matlab and run on a processor with dual 2.8 GHz Xeon CPUs and 12GB memory. For sparse matrix calculations, we used the Matlab function *sparse*. For the likelihood function optimization, we used the Matlab function *fmincon* which implements a Broyden-Fletcher-Goldfarb-Shanno (BFGS) based Quasi-Newton method. The R package *spam* (Furrer and Sain, 2010) for sparse matrix calculation is also available at `http://cran.r-project.org/web/packages/spam/`.

5.1. Simulation study

We considered a nonstationary random field that Banerjee et al. (2008) used in their simulation study on predictive process and for comparison purpose, we considered a Bayesian

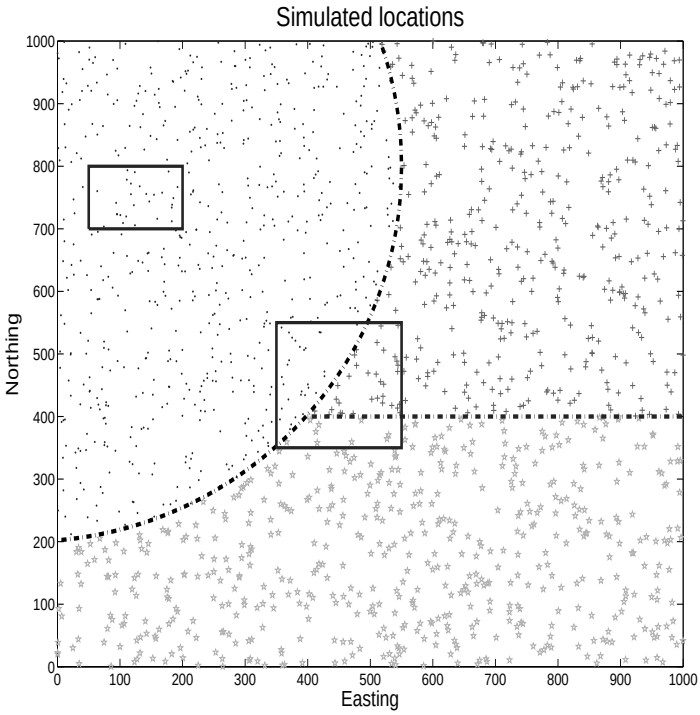

Fig. 4. Spatial distribution of simulated locations.

implementation of the full-scale approximation. We randomly selected 2,000 locations from the region $[0, 1000] \times [0, 1000]$ which in turn is partitioned into three subregions as shown in Fig. 4. We simulated a realization of the spatial process $Y(\mathbf{s})$ at these 2,000 locations using model (1). Three different intercepts were assigned to the three subregions. We adopted a nonstationary version of the Matérn covariance as used in Banerjee et al. (2008, Section 5.1.2) to simulate the spatial random effects (see also Paciorek and Schervish, 2006). The covariance function is

$$C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} |\Sigma_{D(\mathbf{s})}|^{1/4} |\Sigma_{D(\mathbf{s}')}|^{1/4} \left| \frac{\Sigma_{D(\mathbf{s})} + \Sigma_{D(\mathbf{s}')}}{2} \right|^{-1/2} \times \left(2\sqrt{\nu Q(\mathbf{s}, \mathbf{s}')} \right)^{\nu} J_\nu \left(2\sqrt{\nu Q(\mathbf{s}, \mathbf{s}')} \right),$$

where

$$Q(\mathbf{s}, \mathbf{s}') = (\mathbf{s} - \mathbf{s}')^T \left(\frac{\Sigma_{D(\mathbf{s})} + \Sigma_{D(\mathbf{s}')}}{2} \right)^{-1} (\mathbf{s} - \mathbf{s}')$$

with $D(\mathbf{s})$ indicating the subregion where $\mathbf{s}$ belongs to. Anisotropy is introduced to the covariance function by letting Σ depend on $D(\mathbf{s})$. We reparameterize

$$\Sigma_{D(\mathbf{s})} = G(\psi_{D(\mathbf{s})}) \begin{pmatrix} \lambda^2_{D(\mathbf{s}),1} & 0 \\ 0 & \lambda^2_{D(\mathbf{s}),2} \end{pmatrix} G^T(\psi_{D(\mathbf{s})}),$$

where $G(\psi)$ is the rotation matrix with angle ψ, i.e.,

$$\begin{pmatrix} \cos(\psi) & -\sin(\psi) \\ \sin(\psi) & \cos(\psi) \end{pmatrix}.$$

The true values of the parameters are presented in the second column of Table 1.

