

Chapter 3

Geostatistics for Large Datasets

Ying Sun, Bo Li, and Marc G. Genton

Abstract We review various approaches for the geostatistical analysis of large datasets. First, we consider covariance structures that yield computational simplifications in geostatistics and briefly discuss how to test the suitability of such structures. Second, we describe the use of covariance tapering for both estimation and kriging purposes. Third, we consider likelihood approximations in both spatial and spectral domains. Fourth, we explore methods based on latent processes, such as Gaussian predictive processes and fixed rank kriging. Fifth, we describe methods based on Gaussian Markov random field approximations. Finally, we discuss multivariate extensions and open problems in this area.

3.1 Introduction

Due to the advancement of technology, massive amounts of data are often observed at a large number of spatial locations in geophysical and environmental sciences. There are many interesting aspects to discuss for the geostatistical analysis of large spatial datasets. Here we focus on computational issues, that is, how to make the geostatistical analysis of large datasets feasible or how to improve computational efficiency. This is crucial because spatial problems with modern data often overwhelm traditional implementations of spatial statistics, such as maximum likelihood estimation, Bayesian methods, and best linear unbiased prediction (kriging). In particular, large dimensional covariance matrices must be inverted. Moreover, many

Y. Sun (✉) · M.G. Genton

AQ1 Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA
e-mail: sunwards@stat.tamu.edu; genton@stat.tamu.edu

B. Li

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
e-mail: boli@stat.purdue.edu

geophysical processes are observed on a globe and it is common to have spatial data covering a large portion of the Earth. This requires special techniques to deal with large datasets observed over a sphere [27, 28]. Finally, the computational burden is aggravated in spatio-temporal settings and in multivariate situations where multiple observations occur at each location.

For instance, the exact computation of the likelihood of a Gaussian spatial random field observed at n irregularly sited locations generally requires $O(n^3)$ operations and $O(n^2)$ memory [43]. Therefore while sample sizes of $n = 1,000$ are no longer a challenge, $n = 1,000,000$ remains out of reach with classical procedures for even large clusters of processors. In a Bayesian framework, hierarchical models implemented through Markov Chain Monte Carlo (MCMC) methods have become especially popular for spatial modeling, given their flexibility and power to fit models that would be infeasible with classical methods. However, fitting hierarchical spatial models also involves expensive matrix operations whose computational complexity increases in a cubic order of the number of spatial locations n at every iteration of the MCMC algorithm [3], and thus here as well the computations can become problematic for large spatial datasets. Kriging, or spatial best linear unbiased prediction (BLUP), is an optimal interpolation in geostatistics. Solving the kriging equations directly requires the solution of a large linear system and involves inversion of an $n \times n$ covariance matrix \mathbf{C} , where $O(n^3)$ computations are required to obtain \mathbf{C}^{-1} [7, 18].

Because large dataset issues often arise from the difficulty of dealing with large covariance matrices, understanding and modeling covariance structures is the key to tackle this problem. Let $\{Z(\mathbf{x}); \mathbf{x} \in \mathcal{D} \subset \mathbb{R}^d\}$, $d \geq 1$, be a random field and $\mathbf{x}_1, \dots, \mathbf{x}_n$ be the sampling points in \mathcal{D} . For a second-order stationary random field, the covariance function, $C(\mathbf{k}) = \text{cov}\{Z(\mathbf{x}), Z(\mathbf{x} + \mathbf{k})\}$, is determined only by the lag \mathbf{k} but not the location \mathbf{x} . Here \mathbf{x} denotes the location and \mathbf{k} denotes the lag and they can be defined in either a purely spatial domain $\mathcal{D} = \mathcal{S}$ or a spatio-temporal domain $\mathcal{D} = \mathcal{S} \times \mathcal{T}$ depending on the nature of the data. Let \mathbf{h} denote the spatial lag and u the temporal lag, that is, $\mathbf{k} = \mathbf{h}$ for spatial data indexed by $\mathbf{x} = \mathbf{s} \in \mathcal{S}$ while $\mathbf{k} = (\mathbf{h}, u)$ for spatio-temporal data indexed by $\mathbf{x} = (\mathbf{s}, t) \in \mathcal{S} \times \mathcal{T}$. Further, we can define the second-order stationary covariance function for a multivariate random field $\{\mathbf{Z}(\mathbf{x}) = (Z_1(\mathbf{x}), \dots, Z_p(\mathbf{x}))^T; \mathbf{x} \in \mathcal{D} \subset \mathbb{R}^d\}$ by $C_{ij}(\mathbf{k}) = \text{cov}\{Z_i(\mathbf{x}), Z_j(\mathbf{x} + \mathbf{k})\}$, for $i, j = 1, \dots, p$, where p is the number of variables at each location. Compared to the $n \times n$ univariate covariance matrix \mathbf{C} induced by n spatial locations, the size of the multivariate cross-covariance matrix is inflated to $np \times np$.

There have been several approaches to overcome this large matrix problem, such as imposing separability on covariance functions, tapering the covariance matrix, using composite likelihoods, truncating the spectral representation of a random field, modeling the realizations by a latent process with reduced dimension, and approximating the random field with a Gaussian Markov random field. One common feature implied by all these methods is to sacrifice some unimportant information in the data in order to gain computational efficiency. How to define ‘‘unimportant’’ distinguishes these methods. Separability ignores the interaction between different

types of covariances so that the dimension of the covariance matrices to be inverted is reduced dramatically and thus facilitates computational procedures. Covariance tapering makes use of the computational advantage of sparse matrices obtained by zeroing out the “small” values in the covariance matrix. Composite likelihood methods throw away weak correlations between observations that are far apart, and spectral methods estimate parameters by truncating spectral representations. Latent process approaches keep only the most fundamental structure and wash out the details in the random field via adopting a low rank structure on the covariance matrix, which enables to simply deal with a matrix of low dimension rather than a large covariance matrix. While sharing this dimension reduction idea, many low rank models in different forms have been developed. Besides, Gaussian Markov random fields, for which the conditional distributions only depend on nearby neighbors, lead to sparseness of the precision matrix, the inverse of the covariance matrix.

All these methods can be regarded as approximations of the underlying random field. In [48], an efficient computing algorithm was thus developed for calculating different predictive scores which are measures of the predictive performance and assess how well an approximation works. Estimators are then constructed to minimize some prediction scores.

The remainder of this chapter is organized as follows. In Sect. 3.2, we review separable covariance structures and explains how they can facilitate computational procedures for large spatial datasets. Then, in Sect. 3.3, we describe the use of covariance tapering for both maximum likelihood estimation and kriging purposes. We provide some other practical ways to evaluate likelihood functions in both spatial and spectral domains in Sect. 3.4. Next, in Sect. 3.5, we introduce different forms of low rank models for latent process approaches, including Gaussian predictive process models and fixed rank kriging, and in Sect. 3.6 we discuss approximations using Gaussian Markov random fields. We review some existing methods and extensions for multivariate spatial datasets in Sect. 3.7, and the chapter ends with a discussion in Sect. 3.8.

3.2 Separable Covariance Structures

One way to deal with the computational issue of large covariance matrices is to take advantage of some special covariance structures. A class of such covariance structures that has been used widely is that of separable covariance functions. Separability is defined with different notions depending on the context. For example, a space-time separable covariance model is defined as $C(\mathbf{h}, u) = C(\mathbf{h}, 0)C(\mathbf{0}, u)/C(\mathbf{0}, 0)$. That is, the space-time covariance function can be factored into a product of a purely spatial covariance and a purely temporal covariance. Another notion of separability, also called intrinsic model and defined only for multivariate random fields, is that $\mathbf{C}(\mathbf{k}) = \rho(\mathbf{k})\mathbf{A}$ for a spatial or spatio-temporal correlation function $\rho(\mathbf{k})$ and a $p \times p$ positive definite matrix \mathbf{A} . This type of separability indicates that the covariance between variables is independent of the covariance induced by spatial locations.

The aforementioned separable covariance functions can significantly reduce the dimension of the covariance matrices that need to be inverted, hence they can alleviate the computational demand. This is because separability enables the large covariance matrix to be a Kronecker product of smaller matrices and then only the inversions of those smaller matrices are required due to nice properties of Kronecker products. For instance, a spatio-temporal dataset observed at 100 spatial locations and 100 time points leads to a covariance matrix of size $10,000 \times 10,000$, for which the inversion is difficult to compute. However, employing a space-time separable covariance function decomposes this large covariance matrix into a Kronecker product of two square matrices each of size 100×100 with one being the spatial covariance matrix and the other the temporal covariance matrix. Then the inversion of the matrix of size $10,000 \times 10,000$ is reduced to the inversion of two matrices of size 100×100 . Similar gains can be achieved when separable models are used in multivariate spatial or spatio-temporal data analysis.