For the Bayesian posterior inference, flat priors were assigned to each of the three intercepts, $U(0, \pi/2)$ priors were assigned for the rotation angle ψ's, $U(1, d_{\max}/3)$ priors for the λ's, where $d_{\max}$ is the maximum distance of all pairs. The smoothness parameter ν was fixed to be .5. The variance parameters σ^2 and τ^2 were assumed to have inverse gamma priors with parameters based on reasonable guess of their means and variances. We applied the full-scale approximation with 225 knots and taper range 50. Knots were located on a uniform grid over the domain. We also applied the predictive process approximation with the same 225 knot locations, the covariance tapering with taper range 50, and the full covariance model. For each method, we ran 6,000 iterations to collect posterior samples after a burn-in period of 2,000 iterations, thinning using every third iteration. Trace plots of parameters indicated good convergence of the respective marginal distributions.

Table 1 shows the Bayesian posterior sample means and standard deviations of the model parameters for each approach. The posterior means obtained from the full-scale approximation have negligible biases compared with those from the full model. For the mean parameters β's, the three approximations all give results close to those of the full covariance model. For the variance parameters σ^2 and τ^2, and the correlation parameters λ's, the posterior means from the full-scale approximation are in general closer to the corresponding posterior means from the full covariance model, compared with two other approximations.

Table 1 also displays the MSPE for the posterior predictions at 500 randomly selected new locations. The full-scale approximation achieves similar prediction accuracy as using the full model. For example, when $m = 225$ and $\gamma = 50$, the MSPE is 0.6858, which is slightly higher than 0.6814, the MSPE based on the full model. For the predictive process model with $m = 225$, and the covariance tapering method with $\gamma = 50$, the values of MSPE are .7311 and 8328, respectively, which are higher.

We now examine the predictive performance and the computing time of the three approximation approaches for a larger simulated dataset. We generated realizations at 8000 random locations from the nonstationary model configuration described in the earlier part of this subsection. Two different sets of parameter values were used for the simulation respectively: One having the parameter values listed in the second column of Table 1, and the other having $\lambda_{11} = 16.69$, $\lambda_{12} = 66.7$, $\lambda_{21} = 5$, $\lambda_{22} = 50$,$\lambda_{31} = 66.7$, and $\lambda_{32} = 16.69$ with the remaining parameters unchanged. We split the data set into a training set with 7,000 data points and a testing set with 1,000 points. Among the testing points, 439 locations are in the rectangles $[350, 550] \times [350, 550]$ and $[50, 200] \times [700, 800]$ shown in Fig. 4, and another 561 locations are uniformly sampled from the rest of the area. This interpolation scenario is common in analyzing satellite data where missing data often correspond to sizable coverage gaps/holes. The training data are used to predict the testing data by plugging in the

Table 1. Posterior estimations of the model parameters, the MSPE and the DIC in the simulation study 2.

	TRUE value	Full model	$m = 225, r = 50$	$m = 225$	$r = 50$
β_0	50	49.94(0.19)	49.97(0.20)	49.86(0.19)	49.90(0.06)
β_1	10	10.01(0.19)	10.15(0.18)	10.25(0.23)	10.06(0.07)
β_2	25	25.08(0.17)	24.87(0.22)	24.94(0.26)	25.00(0.06)
ψ_1	0.78	0.88(0.04)	0.79(0.08)	0.57(0.08)	0.70(0.30)
ψ_2	1.31	1.28(0.16)	1.25(0.17)	1.32(0.16)	1.29(0.17)
ψ_3	0.44	0.46(0.05)	0.54(0.10)	0.65(0.09)	0.85(0.44)
λ_{11}	33.33	39.45(6.35)	41.92(6.64)	43.06(11.10)	91.19(37.02)
λ_{12}	100	127.43(18.60)	135.70(25.76)	124.81(31.65)	199.31(34.36)
λ_{21}	50	60.11(9.52)	70.13(12.82)	98.08(19.07)	176.47(47.48)
λ_{22}	50	66.63(10.64)	61.14(12.15)	78.87(23.30)	189.51(48.30)
λ_{31}	100	100.82(14.86)	116.57(17.92)	155.11(31.41)	212.19(24.83)
λ_{32}	33.33	31.65(4.73)	57.47(8.24)	70.43(15.26)	203.50(33.09)
σ^2	1	1.01(0.13)	1.04(0.08)	1.31(0.14)	0.82(0.04)
τ^2	0.2	0.21(0.01)	0.17(0.02)	0.54(0.02)	0.13(0.02)
MSPE	-	0.6814	.6858	0.7311	.8328
DIC	-	2325	2338	2443	2457

true parameter values into the BLUP equation. The prediction accuracy of each method is evaluated by the estimated MSPE based on the 1,000 testing data points at the hold-out locations. For each set of parameter values, we recorded the MSPE and the prediction run time under the three approaches for various choices of knot numbers and taper ranges. The results are shown in Fig. 5. The full-scale approximation clearly outperforms the other two approximations: It requires substantially less run time to reach the same MSPE. On the other hand, there is no clear winner between the predictive process and covariance tapering: The covariance tapering performs better than the predictive process for the first set of parameter values, but the opposite is true for the second set of parameter values.

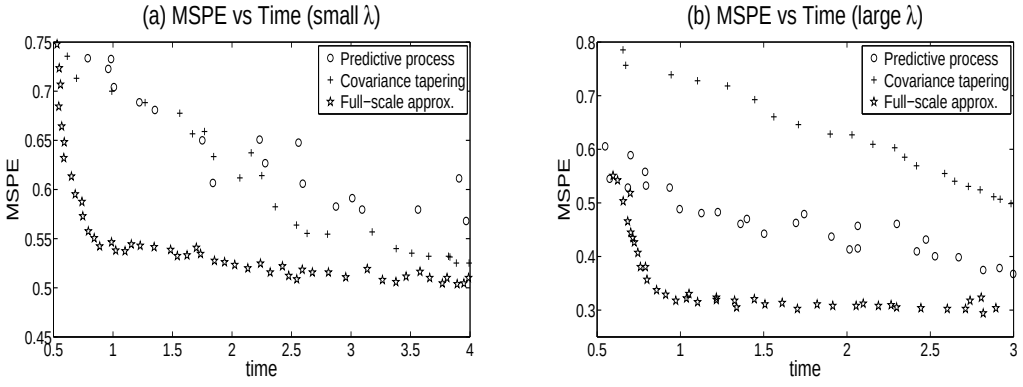

Fig. 5. The MSPE versus time plots for the simulation example under the full-scale approximation (diamond), the predictive process (circle) and the covariance tapering (plus). Figure (a) shows the result using $\lambda_{11} = 16.69$, $\lambda_{12} = 66.7$, $\lambda_{21} = 5$, $\lambda_{22} = 50$,$\lambda_{31} = 66.7$, $\lambda_{32} = 16.69$, and Figure (b) shows the result using $\lambda_{11} = 33.3$, $\lambda_{12} = 100$, $\lambda_{21} = 50$, $\lambda_{22} = 50$,$\lambda_{31} = 100$, and $\lambda_{32} = 33.3$ with the remaining parameters given in the second column of Table 1.

5.2. Real data analysis

We next illustrate our method using annual total precipitation anomalies at 7352 weather stations from the year 1962 in the United States. The annual precipitation anomalies are the annual rainfall totals standardized by subtracting the long-run mean and dividing the standard deviation for each station using the precipitation data from the National Climatic Data Center (NCDC) for the years 1895 to 1997 (Johns et al., 2003), Kaufman et al. (2008) noticed no obvious nonstationarity or anisotropy in this spatial data set. Therefore, we fit to data the spatial regression model (1) with the exponential covariance function, which is stationary and isotropic.