Despite their attractive properties, separable models are not always appropriate for real data due to their lack of flexibility to allow for interactions between different types of correlations. [21] illustrated the lack of space-time separability for an Irish wind data and [6] suggested a nonseparable space-time covariance underlying a tropical wind dataset in the Pacific Ocean. Further, [32] also demonstrated the lack of multivariate separability for a trivariate pollution data over California. A variety of tests have been developed to assess the appropriateness of space-time separability. Among those, [31] proposed a unified nonparametric framework to test many different assumptions, including separability, made on the covariance structure. Their approach is based on the asymptotic normality of covariance estimators and can be easily implemented without assuming any specific marginal or joint distribution of the data other than some mild moment and mixing conditions. Later, this test was extended by [32] to assess separability for multivariate covariance functions, for which no effective and formal methods had been developed. Based on the testing framework in [31, 32], [40] proposed a self-normalization idea in place of subsampling methods to estimate the covariance of empirical covariance estimators. This new estimating method avoids the choice of the optimal block size required by the subsampling method.

In the case of lack of separability, [20] described a nearest Kronecker product approach to find separable approximations of nonseparable space-time covariance matrices. His main idea was to identify two small matrices that minimize the Frobenius norm of the difference between the original covariance matrix and the Kronecker product of those two matrices. His data example with Irish wind speeds showed that the prediction deteriorated only slightly whereas large computational savings were obtained.

Other structures of the covariance function can lead to simplifications too. For instance, a spatial random field in \mathbb{R}^2 with an isotropic stationary covariance function yields a symmetric block Toeplitz covariance matrix with Toeplitz blocks. Recall that a matrix is said to be of Toeplitz form if its entries are constant on each diagonal. Kriging can then be performed more efficiently with such a structured covariance matrix [50]. Stationarity of the random field can be tested with the procedure of [26] and then isotropy with the method of [22].

3.3 Tapering

The idea of tapering is to set the covariances at large distances to zero but still keep the original covariances for proximate locations. This is done in a way such that the new matrix retains the property of positive definiteness while efficient sparse matrix algorithms can be used. However, since tapering strictly zeroes out weak correlations, a natural question is whether statistical inference based on the tapered version shares the same desirable properties with the untapered exact solution. This question is answered separately in two sections: Sect. 3.3.1 focuses on properties of the maximum likelihood estimator (MLE) of the covariance parameters, and Sect. 3.3.2 discusses spatial interpolation using kriging with known covariance functions.

3.3.1 Tapering for Estimation

We consider data drawn from a zero-mean, stationary and isotropic Gaussian random field Z . Let $C(h; \boldsymbol{\theta})$ be the parametric covariance function between any two observations whose locations are apart by a distance h . The parameter vector $\boldsymbol{\theta} \in \mathbb{R}^b$ needs to be estimated from a finite number of observations, $\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))^T$. Since the vector \mathbf{Z} follows a multivariate normal distribution, we have the log-likelihood function for $\boldsymbol{\theta}$

$$l(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{C}(\boldsymbol{\theta})| - \frac{1}{2} \mathbf{Z}^T \mathbf{C}(\boldsymbol{\theta})^{-1} \mathbf{Z}, \quad (3.1)$$

where $\mathbf{C}(\boldsymbol{\theta})$ is a $n \times n$ covariance matrix with the (i, j) th element equal to $C(\|\mathbf{s}_i - \mathbf{s}_j\|; \boldsymbol{\theta})$, $i, j = 1, \dots, n$. [30] proposed a method of covariance tapering to approximate the log-likelihood (3.1). Then, focusing on the particular case of the Matérn class of covariance functions, they illustrated the behavior of the MLE.

The tapering idea is particularly suitable if the correlations between observations far apart are negligible. This type of structure can then be modeled by a compactly supported covariance function. Let a tapering function $C_{tap}(h; \gamma)$ be an isotropic correlation function of compact support, that is, $C_{tap}(h; \gamma) = 0$ whenever $h \geq \gamma$ for some $\gamma > 0$. Denote a tapered covariance function by

$$\tilde{C}(h; \boldsymbol{\theta}, \gamma) = C(h; \boldsymbol{\theta}) C_{tap}(h; \gamma), \quad h > 0. \quad (3.2)$$

The tapered covariance matrix defined by $\tilde{\mathbf{C}}$ is a Schur (or Hadamard) product $\tilde{\mathbf{C}}(\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)$, where $\mathbf{T}(\gamma)_{ij} = C_{tap}(\|\mathbf{s}_i - \mathbf{s}_j\|; \gamma)$, or $\tilde{C}_{ij} = \tilde{C}(\|\mathbf{s}_i - \mathbf{s}_j\|; \boldsymbol{\theta}, \gamma)$. The tapered covariance matrix is positive definite, since the elementwise matrix product of two covariance matrices is again a valid covariance matrix. In addition, it has high proportion of zero elements when γ is small and is, therefore, a sparse matrix. Hence it can be inverted much more efficiently than inverting a full matrix of the same dimension when evaluating the log-likelihood.

Using covariance tapering, [30] proposed two approximations of the log-likelihood in (3.1). The first approximation simply replaces the model covariance matrix $\mathbf{C}(\boldsymbol{\theta})$ by $\mathbf{C}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)$, yielding

$$l_{1tap}(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)| - \frac{1}{2}\mathbf{Z}^T[\mathbf{C}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]^{-1}\mathbf{Z} \quad (3.3)$$

with biased score function, that is, $E[\frac{\partial}{\partial\boldsymbol{\theta}}l_{1tap}(\boldsymbol{\theta})] \neq \mathbf{0}$. This means that there is no guarantee that the estimator that maximizes (3.3) is asymptotically unbiased. To correct the bias, the second approximation takes an estimating equations approach. First, rewrite $\mathbf{Z}^T\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{Z} = \text{tr}\{\widehat{\mathbf{C}}\mathbf{C}(\boldsymbol{\theta})^{-1}\}$, where $\widehat{\mathbf{C}} = \mathbf{Z}\mathbf{Z}^T$ is the sample covariance matrix. Then replace both the model and sample covariance matrices with tapered versions yielding

$$\begin{aligned} l_{2tap}(\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)| - \frac{1}{2}\text{tr}\left\{[\widehat{\mathbf{C}} \circ \mathbf{T}(\gamma)][\mathbf{C}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]^{-1}\right\} \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)| - \frac{1}{2}\mathbf{Z}^T\left\{[\mathbf{C}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)]^{-1} \circ \mathbf{T}(\gamma)\right\}\mathbf{Z}. \end{aligned}$$

Maximizing $l_{2tap}(\boldsymbol{\theta})$ now corresponds to solving an unbiased estimating equation for $\boldsymbol{\theta}$, that is, $E[\frac{\partial}{\partial\boldsymbol{\theta}}l_{2tap}(\boldsymbol{\theta})] = \mathbf{0}$.

In both approximations, small values of γ correspond to more severe tapering. When $\gamma = 0$, observations are treated as independent, and as $\gamma \rightarrow \infty$, we recover the original likelihood. For the particular case of the Matérn class of covariance functions, it has been shown that the estimators maximizing the tapering approximations, such as the MLE, are strongly consistent under certain conditions.

[11] then investigated how the tapering affects the asymptotic efficiency of the MLE for parameters in the Matérn covariance function under the assumption that data are collected along a line in a bounded region. Their results imply that, under some conditions on the taper, the tapered MLE is asymptotically as efficient as the true MLE. Recently, [39] showed that under suitable asymptotics, maximum tapered likelihood estimators are consistent and asymptotically normal for a wide range of covariance models.

[19] proposed a combination of tapering and backfitting to estimate the fixed and random spatial component parameters in a very general type of mixed model. They were able to model and analyze spatial datasets several orders of magnitude larger than those analyzed with classical approaches. Tapering techniques in Kalman filter updates were studied in [16].

3.3.2 Tapering for Kriging

Instead of parameter estimation, [18] addressed the problem of covariance tapering for interpolation of large spatial datasets. In geostatistics the standard approach, termed kriging, is based on the principle of minimum mean squared prediction

error. Starting with the simplest spatial model, we assume that $Z(\mathbf{s})$ is observed without any measurement error. Then the best linear unbiased prediction (BLUP) at an unobserved location $\mathbf{s}_0 \in \mathcal{S}$ is

$$\widehat{Z}(\mathbf{s}_0) = \mathbf{c}_0 \mathbf{C}^{-1} \mathbf{Z}, \quad (3.4)$$

where $\mathbf{C}_{ij} = C(\mathbf{s}_i, \mathbf{s}_j)$ and $\mathbf{c}_{0i} = C(\mathbf{s}_i, \mathbf{s}_0)$ are based on a possibly nonstationary covariance function C . The mean squared prediction error $\text{MSPE}(\mathbf{s}_0, C)$ has the form

$$\text{MSPE}(\mathbf{s}_0, C) = C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{c}_0^T \mathbf{C}^{-1} \mathbf{c}_0. \quad (3.5)$$

Similar to (3.2), let $\widetilde{C}(\mathbf{s}, \mathbf{s}_0) = C(\mathbf{s}, \mathbf{s}_0) C_{\text{taper}}(\mathbf{s}, \mathbf{s}_0; \gamma)$. By replacing the covariance matrix \mathbf{C} by the tapered version defined by \widetilde{C} , the linear system defining the weights in (3.4) can be solved efficiently. The implication is that we limit the covariance function to a local neighborhood. In general we expect the weights $\mathbf{c}_0 \mathbf{C}^{-1}$ in (3.4) to be close to zero for observations whose locations are far from \mathbf{s}_0 . The localization of the weights in the prediction equation motivates kriging using only a neighborhood of locations.

However, if the BLUP (3.4) is calculated under the covariance function \widetilde{C} , the mean squared prediction error is of the form

$$\text{MSPE}(\mathbf{s}_0, \widetilde{C}) = C(\mathbf{s}_0, \mathbf{s}_0) - 2\widetilde{\mathbf{c}}_0^T \widetilde{\mathbf{C}}^{-1} \mathbf{c}_0 + \widetilde{\mathbf{c}}_0^T \widetilde{\mathbf{C}}^{-1} \mathbf{C} \widetilde{\mathbf{C}}^{-1} \widetilde{\mathbf{c}}_0, \quad (3.6)$$

where the tilde terms are based on \widetilde{C} . For the Matérn covariance family, [18] showed that under specific conditions the asymptotic mean squared error of the predictions in (3.6) using the tapered covariance converges to the minimal error in (3.5). It was also shown that the naive prediction error $\text{MSPE}(\mathbf{s}_0, \widetilde{C})$, assuming that \widetilde{C} is the true covariance function, has the correct convergence rate as well. As can be seen, covariance tapering for kriging purpose is an approximation to the standard linear spatial predictor which is justified to be both asymptotically accurate and computationally efficient.

3.4 Likelihood Approximations

Likelihood approaches for large irregularly spaced spatial datasets are often very difficult, if not infeasible, to implement due to computational limitations. Tapering methods in Sect. 3.3 approximate the Gaussian likelihood through sparse covariance matrices. In this section, we review some other practical ways to evaluate likelihood functions in both spatial and spectral domains.

3.4.1 Likelihood Approximations in the Spatial Domain

In a spatial setting, [46] suggested a simple likelihood approximation. The idea is that any joint density can be written as a product of conditional densities based on some ordering of the observations. Then, one way to lessen the computations is to condition on only some of the “past” observations when computing the conditional densities. Specifically, suppose that $\mathbf{Z} = (Z_1, \dots, Z_n)^T$ has joint density $p(\mathbf{z}; \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is a vector of unknown parameters. By partitioning \mathbf{Z} into subvectors $\mathbf{Z}_1, \dots, \mathbf{Z}_b$ of possibly different lengths and defining $\mathbf{Z}_{(j)}^T = (\mathbf{Z}_1^T, \dots, \mathbf{Z}_j^T)$, we always have

$$p(\mathbf{z}; \boldsymbol{\phi}) = p(\mathbf{z}_1; \boldsymbol{\phi}) \prod_{j=2}^b p(\mathbf{z}_j | \mathbf{z}_{(j-1)}; \boldsymbol{\phi}). \quad (3.7)$$

To calculate the conditional densities $p(\mathbf{z}_j | \mathbf{z}_{(j-1)}; \boldsymbol{\phi})$, it may not be crucial to condition on all components of $\mathbf{z}_{(j-1)}$ for the purpose of reducing the computational effort. In particular, if, for $j = 1, \dots, b-1$, $\mathbf{V}_{(j)}$ is some subvector of $\mathbf{Z}_{(j)}$, then we have the approximation:

$$p(\mathbf{z}; \boldsymbol{\phi}) \approx p(\mathbf{z}_1; \boldsymbol{\phi}) \prod_{j=2}^b p(\mathbf{z}_j | \mathbf{v}_{(j-1)}; \boldsymbol{\phi})$$

which is the general form of Vecchia’s approximation to the likelihood. For Gaussian \mathbf{Z} , the best linear predictor (BLP) of \mathbf{Z}_j given $\mathbf{Z}_{(j-1)}$ is the conditional expectation $E(\mathbf{Z}_j | \mathbf{Z}_{(j-1)}; \boldsymbol{\phi})$ as a function of $\boldsymbol{\phi}$, and therefore, $p(\mathbf{z}_j | \mathbf{z}_{(j-1)}; \boldsymbol{\phi})$ is the density of the error of the BLP of \mathbf{Z}_j given $\mathbf{Z}_{(j-1)}$. Vecchia’s approximation is accomplished by replacing this density with the one for errors of the BLP of \mathbf{Z}_j given $\mathbf{V}_{(j-1)}$.

[44] adapted Vecchia’s approach for the full likelihood to approximate the restricted likelihood of a Gaussian process and showed that the approximation gives unbiased estimating equations. Suppose that $\mathbf{Z} \sim_n (\mathbf{X}\boldsymbol{\beta}, \mathbf{C}(\boldsymbol{\theta}))$, where \mathbf{X} is a known $n \times q$ matrix of rank q , $\boldsymbol{\beta} \in \mathbb{R}^q$ is a vector of unknown coefficients and $\boldsymbol{\theta} \in \Theta$ is a length r vector of unknown parameters for the covariance matrix of \mathbf{Z} , then $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\theta})$. For estimating $\boldsymbol{\theta}$, the maximum likelihood acts as if $\boldsymbol{\beta}$ were known and, hence, tends to underestimate the variation of the spatial process. Restricted maximum likelihood (REML) avoids this problem by considering $\boldsymbol{\beta}$ as nuisance and estimating $\boldsymbol{\theta}$ by using only contrasts, or linear combinations of the observations whose means do not depend on $\boldsymbol{\beta}$.

Just as the full likelihood can be written in terms of the densities of errors of BLPs, the restricted likelihood can also be written in terms of the densities of errors of best linear unbiased predictors (BLUPs) similar to equation (3.7). Specifically, let \mathbf{Z}_i have length n_i and take \mathbf{X}_i to be the corresponding n_i rows of \mathbf{X} assuming that $\text{rank}(\mathbf{X}_1) = q$. For $j > 1$, let $\mathbf{B}_j(\boldsymbol{\theta})$ be the $n_j \times n$ matrix such that

$\mathbf{W}_j(\boldsymbol{\theta}) = \mathbf{B}_j(\boldsymbol{\theta})\mathbf{Z}$ is the vector of errors of the BLUP of \mathbf{Z}_j based on $\mathbf{Z}_{(j-1)}$. For $j = 1$, take $\mathbf{B}_1(\boldsymbol{\theta})$ to be a matrix independent of $\boldsymbol{\theta}$ of size $(n_1 - q) \times n$ such that $\mathbf{W}_1(\boldsymbol{\theta}) = \mathbf{B}_1(\boldsymbol{\theta})\mathbf{Z}$ is a set of contrasts of \mathbf{Z}_1 . Then $\mathbf{W}_j(\boldsymbol{\theta}) \sim N_{n_j}(\mathbf{0}, \mathbf{A}_j(\boldsymbol{\theta}))$ where $\mathbf{A}_j(\boldsymbol{\theta}) = \mathbf{B}_j(\boldsymbol{\theta})\mathbf{C}(\boldsymbol{\theta})\mathbf{B}_j^T(\boldsymbol{\theta})$. We then could obtain the restricted likelihood, which only depends on $\boldsymbol{\phi}$ through $\boldsymbol{\theta}$:

$$rl(\boldsymbol{\theta}; \mathbf{Z}) = \frac{n-q}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^b \left[\log\{\det(\mathbf{A}_j)\} + \mathbf{W}_j^T \mathbf{A}_j^{-1} \mathbf{W}_j \right].$$

Now consider approximating the restricted likelihood by computing the BLUP of \mathbf{Z}_j in terms of some subvector $\mathbf{V}_{(j-1)}$ of $\mathbf{Z}_{(j-1)}$. The BLUP of, say, Z_ℓ given some subvector \mathbf{S} of \mathbf{Z} that does not contain Z_ℓ is just the linear combination $\boldsymbol{\lambda}^T \mathbf{S}$ that minimizes $\text{var}(Z_\ell - \boldsymbol{\lambda}^T \mathbf{S})$ subject to $E(Z_\ell - \boldsymbol{\lambda}^T \mathbf{S}) = 0$ for all values of $\boldsymbol{\beta}$. Let \mathbf{V} be the collection of subvectors $\mathbf{V}_{(1)}, \dots, \mathbf{V}_{(b-1)}$. Define $\mathbf{W}_1(\mathbf{V}) = \mathbf{W}_1$ and, for $j > 1$, $\mathbf{W}_j(\mathbf{V})$ is the error of the BLUP of \mathbf{Z}_j based on $\mathbf{V}_{(j-1)}$. Let $\mathbf{A}_j(\mathbf{V})$ be the covariance matrix of $\mathbf{W}_j(\mathbf{V})$. Then the approximation to $rl(\boldsymbol{\theta}; \mathbf{Z})$ is of the form

$$rl(\boldsymbol{\theta}; \mathbf{V}) = \frac{n-q}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^b \left[\log\{\det(\mathbf{A}_j(\mathbf{V}))\} + \mathbf{W}_j^T(\mathbf{V}) \mathbf{A}_j(\mathbf{V})^{-1} \mathbf{W}_j(\mathbf{V}) \right]. \quad (3.8)$$

Having this restricted likelihood approximation, [44] showed that equation (3.8) gives a set of unbiased estimating equations for $\boldsymbol{\theta}$. The properties of its solutions were studied using the well-developed theory of estimating equations, and the effectiveness of various choices for \mathbf{V} was also investigated. [46] only considered prediction vectors of length 1 such that $\mathbf{Z}_j = Z_j$, whereas [44] considered prediction vectors \mathbf{Z}_j of length greater than 1 and added more observations in the conditioning set rather than just the nearest neighbors in order to further reduce the computational effort and to improve the efficiency of the estimated parameters. However, difficulties with the composite likelihoods of [46] and [44] arise with the selection of the observation order and of the conditioning sets as pointed out by [45], who reviewed recent developments of composite likelihood. To overcome such complications, three different likelihood approximations together with their statistical properties all based on splitting the data into blocks were proposed and investigated by [4, 5].