Our approach requires choice of the knot intensity and the taper range. It is evident that selection of knots and taper range involves tradeoff analysis between inference accuracy and computational cost. A full discussion on this issue is beyond the scope of this paper. For this data set, we did a crude tradeoff analysis using a training subset of size 2,000 and a test subset of size 1,000. Our pilot analysis calculated the MSPE on the test subset and the corresponding run time for prediction over different choices of m and γ. Weighing the tradeoff between MSPE and run time, we chose $m = 460$ and $\gamma = 25$ kilometers for further analysis on the full data set. The resulting tapered covariance matrix [i.e., (21)] is very sparse, only 0.086% of the off-diagonal entries are nonzero.

We split the full data set into a training set of 7,000 observations and a test set of 352 observations. The training set is used for parameter estimation and for computing the spatial prediction. The test set is used to evaluate prediction. Table 2 displays both the MLEs and the Bayesian estimates for the model parameters under the full model and each of the three approximations. We report the standard deviations for the MLEs based on the Fisher information and the standard deviations for the Bayesian estimates based on the posterior samples. However, for the predictive process estimates, it is inappropriate to use the Fisher information to estimate the standard deviations because of the biasness of the predictive process MLEs. Therefore, we do not report the standard deviations for the predictive process estimates. From Table 2, no significant difference is observed between the MLEs and Bayesian estimates. We can use the results from the full model as the standard to evaluate the quality of the three approximations. The full-scale approximation and the two-taper covariance approximation produce parameter estimates that are closer to the full model estimates than the predictive process. However, the estimate for the range parameter obtained from the two-taper approximation seems to be associated with larger standard deviation. Table 2 also displays the maximized log-likelihood for the MLEs and the deviance information criterion (DIC) (Spiegelhalter et al., 2002) for the Bayesian approach to compare the quality of model fitting. As is expected, the true covariance model has the highest log-likelihood and the lowest DIC score, both indicating the best model fitting. It is evident that the full-scale approximation outperforms the other two approaches in that it provides the maximized log-likelihood and the DIC score that are closest to those by the true covariance model.

Table 3 shows the MSPE on the test set, the run time for the likelihood evaluation on the training set, and the run time for the prediction on the test set. All three approaches reduce the computation time substantially, while the full-scale approximation leads to more accurate prediction in terms of MSPE with the same (or shorter) run time. Using 460 knots and the taper range 25 kilometers, the MSPE using the full-scale approximation is .2254 and it only takes about 1.7s each for the likelihood evaluation and prediction. In order to achieve the same level of accuracy as the full-scale approximation, either the knot intensity

Table 2. Real data example. MLEs and Bayesian inference results. The unit of the range parameter ϕ is in kilometers.

MLE	σ^2	τ^2	ϕ	Log lik
Full model	.6704 (.0649)	.1059 (.0044)	107.25 (11.97)	-5160.9
$m = 460, \gamma = 25$	.7825(.0715)	.0458(.0053)	139.39(13.88)	-5364.8
$m = 460$	.9976 (-)	.2601 (-)	186.20 (-)	-5999.1
$\gamma = 25$	.7569 (.0516)	.0296 (.0188)	81.11 (26.18)	-8548.3
Bayesian	σ^2	τ^2	ϕ	DIC
Full model	.6706 (.0671)	.1054 (.0056)	106.99 (11.70)	5164.3
$m = 460, \gamma = 25$	.7559(.0584)	.0453(.0049)	133.93(13.06)	5367.2
$m = 460$	1.0838 (.1208)	.2598 (.0069)	208.98 (32.02)	5733.9
$\gamma = 25$	.7539 (.0232)	.0362 (.0122)	113.17 (35.53)	8550.1

Table 3. MSPE, likelihood evaluation time and prediction time of the full-scale approximation, the predictive process and the covariance tapering.

	Full-scale Approximation	Predictive Process		Covariance Tapering	
	$m = 460, \gamma = 25$	$m = 460$	$m = 1705$	$\gamma = 25$	$\gamma = 100$
MSPE	0.2254	0.2851	.2400	.4205	0.2351
Lik. eval. time	1.69	.50	3.91	2.64	24.11
Pred. time	1.77	.54	4.27	.65	3.28

has to be increased for the predictive process approximation or the taper range has to be expanded for the covariance tapering approximation, which typically yield rapid growth in computational times. For the covariance tapering approach, the similar MSPE (.2351) is achieved when the taper range is increased to 100, the run time for the likelihood evaluation is more than 10 times of using our approach. For the predictive process, a similar MSPE (.2400) is obtained when the knot intensity is 1705 and the run time is nearly doubled compared with our approach.