3.4.2 Likelihood Approximations in the Spectral Domain

The method proposed by [44] is a spatial domain approach. There are also some spectral methods which give another way to approximate the likelihood without involving the calculation of determinants, and to obtain the MLEs of the covariance

parameters θ . These methods are based on [47]’s approximation to the Gaussian negative log-likelihood, which can only be used for datasets observed on a regular complete lattice. In this situation, fewer calculations are required. For irregularly spaced data, [15] presented a version of Whittle’s approximation to the Gaussian negative log-likelihood by introducing a lattice process. Additional computational savings were obtained by truncating the spectral representation of the lattice process.

Suppose Z is a continuous Gaussian spatial process of interest with a covariance function C , observed at m irregularly spaced locations in \mathbb{R}^2 . Let f_Z be the stationary spectral density of Z , which is the Fourier transform of the covariance function:

$$f_Z(\boldsymbol{\omega}) = \frac{1}{(2\pi)^2} \int_{\mathbb{R}^2} \exp(-i\mathbf{h}^T \boldsymbol{\omega}) C(\mathbf{h}) d\mathbf{h}.$$

We define a process Y at location \mathbf{s} as the integral of Z in a block of area Δ^2 centered at \mathbf{s} ,

$$Y(\mathbf{s}) = \Delta^{-2} \int h(\mathbf{s} - \tilde{\mathbf{s}}) Z(\tilde{\mathbf{s}}) d\tilde{\mathbf{s}}, \quad (3.9)$$

where for $\mathbf{u} = (u_1, u_2)$ we have

$$h(\mathbf{u}) = \begin{cases} 1, & \text{if } |u_1| < \Delta/2, \quad |u_2| < \Delta/2, \\ 0, & \text{otherwise.} \end{cases}$$

Then Y is also a stationary process with spectral density f_Y given by

$$f_Y(\boldsymbol{\omega}) = \Delta^{-2} |\Gamma(\boldsymbol{\omega})|^2 f_Z(\boldsymbol{\omega}),$$

where $\Gamma(\boldsymbol{\omega}) = \int h(\mathbf{u}) e^{-i\boldsymbol{\omega}^T \mathbf{u}} d\mathbf{u} = [2 \sin(\Delta\omega_1/2)/\omega_1][2 \sin(\Delta\omega_2/2)/\omega_2]$ and $\boldsymbol{\omega} = (\omega_1, \omega_2)^T$.

For small values of Δ , $f_Y(\boldsymbol{\omega})$ is approximately $f_Z(\boldsymbol{\omega})$, because we have

$$\lim_{\Delta \rightarrow 0} \Delta^{-2} |\Gamma(\boldsymbol{\omega})|^2 = 1.$$

By (3.9), $Y(\mathbf{s})$ can be treated as a continuous spatial process defined for all $\mathbf{s} \in \mathcal{S}$, but here we consider the process Y only on a lattice $(n_1 \times n_2)$ of sample size $m = n_1 n_2$. That is, the values of \mathbf{s} in (3.9) are the centroids of the m grid cells in the lattice, where the spacing between neighboring sites is Δ . Then the spectrum of observations of the sample sequence $Y(\Delta\mathbf{s})$, for $\mathbf{s} \in \mathbb{Z}^2$, is concentrated within the finite-frequency band $-\pi/\Delta \leq \boldsymbol{\omega} < \pi/\Delta$ (aliasing phenomenon). The spectral density $f_{\Delta, Y}$ of the process on the lattice can be written in terms of the spectral density f_Y of the process Y as

$$f_{\Delta,Y}(\boldsymbol{\omega}) = \sum_{\mathbf{Q} \in \mathbb{Z}^2} f_Y(\boldsymbol{\omega} + 2\pi\mathbf{Q}/\Delta). \quad (3.10)$$

[15] justified that in practice the sum in (3.10) can be truncated after $2m$ terms.

Whittle's approximation to the Gaussian negative log-likelihood is of the form

$$\frac{m}{(2\pi)^2} \sum_{\boldsymbol{\omega}} \{\log f_{\Delta}(\boldsymbol{\omega}) + I_m(\boldsymbol{\omega}) f_{\Delta}(\boldsymbol{\omega})^{-1}\}, \quad (3.11)$$

where the sum is evaluated at the Fourier frequencies, I_m is the periodogram, and f_{Δ} is the spectral density of the lattice process. Now for the lattice process of Y , by computing the periodogram, Whittle's approximate likelihood (3.11) can be applied to $f_{\Delta,Y}$, written in terms of f_Y , then f_Z . Therefore, we can obtain the MLE for the covariances/spectral density parameters of Z . [15] also showed that this version of Whittle's approximation converges to the exact negative log-likelihood for Y , and if n is the total number of observations of the process Z , the calculation requires $O(m \log_2 m + n)$ operations rather than $O(n^3)$ for the exact likelihood of Z .

Another spectral method proposed by [34] extends the definition of a periodogram for time series to the situation where the sampling locations are irregularly spaced. They showed that the well-known property for time series that the periodogram at different frequencies are asymptotically independent still holds for irregularly spaced data. Therefore, it allows for nonparametric and parametric spatial spectral estimators similar to the classical time series analysis setting.

For a stationary random field Z observed at irregularly spaced locations in \mathbb{R}^d , by assuming some distribution of the sampling locations, [34] defined the periodogram based on a finite Fourier transform of $Z(\mathbf{s})$ as well as a tapered version of the periodogram. Just as the methods of estimating spectral densities in time series analysis, both nonparametric and parametric estimators were then proposed. The nonparametric method is the spectral window estimator and the parametric approach is based on Whittle's likelihood approximation using the proposed periodogram. Their asymptotic properties were studied and comparisons with [44] and [15] were reported on numerical examples. [44] focused on high frequency components to estimate parameters which better capture the behavior at very short distances, while [34] focused on low frequency components and [15] did on both high and low frequency components. In terms of computational considerations, the latter two spectral methods have a clear advantage.

3.5 Latent Processes

Statistics for spatial data also faces the problem of dealing with noisy data when the interest is in inference on unobserved latent processes. For large spatial datasets, one way to speed up computation is from the perspective of data dimension reduction. [3] developed a spatial model, called a Gaussian predictive process model, which we

introduce in Sect. 3.5.1. There, an approximation of the latent process is proposed to achieve dimension reduction. In Sect. 3.5.2, another solution is given by [7] who defined a spatial mixed effects model for the latent process and proposed fixed rank kriging within a flexible family of nonstationary covariance functions.

First, we define a latent process. Let $\{Y(\mathbf{s}); \mathbf{s} \in \mathcal{S} \subset \mathbb{R}^d\}$ be a real-valued spatial process. We are interested in making inference on the latent process Y on the basis of data that have measurement error. Consider the process Z defined by

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \mathbf{s} \in \mathcal{S},$$

where $\{\epsilon(\mathbf{s}); \mathbf{s} \in \mathcal{S}\}$ is a spatial white noise process with mean 0, $\text{var}\{\epsilon(\mathbf{s})\} = \tau^2 \nu(\mathbf{s}) \in (0, \infty)$ for $\tau^2 > 0$ and a known $\nu(\cdot)$. In fact, the process Z is observed only at a finite number of spatial locations. Define the vector of available data to be $\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))^T$.

The hidden process Y is assumed to have a linear mean structure,

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + \omega(\mathbf{s}), \quad \mathbf{s} \in \mathcal{S},$$

where $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), \dots, x_q(\mathbf{s}))^T$ represents a $q \times 1$ vector of known covariates; the coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T$ are unknown, and the process ω has zero mean, $0 < \text{var}\{\omega(\mathbf{s})\} < \infty$, for all $\mathbf{s} \in \mathcal{S}$, and a generally nonstationary spatial covariance function:

$$\text{cov}\{\omega(\mathbf{s}), \omega(\mathbf{s}')\} = C(\mathbf{s}, \mathbf{s}'), \quad \mathbf{s}, \mathbf{s}' \in \mathcal{S}.$$

We discuss techniques to reduce computational burden for large spatial datasets under the model

$$Z(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + \omega(\mathbf{s}) + \epsilon(\mathbf{s}). \quad (3.12)$$

That is, $\mathbf{Z} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \mathbf{C} + \tau^2\mathbf{V}$, where $\mathbf{X} = [\mathbf{x}^T(\mathbf{s}_i)]_{i=1}^n$ is a matrix of regressors, $\mathbf{C} = [C(\mathbf{s}_i, \mathbf{s}_j)]_{i,j=1}^n$ and $\mathbf{V} = \text{diag}\{\nu(\mathbf{s}_1), \dots, \nu(\mathbf{s}_n)\}$.

3.5.1 Gaussian Predictive Processes

With regard to the challenge of computational cost on covariance matrices, [3] proposed a class of models based on the idea of a spatial predictive process which is motivated by kriging ideas. The predictive process projects the original process onto a subspace generated by realizations of the original process at a specified set of locations (or knots). The approach is in the same spirit as process modeling approaches using basis functions and kernel convolutions, that is, specifications which attempt to facilitate computations through lower dimensional process representations.

Assume at locations $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^2$, a response variable $Z(\mathbf{s})$ is observed from model (3.12), where $\omega(\mathbf{s})$ is a zero-centered Gaussian Process (GP) with covariance function $C(\mathbf{s}, \mathbf{s}')$ capturing the effect of unobserved covariates with spatial pattern, and $\epsilon(\mathbf{s}) \sim N(0, \tau^2)$ is an independent process, often called the nugget, that models the measurement error. In applications, we usually specify $C(\mathbf{s}, \mathbf{s}') = \sigma^2 \rho(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$ where $\rho(\cdot; \boldsymbol{\theta})$ is a correlation function and $\boldsymbol{\theta}$ includes parameters in the covariance function. The likelihood for n observations is based on $\mathbf{Z} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_Z)$, with $\boldsymbol{\Sigma}_Z = \mathbf{C}(\boldsymbol{\theta}) + \tau^2 \mathbf{I}_n$, where $\mathbf{C}(\boldsymbol{\theta}) = [C(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$. To project the original or parent process onto a subspace, consider the lower-dimensional subspace chosen by the user by selecting a set of knots, $\mathcal{S}^* = \{\mathbf{s}_1^*, \dots, \mathbf{s}_m^*\}$, which may or may not form a subset of the entire collection of observed locations \mathcal{S} . The predictive process $\tilde{\omega}(\mathbf{s})$ is defined as the kriging interpolator

$$\tilde{\omega}(\mathbf{s}) = E[\omega(\mathbf{s}) | \boldsymbol{\omega}^*] = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \boldsymbol{\omega}^*, \quad (3.13)$$

where $\boldsymbol{\omega}^* = [\omega(\mathbf{s}_i^*)]_{i=1}^m \sim N_m(\mathbf{0}, \mathbf{C}^*(\boldsymbol{\theta}))$ is derived from the parent process realization over the knots in \mathcal{S}^* , $\mathbf{C}^*(\boldsymbol{\theta}) = [C(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\theta})]_{i,j=1}^m$ is the corresponding $m \times m$ covariance matrix, and $\mathbf{c}(\mathbf{s}; \boldsymbol{\theta}) = [C(\mathbf{s}, \mathbf{s}_j^*; \boldsymbol{\theta})]_{j=1}^m$.

The predictive process $\tilde{\omega}(\mathbf{s}) \sim GP(0, \tilde{C}(\cdot))$ defined in (3.13) has nonstationary covariance function,

$$\tilde{C}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \mathbf{c}(\mathbf{s}'; \boldsymbol{\theta}),$$

and is completely specified by the parent covariance function. Realizations associated with \mathbf{Z} are given by $\tilde{\boldsymbol{\omega}} = [\tilde{\omega}(\mathbf{s}_i)]_{i=1}^n \sim N_n(\mathbf{0}, \mathbf{c}^T(\boldsymbol{\theta}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \mathbf{c}(\boldsymbol{\theta}))$, where $\mathbf{c}^T(\boldsymbol{\theta})$ is the $n \times m$ matrix whose i th row is given by $\mathbf{c}^T(\mathbf{s}_i; \boldsymbol{\theta})$. The theoretical properties of the predictive process including its role as an optimal projection have been discussed in [3].

Replacing $\omega(\mathbf{s})$ in model (3.12) with $\tilde{\omega}(\mathbf{s})$, we obtain the predictive process model

$$Z(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + \tilde{\omega}(\mathbf{s}) + \epsilon(\mathbf{s}). \quad (3.14)$$

Since in (3.13), $\tilde{\omega}(\mathbf{s})$ is a spatially varying linear transformation of $\boldsymbol{\omega}^*$, the dimension reduction is seen immediately. In fitting model (3.14), the n random effects $\{\omega(\mathbf{s}_i), i = 1, \dots, n\}$ are replaced with only m random effects in $\boldsymbol{\omega}^*$. So we can work with an m -dimensional joint distribution involving only $m \times m$ matrices. Although we introduced the same set of parameters in both models (3.12) and (3.14), they are not identical.

The predictive process systematically underestimates the variance of the parent process $\omega(\mathbf{s})$ at any location \mathbf{s} , since we have $\text{var}\{\tilde{\omega}(\mathbf{s})\} = \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \mathbf{c}(\mathbf{s}; \boldsymbol{\theta})$, $\text{var}\{\omega(\mathbf{s})\} = C(\mathbf{s}, \mathbf{s})$ and $0 \leq \text{var}\{\omega(\mathbf{s}) | \boldsymbol{\omega}^*\} = C(\mathbf{s}, \mathbf{s}) - \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \mathbf{c}(\mathbf{s}; \boldsymbol{\theta})$. In practical implementations, this often reveals itself by overestimating the nugget variance in model (3.12), where the estimated τ^2 roughly captures $\tau^2 + E\{C(\mathbf{s}, \mathbf{s}) - \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \mathbf{c}(\mathbf{s}; \boldsymbol{\theta})\}$. Here $E\{C(\mathbf{s}, \mathbf{s}) - \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta}) \mathbf{C}^{*-1}(\boldsymbol{\theta}) \mathbf{c}(\mathbf{s}; \boldsymbol{\theta})\}$ denotes the

averaged bias underestimation over the observed locations. Indeed, [3] observed that while predictive process models employing a few hundred knots excelled in estimating most parameters in several complex high-dimensional models for datasets involving thousands of data points, reducing this upward bias in τ^2 was problematic.

To remedy this problem, [14] proposed a modified predictive process, defined as $\tilde{\omega}(\mathbf{s}) = \hat{\omega}(\mathbf{s}) + \tilde{\epsilon}(\mathbf{s})$, where now $\tilde{\epsilon}(\mathbf{s}) \sim N(0, C(\mathbf{s}, \mathbf{s}) - \mathbf{c}^T(\mathbf{s}; \boldsymbol{\theta})\mathbf{C}^{*-1}(\boldsymbol{\theta})\mathbf{c}(\mathbf{s}; \boldsymbol{\theta}))$ is a process of independent variables but with spatially adaptive variances. It is then easy to see that $\text{var}\{\tilde{\omega}(\mathbf{s})\} = C(\mathbf{s}, \mathbf{s}) = \text{var}\{\omega(\mathbf{s})\}$, as desired. Furthermore, $E\{\tilde{\omega}(\mathbf{s})|\omega^*\} = \hat{\omega}(\mathbf{s})$ which ensures that $\tilde{\omega}(\mathbf{s})$ inherits the attractive properties of $\hat{\omega}$ discussed by [3]; see also [2]. A recent application of predictive process models to a complex dataset from forestry can be found in [13].

3.5.2 Fixed Rank Kriging

Kriging, the spatial optimal interpolation, involves the inversion of covariance matrices. Straightforward kriging of massive datasets is not possible and ad hoc local kriging neighborhoods are typically used. One remedy is to approximate the kriging equations, for example by means of covariance tapering as we discussed in Sect. 3.3.2. Another approach is to choose classes of covariance functions for which kriging can be done exactly, even though the spatial datasets are large. One advantage of having a spatial model that allows exact computations is that there is no concern about how close the approximate kriging predictors and approximate mean squared prediction errors are to the corresponding theoretical values. For exact methods, two important questions arise: how flexible are the spatial covariance functions that are used for kriging and how are they fitted. [7] constructed a multiresolution spatial process and showed that there is a very rich class of spatial covariances for which kriging of large spatial datasets can be carried out both exactly and extremely rapidly, with computational complexity linear in the size of the data. They showed how to specify the $n \times n$ covariance matrix $\boldsymbol{\Sigma}$ so that $\boldsymbol{\Sigma}^{-1}$ can be obtained by inverting $m \times m$ positive definite matrices, where m is fixed and independent of n . The result is a spatial BLUP (kriging) procedure which they called Fixed Rank Kriging (FRK); see also [41].