6. Discussion

We have proposed a new approximation of covariance functions for modeling and analysis of very large point-referenced spatial data sets. This "full-scale" approximation can effectively capture both the large scale and small scale spatial variations. Through simulation studies, we have shown that the new approximation generally provides more efficient computation and substantially better performance in both model inference and prediction, compared with the reduced rank approximation and the covariance tapering approximation. While each of the two existing approaches has its own failure modes, our new approximation consistently performs well regardless of the dependence properties of the spatial covariance functions.

The full-scale approximation has two tuning parameters: the knot intensity m and the taper range γ. As we demonstrated in Section 5, it is evident that larger m and longer γ offer better approximation to the original covariance function, which, unfortunately, will result in higher computational cost. In the real data example, we used a subset of the data to decide on how to weigh the tradeoff between inference accuracy and computational cost. A more comprehensive study on the selection of knots and taper range is left for future research. It is also interesting to extend the current work to other contexts of spatial statistics, such

as analysis of data sets observed on a sphere, non-Gaussian spatial processes, multivariate and spatio-temporal processes, in which many existing spatial models face computational challenges when the size of the data set is large.

Acknowledgement

The research of Huiyan Sang and Jianhua Z. Huang was partially sponsored by NSF grant DMS-1007618. Jianhua Z. Huang's work was also partially supportedby NSF grant DMS-09-07170 and the NCI grant CA57030. Both authors were supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors thank the referees and the editors for valuable comments. The authors also thank Dr. Sudipto Banerjee, Dr. Reinhard Furrer, and Dr. Lan Zhou for several useful discussions regarding this work, and thank Dr. Cari Kaufman for providing the precipitation data set.

References

Baker, C. (1977). *The Numerical Treatment of Integral Equations*. Oxford: Clarendon press.

Banerjee, S., B. Carlin, and A. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall-CRC.

Banerjee, S., A. Finley, P. Waldmann, and T. Ericsson (2010). Hierarchical spatial process models for multiple traits in large genetic trials. *Journal of the American Statistical Association 105*(490), 506–521.

Banerjee, S., A. Gelfand, A. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Chiles, J. and P. Delfiner (1999). *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Cressie, N. (1993). *Statistics for Spatial Data, 2nd edn*. New York: Wiley.

Cressie, N. and G. Johannesson (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 209–226.

Diggle, P. and S. Lophaven (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics 33*(1), 53–64.

Diggle, P., J. Tawn, and R. Moyeed (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 47*(3), 299–350.

Finley, A., H. Sang, S. Banerjee, and A. Gelfand (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis 53*(8), 2873–2884.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association 102*(477), 321–331.

Furrer, R., M. Genton, and D. Nychka (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics 15*(3), 502–523.

Furrer, R. and S. Sain (2010). spam: A sparse matrix r package with emphasis on mcmc methods for gaussian markov random fields. *Journal of Statistical Software 36*(10), 1–25.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis 83*(2), 493–508.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In C. W. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi (Eds.), *Quantitative methods for current environmental issues*, pp. 37–54. London: Springer.

Horn, G. A. and C. R. Johnson (1985). *Matrix Analysis*. Cambridge: Cambridge University Press.

Johns, C., D. Nychka, T. Kittel, and C. Daly (2003). Infilling sparse records of spatial fields. *Journal of the American Statistical Association 98*(464), 796–806.

Kammann, E. and M. Wand (2003). Geoadditive models. *Journal of the Royal Statistical Society Series C(Applied Statistics) 52*(1), 1–18.

Kaufman, C., M. Schervish, and D. Nychka (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association 103*(484), 1545–1555.

Kaufman, L. and P. Rousseeuw (1990). *Finding groups in data*, Volume 16. Wiley New York.

Mardia, K. and R. Marshall (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika 71*(1), 135.

Paciorek, C. and M. Schervish (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics 17*(5), 483–506.