For model (3.12), the kriging predictor of $Y(\mathbf{s}_0)$ in terms of the covariance function is

$$\hat{Y}(\mathbf{s}_0) = \mathbf{x}^T(\mathbf{s}_0)\hat{\boldsymbol{\beta}} + \mathbf{g}^T(\mathbf{s}_0)(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}), \quad (3.15)$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z}$, $\mathbf{g}^T(\mathbf{s}_0) = \mathbf{c}^T(\mathbf{s}_0) \boldsymbol{\Sigma}^{-1}$ and $\mathbf{c}(\mathbf{s}_0) = [C(\mathbf{s}_0, \mathbf{s}_j)]_{j=1}^n$.

The FRK method captures the scales of spatial dependence through a set of m (not necessarily orthogonal) basis functions, $\mathbf{B}(\mathbf{s}) = (B_1(\mathbf{s}), \dots, B_m(\mathbf{s}))^T$, for $\mathbf{s} \in \mathbb{R}^d$, where m is fixed. For any $m \times m$ positive definite matrix \mathbf{G} , we model $\text{cov}\{Y(\mathbf{s}), Y(\mathbf{s}')\}$ according to

$$C(\mathbf{s}, \mathbf{s}') = \mathbf{B}^T(\mathbf{s})\mathbf{G}\mathbf{B}(\mathbf{s}'), \quad \mathbf{s}, \mathbf{s}' \in \mathbb{R}^d, \quad (3.16)$$

which can be shown to be a valid covariance function. It is easy to see that expression (3.16) is a consequence of writing $\omega(\mathbf{s}) = \mathbf{B}^T(\mathbf{s})\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is an m -dimensional vector with $\text{var}(\boldsymbol{\eta}) = \mathbf{G}$. [7] called the model for ω a spatial random effects model. Hence, $Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\beta} + \mathbf{B}^T(\mathbf{s})\boldsymbol{\eta}$ is a spatial mixed effects model.

From expression (3.16), we can write the $n \times n$ covariance matrix as

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{G}\mathbf{B}^T + \tau^2\mathbf{V}. \quad (3.17)$$

Both \mathbf{B} , the $n \times m$ matrix whose (i, l) th element is $\mathbf{B}_l(\mathbf{s}_i)$, and \mathbf{V} , a diagonal matrix with entries given by the measurement error variances, are assumed known. Further,

$$\text{cov}\{Y(\mathbf{s}_0), \mathbf{Z}\} = \mathbf{c}^T(\mathbf{s}_0) = \mathbf{B}^T(\mathbf{s}_0)\mathbf{G}\mathbf{B}^T.$$

[7] showed that the choice of the covariance function (3.16) allows alternative ways of computing the kriging equations involving inversion of only $m \times m$ matrices. From equation (3.17),

$$\boldsymbol{\Sigma}^{-1} = \tau^{-1}\mathbf{V}^{-1/2} \left\{ \mathbf{I} + (\tau^{-1}\mathbf{V}^{-1/2}\mathbf{B})\mathbf{G}(\tau^{-1}\mathbf{V}^{-1/2}\mathbf{B})^T \right\}^{-1} \tau^{-1}\mathbf{V}^{-1/2}. \quad (3.18)$$

Then we have that, for any $n \times m$ matrix \mathbf{P} ,

$$\mathbf{I} + \mathbf{P}\mathbf{G}\mathbf{P}^T = \mathbf{I} + (\mathbf{I} + \mathbf{P}\mathbf{G}\mathbf{P}^T)\mathbf{P}\mathbf{G}(\mathbf{I} + \mathbf{P}^T\mathbf{P}\mathbf{G})^{-1}\mathbf{P}^T.$$

Premultiplying by $(\mathbf{I} + \mathbf{P}\mathbf{G}\mathbf{P}^T)^{-1}$ yields

$$(\mathbf{I} + \mathbf{P}\mathbf{G}\mathbf{P}^T)^{-1} = \mathbf{I} - \mathbf{P}(\mathbf{G}^{-1} + \mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T,$$

which is a particular case of the well-known Sherman-Morrison-Woodbury formulae. Using this in equation (3.18), it yields the computational simplification

$$\boldsymbol{\Sigma}^{-1} = (\tau^2\mathbf{V})^{-1} - (\tau^2\mathbf{V})^{-1}\mathbf{B}\{\mathbf{G}^{-1} + \mathbf{B}^T(\tau^2\mathbf{V})^{-1}\mathbf{B}\}^{-1}\mathbf{B}^T(\tau^2\mathbf{V})^{-1}. \quad (3.19)$$

The formula (3.19) for $\boldsymbol{\Sigma}^{-1}$ involves inverting only fixed rank $m \times m$ positive definite matrices and the $n \times n$ diagonal matrix \mathbf{V} . Finally, the kriging predictor (3.15) is

$$\hat{Y}(\mathbf{s}_0) = \mathbf{x}^T(\mathbf{s}_0)\hat{\boldsymbol{\beta}} + \mathbf{B}^T(\mathbf{s}_0)\mathbf{G}\mathbf{B}^T\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}$ and $\boldsymbol{\Sigma}^{-1}$ is given by equation (3.19).

For a fixed number of regressors q and a fixed rank m of \mathbf{G} in the covariance model that is defined by (3.16), [7] showed that the computational burden of FRK is only linear in n . The results rely on using a rich class of nonstationary covariance

functions (3.16) that arise from a spatial random effects model. Further, microscale variation in the hidden process Y could be modeled by including another diagonal matrix in equation (3.17),

$$\Sigma = \mathbf{BGB}^T + \xi^2 \mathbf{I} + \tau^2 \mathbf{V}.$$

When both diagonal matrices are proportional to each other, the measurement error parameter τ^2 and the microscale parameter ξ^2 are not individually identifiable, although their sum $\xi^2 + \tau^2$ is. The presence of the microscale variance ξ^2 was discussed by [7] but they have assumed that the process Y is smooth (i.e. $\xi^2 = 0$). The FRK formulae including ξ^2 were given by [8].

Statistics for spatio-temporal data inherits a similar need for data dimension reduction as what we saw for spatial data, possibly more so since the data size quickly becomes massive as time progresses. [9] built a spatio-temporal random effects model that allows both dimension reduction (spatially) and rapid smoothing, filtering, or forecasting (temporally). They focused on filtering and developed a methodology called Fixed Rank Filtering (FRF); see also [29]. With a similar idea as FRK, the fast statistical prediction of a hidden spatio-temporal process is achieved through spatio-temporal models defined on a space of fixed dimension; the space is defined by the random coefficients of prespecified spatio-temporal basis functions, and the coefficients are assumed to evolve dynamically. By reducing the dimensionality, FRF was proposed as a spatio-temporal Kalman filter, which is able to use past data as well as current data to great effect when estimating a process from a noisy, incomplete, and very large spatio-temporal dataset. [9] showed that the gains can be substantial when the temporal dependence is strong and there are past data at or near locations where the current data have gaps.

3.6 Gaussian Markov Random Field Approximations

Gaussian Markov random fields (GMRFs) possess appealing computational properties due to the sparse pattern of their precision matrices. The numerical factorization of the precision matrix using sparse matrix algorithms can be done at a typical cost of $O(n^{3/2})$ for two-dimensional GMRFs. The computational gains from GMRFs have been exploited to provide fast and accurate Bayesian inference for latent Gaussian field models through integrated nested Laplace approximation (INLA) by [35]. [12] made a further step of applying INLA on top of predictive process models [3] to dramatically reduce the computation in making inference for large spatial datasets. More generally, [36] demonstrated empirically that GMRFs could closely approximate some commonly used covariance functions in geostatistics, and proposed to use GMRFs as computational replacements for Gaussian random fields for example when making kriging predictions [24]. However, their approximation was restricted to Gaussian random fields that are observed over a regular lattice (or torus) and the fit itself had to be precomputed for a discrete set of parameter values. Other literature following this idea includes [1, 10, 25, 42].

[33] recently proposed an approach to find GMRFs with local neighborhood and precision matrix to represent certain Gaussian random fields with Matérn covariance structure. This is accomplished through the following two facts:

- (a) The solution of a particular type of stochastic partial differential equation (SPDE) driven by Gaussian white noise is a Gaussian random field with Matérn covariance function. Specifically, let $X(\mathbf{s})$ be the solution of the following linear fractional SPDE:

$$(\kappa^2 - \Delta)^{\alpha/2} X(\mathbf{s}) = W(\mathbf{s}), \quad \mathbf{s} \in \mathbb{R}^d, \quad \alpha = \nu + d/2, \quad \kappa > 0, \quad \nu > 0,$$

where $W(\mathbf{s})$ is a spatial Gaussian white noise with unit variance. Then $X(\mathbf{s})$ is a Gaussian random field with the Matérn covariance

$$C(\mathbf{k}) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} (\kappa \|\mathbf{k}\|)^\nu K_\nu(\kappa \|\mathbf{k}\|),$$

where

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}}.$$

Here the Laplacian $\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial x_i^2}$, K_ν is the modified Bessel function of second kind with order $\nu > 0$, $\kappa > 0$ is a scaling parameter and σ^2 is the marginal variance. The integer value of ν determines the mean square differentiability of the underlying random field.

- (b) Let X be a GMRF on a regular two-dimensional lattice indexed by (i, j) , where the Gaussian full conditionals are

$$E(X_{i,j} | \mathbf{X}_{-(i,j)}) = \frac{1}{a} (X_{i-1,j} + X_{i+1,j} + X_{i,j-1} + X_{i,j+1}),$$

and $\text{var}(X_{i,j} | \mathbf{X}_{-(i,j)}) = 1/a$ for $|a| > 4$, where $\mathbf{X}_{-(i,j)}$ denotes the vector of X 's on the lattice except at location (i, j) . The coefficients in the GMRF representation of the SPDE in (a) over a regular unit-distance two-dimensional infinite lattice for $\nu = 1, 2, \dots$, can be found by convolving the elements of the precision matrix corresponding to X by itself ν times.

The authors then generalized the above results to enable the construction of the corresponding GMRFs representation of the Matérn field on a triangulated lattice, hence the Gaussian random fields in [33] are no longer restricted to lattice data. This avoids the interpolation of irregularly spaced observations to grid points and allows for finer resolution where details are required. The drawback of this approach is that we can only find the explicit form of GMRFs for those Gaussian random fields that have a Matérn covariance structure at certain integer smoothnesses. Nevertheless, the main results in [33] cover the most important and used covariance models in

spatial statistics, and they can be extended to model Matérn covariances on the sphere, nonstationary locally isotropic Gaussian random fields, Gaussian random fields with oscillating correlation functions, and non-isotropic fields.

3.7 Multivariate Geostatistics

Multivariate spatial data have been increasingly employed in various scientific areas. We have introduced the intrinsic model, a separable covariance structure for multivariate data, to reduce computational cost in Sect. 3.2. In addition, in many statistical applications, predicting a geophysical quantity based on observations at nearby locations of the same quantity and on other related variables, so-called covariables, is of prime interest. Obviously, the analysis should take advantage of covariances between locations as well as covariances among variables. For a single variable of interest, we have discussed the tapering technique for kriging purpose in Sect. 3.3.2 and the fixed rank kriging method in Sect. 3.5.2. When information from several different variables is also available, it should be used for prediction as well. The problems implied by large amounts of data are then further amplified since many observations occur at each location. Therefore, we need methodologies to keep the analysis computationally feasible.

For spatially correlated multivariate random fields, the best linear unbiased prediction (BLUP) is often called cokriging in geostatistics. Assume that we have a primary variable and two or more secondary variables and aim at predicting the primary variable at some location. It is well-known that in a mean squared prediction error (MSPE) sense, the best predictor is the conditional expectation given variables at the other locations, where the set of conditioning variables can be either just the primary variable (i.e., kriging) or some or all of the secondary variables (i.e., cokriging).

Thus, the cokriging technique requires the solution of a large linear system based on the covariance and cross-covariance matrix of all involved variables. For large amounts of data, it is impossible to solve the linear system with direct methods. [17] proposed aggregation-cokriging for highly-multivariate spatial datasets to reduce the computational burden. This method is based on a linear aggregation of the covariables with carefully chosen weights, so that the resulting linear combination of secondary variables contributes as much as possible to the prediction of the primary variable in the MSPE sense. In other words, the secondary variables should be weighted by the strength of their correlation with the location of interest. The prediction is then performed using a simple cokriging approach with the primary variable and the aggregated secondary variables. This reduces the computational burden of the prediction from solving a $(n + \ell m) \times (n + \ell m)$ to solving a $(n + m) \times (n + m)$ linear system, where n and m are the numbers of observations of the primary and secondary variables, respectively, and ℓ is the number of secondary variables. The computational complexity is now comparable with simple bikriging, i.e., simple cokriging with only one of the secondary variables, and its optimality was discussed by [17] under different covariance structures.

Besides cokriging, Gaussian predictive process models can also be generalized to multivariate settings. In Sect. 3.5.1, we have discussed a class of models based on the idea of a univariate spatial predictive process which is motivated by kriging ideas. The predictive process projects the original process onto a subspace generated by realizations of the original process at a specified set of locations (or knots). Similarly, for multivariate spatial processes, the multivariate predictive process extends the preceding concepts to multivariate Gaussian processes. Now $\omega(\mathbf{s})$ from the model (3.12) is assumed to be a p -dimensional zero-centered multivariate Gaussian process $\omega(\mathbf{s})$, where p is the number of variables at each location. For locations $\mathbf{s}_1, \dots, \mathbf{s}_n$, we write the multivariate realizations as a vector $\omega = [\omega(\mathbf{s}_i)]_{i=1}^n \in \mathbb{R}^{np}$. Analogous to the univariate setting, [3] again considered a set of knots \mathcal{S}^* and denoted by ω^* the realization of $\omega(\mathbf{s})$ over \mathcal{S}^* . Then similar to (3.13), the multivariate predictive process is defined as

$$\tilde{\omega}(\mathbf{s}) = \text{cov}\{\omega(\mathbf{s}), \omega^*\} \text{var}^{-1}(\omega^*) \omega^*,$$

and $\tilde{\omega}(\mathbf{s})$ has properties that are analogous to its univariate counterpart. The projection onto a lower dimensional subspace allows the flexibility to accommodate multivariate processes in the context of large datasets.

3.8 Discussion

Whenever we deal with large spatial datasets, we face problems with storage and computation. When covariance functions produce covariance matrices that are neither sparse nor low rank, it is sometimes possible to compute the exact likelihood even for quite large datasets if they are spatially gridded data. However, exact likelihood calculations are not possible with large numbers of irregularly sited observations.

One solution that has been discussed is to truncate the covariance function to zero and use well-established algorithms to handle sparse systems. Such libraries or toolboxes are available in widely used software packages such as R or Matlab. The tapering for kriging purpose presented in Sect. 3.3.2 is based on the assumption of Matérn covariances which could be weakened, but not entirely eliminated. However, the Matérn family is already sufficiently flexible to model a broad class of processes. The tapering techniques also work for nonstationarity or anisotropic processes at least with conservative taper ranges, but the accuracy of the tapering approximation for nonstationary problems remains an open question [18].

The approximation of the likelihood in either the spatial or spectral domain is another solution to overcome computational obstacles. In the spatial domain, the composite likelihood method in Sect. 3.4.1 is based on conditional densities, which points out the difficulty in choosing conditioning sets and the need for less haphazard rules. Furthermore, the use of this approximation in Bayesian analysis poses considerable challenges, since the approximation accuracy needs to be evaluated especially if the likelihood calculation is just one part of a single step

in a MCMC algorithm [44]. The spectral methods in Sect. 3.4.2 are computationally efficient by avoiding the calculation of determinants and can be easily adapted to model nonstationary processes as a mixture of independent stationary processes [15]. However, they do not overcome the difficulty in prediction with massive data. The Gaussian predictive process model in Sect. 3.5.1 projects the parent process onto a lower dimensional subspace. Since every spatial or spatio-temporal process induces a predictive process model, it is flexible to accommodate nonstationary, non-Gaussian, possibly multivariate, possibly spatio-temporal processes in the context of large datasets. In the same spirit, the fixed rank kriging in Sect. 3.5.2 relies on using a class of nonstationary covariances where kriging can be done exactly. Those techniques can be implemented on very large datasets as the computations are linear in the size of the dataset, and they are highly flexible since they allow the underlying spatial covariances to be nonstationary. Section 3.6 described another approach based on Gaussian Markov random field approximations.

The covariance tapering and the reduced rank based methods have shown great computational gains, but they also have their own drawbacks. The covariance tapering may not be effective in accounting for spatial dependence with long range while the reduced rank based methods usually fail to accurately capture the local, small scale dependence structure. To capture both the large and small scale spatial dependence, [37] proposed to combine the ideas of the reduced rank process approximation and the sparse covariance approximation. They decomposed 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. This idea was then extended to the multivariate setting by [38]. However, the application of tapering techniques to multivariate random fields remains to be explored due to the lack of flexible compactly supported cross-covariance functions.

[49] discussed some strategies to deal with computations for large datasets within the context of their convolution-based spatial nonstationary models. Basically, for parameter estimation they proposed to smooth raw estimates obtained over apriori determined grids, and for predictions they discussed the idea of local window as in [23] and tapering as in [18].

For spatio-temporal processes, we have reviewed the methods to compute separable approximations of space-time covariance matrices. A well-known shortcoming of separable covariance functions is that they do not allow for space-time interactions in the covariance. Nevertheless, the separable space-time structure allows for a simple construction of valid space-time parametric models. By assuming separability, one can further combine separable approximations with the tapering approach. In this case, it is expected that a combination of computational gains can be achieved [20]. When dealing with several variables evolving in space and time, that is, multivariate space-time random fields, the cokriging approaches are even more computationally costly. In this context, the separable approximation techniques can also be combined with other approaches for multivariate spatial processes to further facilitate computational procedures.