Pissanetzky, S. (2007). *Sparse Matrix Technology*. London: Academic Press.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman & Hall-CRC.

Rue, H. and H. Tjelmeland (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics 29*, 31–49.

Schabenberger, O. and C. Gotway (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton: Chapman & Hall.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64*(4), 583–639.

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

Stein, M. (2008). A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society 37*(1), 3–10.

Stein, M., Z. Chi, and L. Welty (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society. Series B, Statistical Methodology 66*, 275–296.

Vecchia, A. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological) 50*, 297–312.

Ver Hoef, J., N. Cressie, and R. Barry (2004). Flexible spatial models for kriging and cokriging using moving averages and the Fast Fourier Transform (FFT). *Journal of Computational and Graphical Statistics 13*(2), 265–282.

Wendland, H. (1995). Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics 4*(1), 389–396.

Wendland, H. (1998). Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *Journal of Approximation Theory 93*(2), 258–272.

Wikle, C. and N. Cressie (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika 86*(4), 815–829.

Williams, C. and M. Seeger (2001). Using the nystrom method to speed up kernel machines. In T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13*, pp. 682–688. MIT Press.

Appendix A: Proof of Proposition 1

We need only show that the Gram matrix of $C^\dagger(\cdot,\cdot)$ with respect to a set of spatial locations is positive semi-definite or positive definite if the corresponding Gram matrix of $C^\dagger(\cdot,\cdot)$ has the same property. We can assume that this set of locations contains all m knots used to definite the finite-rank covariance function $C_l(\cdot,\cdot)$ given in (15) because, if the Gram matrix with respect to a set of locations is positive semi-definite or positive definite, the Gram matrix with respect to a subset has the same property.

Partition the Gram matrix to 2×2 blocks

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix},$$

where the block $\mathbf{C}_{11}$ is the covariance matrix of $(w(\mathbf{s}_1),\ldots,w(\mathbf{s}_m))^T$ where $\mathbf{s}_1,\ldots,\mathbf{s}_m$ are the m knots. The Gram matrix of the full-scale approximated covariance function (17) can be written as

$$\begin{pmatrix} \mathbf{C}_{11} \\ \mathbf{C}_{21} \end{pmatrix} \mathbf{C}_{11}^{-1} \left(\mathbf{C}_{11}\mathbf{C}_{12}\right) + \left[\begin{pmatrix} \mathbf{C}_{11}\mathbf{C}_{12} \\ \mathbf{C}_{21}\mathbf{C}_{22} \end{pmatrix} - \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} \end{pmatrix} \right] \circ \begin{pmatrix} \mathbf{T}_{11}\mathbf{T}_{12} \\ \mathbf{T}_{21}\mathbf{T}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{11}\mathbf{C}_{12} \\ \mathbf{C}_{21}\widetilde{\mathbf{C}}_{22} \end{pmatrix},$$

where $\widetilde{\mathbf{C}}_{22} = \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} + \{\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\} \circ \mathbf{T}_{22}$, and $\mathbf{T}_{ij}$'s correspond to the partition of the Gram matrix of the tapering function.

Standard matrix calculation yields

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{C}_{21}\mathbf{C}_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\mathbf{C}_{11}^{-1}\mathbf{C}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12} \end{pmatrix}$$

and

$$\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{C}_{21}\mathbf{C}_{11}^{-1} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \widetilde{\mathbf{C}}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\mathbf{C}_{11}^{-1}\mathbf{C}_{12} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{0} \\ \mathbf{0} & \{\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\} \circ \mathbf{T}_{22} \end{pmatrix}.$$

Thus the matrices

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \widetilde{\mathbf{C}}_{22} \end{pmatrix}$$

are positive (semi-)definite if and only if the matrices $\mathrm{diag}(\mathbf{C}_{11}, \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})$ and $\mathrm{diag}\{\mathbf{C}_{11}, (\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}) \circ \mathbf{T}_{22}\}$ are positive (semi-)definite, respectively. According to the Schur product theorem (Horn and Johnson, 1985, Theorem 7.5.3), $\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$ is positive (semi-)definite if and only if $(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}) \circ \mathbf{T}_{22}$ is positive (semi-)definite. The desired results follow.