In summary, we have compared various methods for handling large spatial datasets from the literature reviewed in this chapter. However, it would be interesting to further compare all of them under different situations with Monte Carlo simulation studies and real data examples. We look forward to the emergence of such work.

Acknowledgements The authors thank Reinhard Furrer for valuable comments on the manuscript. Li's research was partially supported by NSF grant DMS-1007686. Genton's research was partially supported by NSF grants DMS-1007504 and DMS-1100492, and by Award No. KUSC1-016-04, made by King Abdullah University of Science and Technology (KAUST).

References

1. Allcroft, D.J., Glasbey, C.A.: A latent Gaussian Markov random field model for spatio-temporal rainfall disaggregation. *J. R. Stat. Soc. Ser. C* **52**, 487–498 (2003)
2. Banerjee, S., Finley, A.O., Waldmann, P., Ericsson, T.: Hierarchical spatial process models for multiple traits in large genetic trials. *J. Amer. Statist. Assoc.* **105**, 506–521 (2010)
3. Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H.: Gaussian predictive process models for large spatial datasets. *J. R. Stat. Soc. Ser. B* **70**, 825–848 (2008)
4. Caragea, P.C., Smith, R.L.: Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *J. Multivariate Anal.* **98**, 1417–1440 (2007)
5. Caragea, P.C., Smith, R.L.: Approximate likelihoods for spatial processes. *Technometrics*. In press (2011)
6. Cressie, N., Huang, H.C.: Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94**, 1330–1340 (1999)
7. Cressie, N., Johannesson, G.: Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B* **70**, 209–226 (2008)
8. Cressie, N., Kang, E.L.: High-resolution digital soil mapping: Kriging for very large datasets. In *Proximal Soil Sensing*, eds R. Viscarra-Rossel, A. B. McBratney, and B. Minasny. Elsevier, Amsterdam, 49–66 (2010)
9. Cressie, N., Shi, T., Kang, E.L.: Fixed rank filtering for spatio-temporal data. *J. Comput. Graph. Statist.* **19**, 724–745 (2010)
10. Cressie, N., Verzelen, N.: Conditional-mean least-squares fitting of Gaussian Markov random fields to Gaussian fields. *Comput. Statist. Data Anal.* **52**, 2794–2807 (2008)
11. Du, J., Zhang, H., Mandrekar, V.S.: Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann. Statist.* **37**, 3330–3361 (2009)
12. Eidsvik, J., Finley, A., Banerjee, S., Rue, H.: Approximate Bayesian inference for large spatial datasets using predictive process models. Manuscript (2011)
13. Finley, A.O., Banerjee, S., MacFarlane, D.W.: A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *J. Amer. Statist. Assoc.* **106**, 31–48 (2011)
14. Finley, A.O., Sang, H., Banerjee, S., Gelfand, A.E.: Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53**, 2873–2884 (2009)
15. Fuentes, M.: Approximate likelihood for large irregularly spaced spatial data. *J. Amer. Statist. Assoc.* **102**, 321–331 (2007)
16. Furrer, R., Bengtsson, T.: Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.* **98**, 227–255 (2007)
17. Furrer, R., Genton, M.G.: Aggregation-cokriging for highly-multivariate spatial data. *Biometrika*. **98**, 615–631 (2011)

18. Furrer, R., Genton, M.G., Nychka, D.: Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Statist.* **15**, 502–523 (2006)
19. Furrer, R., Sain, S.: Spatial model fitting for large datasets with applications to climate and microarray problems. *Stat. Comput.* **19**, 113–128 (2009)
20. Genton, M.G.: Separable approximations of space-time covariance matrices. *Environmetrics. Special Issue for METMA3* **18**, 681–695 (2007)
21. Gneiting, T., Genton, M.G., Guttorp, P.: Geostatistical space-time models, stationarity, separability and full symmetry. In Finkenstaedt, B., Held, L. and Isham, V. (eds.), *Statistics of Spatio-Temporal Systems*, Chapman & Hall/CRC Press, 151–175 (2007)
22. Guan, Y., Sherman, M., Calvin, J.A.: A nonparametric test for spatial isotropy using subsampling. *J. Amer. Statist. Assoc.* **99**, 810–821 (2004)
23. Haas, T.C.: Lognormal and moving window methods of estimating acid deposition. *J. Amer. Statist. Assoc.* **85**, 950–963 (1990)
24. Hartman, L., Hössjer, O.: Fast kriging of large data sets with Gaussian Markov random fields. *Comput. Statist. Data Anal.* **52**, 2331–2349 (2008)
25. Hrafnkelsson, B., Cressie, N.: Hierarchical modeling of count data with application to nuclear fall-out. *Environ. Ecol. Stat.* **10**, 179–200 (2003)
26. Jun, M., Genton, M.G.: A test for stationarity of spatio-temporal processes on planar and spherical domains. *Statist. Sinica*. In press (2012)
27. Jun, M., Stein, M.L.: An approach to producing space-time covariance functions on spheres. *Technometrics.* **49**, 468–479 (2007)
28. Jun, M., Stein, M.L.: Nonstationary covariance models for global data, *Ann. Appl. Stat.* **2**, 1271–1289 (2008)
29. Kang, E.L., Cressie, N., Shi, T.: Using temporal variability to improve spatial mapping with application to satellite data. *Canad. J. Statist.* **38**, 271–289 (2010)
30. Kaufman, C., Schervish, M., Nychka, D.: Covariance tapering for likelihood-based estimation in large spatial datasets. *J. Amer. Statist. Assoc.* **103**, 1556–1569 (2008)
31. Li, B., Genton, M.G., Sherman, M.: A nonparametric assessment of properties of space-time covariance functions. *J. Amer. Statist. Assoc.* **102**, 736–744 (2007)
32. Li, B., Genton, M.G., Sherman, M.: Testing the covariance structure of multivariate random fields. *Biometrika.* **95**, 813–829 (2008)
33. Lindgren, F., Rue, H., Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach (with discussion). *J. R. Stat. Soc. Ser. B*, **73**, 423–498 (2011)
34. Matsuda, Y., Yajima, Y.: Fourier analysis of irregularly spaced data on \mathbb{R}^d . *J. R. Stat. Soc. Ser. B* **71**, 191–217 (2009)
35. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. R. Stat. Soc. Ser. B* **71**, 319–392 (2009)
36. Rue, H., Tjelmeland, H.: Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.* **29**, 31–50 (2002)
37. Sang, H., Huang, J.Z.: A full-scale approximation of covariance functions for large spatial data sets. *J. R. Stat. Soc. Ser. B*, in press (2011)
38. Sang, H., Jun, M., Huang, J.Z.: Covariance approximation for large multivariate spatial datasets with an application to multiple climate model errors. *Ann. Appl. Stat.*, in press (2011)
39. Shaby, B., Ruppert, D.: Tapered covariance: Bayesian estimation and asymptotics. *J. Comput. Graph. Statist.*, in press (2011)
40. Shao, X., Li, B.: A tuning parameter free test for properties of space-time covariance functions. *J. Statist. Plann. Inference.* **139**, 4031–4038 (2009)
41. Shi, T., Cressie, N.: Global statistical analysis of MISR aerosol data: a massive data product from NASA's Terra satellite. *Environmetrics.* **18**, 665–680 (2007)
42. Song, H., Fuentes, M., Gosh, S.: A comparative study of Gaussian geostatistical models and Gaussian Markov random field models. *J. Multivariate Anal.* **99**, 1681–1697 (2008)

43. Stein, M.L.: A modeling approach for large spatial datasets. *J. Korean Statist. Soc.* **37**, 3–10 (2008)
44. Stein, M.L., Chi, Z., Welty, L.J.: Approximating likelihoods for large spatial datasets. *J. R. Stat. Soc. Ser. B* **66**, 275–296 (2004)
45. Varin, C., Reid, N., Firth, D.: An overview on composite likelihood methods. *Statist. Sinica*. **21**, 5–42 (2011)
46. Vecchia, A.V.: Estimation and model identification for continuous spatial process. *J. R. Stat. Soc. Ser. B* **50**, 297–312 (1998)
47. Whittle, P.: On stationary processes in the plane. *Biometrika*. **41**, 434–449 (1954)
48. Zhang, H., Wang, Y.: Kriging and cross-validation for massive spatial data. *Environmetrics*. **21**, 290–304 (2010)
49. Zhu, Z., Wu, Y.: Estimation and prediction of a class of convolution-based spatial nonstationary models for large spatial data. *J. Comput. Graph. Statist.* **19**, 74–95 (2010)
50. Zimmerman, D.: Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models. *J. Statist. Comput. Simul.* **32**, 1–15 (1989)

AUTHOR QUERY

AQ1. Please note that the first author has been consider as corresponding author.
Please check.