Estimating Spatial Covariance using Penalized Likelihood with Weighted L_1 Penalty

Zhengyuan Zhu and Yufeng Liu*

July 5, 2007

Abstract

In spatial statistics, estimation of large covariance matrices are of great importance because of their role in spatial prediction and design. The traditional approaches typically assume that the spatial process is stationary, the covariance function takes some well known parametric form, and estimates the parameters of the covariance functions using likelihood based methods. In this paper we propose a nonparametric approach to estimate the covariance matrix for spatial nonstationary Gaussian Markov random field models. By exploiting the sparsity structure in the inverse covariance matrix, we show that a LASSO-type of approach gives improved covariance estimators measured by several criteria. Both our simulated examples and an application to the rainfall dataset show that the proposed method performs competitively.

Key words: Cholesky decomposition; LASSO; maximum likelihood; nonstationarity; Gaussian Markov random fields

*Zhengyuan Zhu and Yufeng Liu are both Assistant Professors, Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599 (E-mails: zhuz/yfliu@email.unc.edu). This research is partially supported by NSF DMS grants 0605434 and 0606577.

1 Introduction

The Gaussian Markov random field (GMRF) models has been used extensively in spatial statistics for data observed on grids. See for example Cressie (1993); Banerjee et al. (2004), and the references therein. Typical application areas include epidemiology, environmetrics, ecology, econometrics, and social science. Most of the analysis of spatial data using GMRF are based on strict stationarity assumptions which are necessary for one to use standard estimation methods. However, the stationarity assumption rarely holds in practice. Modelling non-stationarity is important when the objective of a study is to gain some insight about the dependence structure across the grid and how they vary in space. It can also improve the quality of spatial prediction when serious deviation from the stationarity assumption is presented in the data. While the mean nonstationarity can often be handled using detrending methods, it is more difficult to model the covariance nonstationarity, and few work in the statistics literature has addressed this problem for spatial GMRF. This is in sharp contrast with the amount of literature on spatial non-stationary models for general Gaussian random fields (GRF) (see, for example, Sampson and Guttorp (1992); Higdon et al. (1999); Fuentes and Smith (2001); Paciorek and Schervish (2006); Stein (2005); Pintore and Holmes (2003)).

In this paper we propose a nonparametric penalized likelihood method to identify spatial GMRF model through estimation of the precision matrix when multiple realizations of the GMRF are available. Our method does not need any stationarity assumption in space. By allowing for irregular zero patterns in the precision matrix, our method is effective for modelling the variation in neighborhood size and strength of correlation across the space.

Estimation of the precision matrix using penalized likelihood with different penalties has been studied in the context of longitudinal data analysis (Huang et al., 2006;

Levina and Zhu, 2006) and graphical models (Meinshusen and Buhlmann, 2006; Yuan and Lin, 2007). The estimation of precision matrix for spatial data is more difficult than that for time series data because there is no natural ordering of the observations, while it is easier than estimating the precision matrix in graphical models since one can use the location information of the spatial observations. The basic idea of our approach is as follows. First, we make use of the spatial location information to order the observations so that the corresponding precision matrix can be estimated optimally. Once the ordering is determined, we reparameterize the precision matrix using the modified Cholesky decomposition which helps to remove the positive definite constraint of the precision matrix (Huang et al., 2006; Levina and Zhu, 2006), and then use the L_1 penalized likelihood method to estimate elements in the decomposed triangular matrix.

Ordering the observations is an important step for our estimation procedure of the precision matrix since the likelihood procedure is not permutation invariant. In this research, we propose an ordering strategy which uses the location information to minimize the bandwidth of the decomposed triangular matrix of the precision matrix for certain spatial autoregressive models. A decomposed triangular matrix with a smaller bandwidth contains more zero elements. As a result, the corresponding L_1 penalized likelihood procedure can be more effective since the L_1 penalty encourages shrinkage of some estimators to be exactly zero (Tibshirani, 1996; Fan and Li, 2001).

The standard L_1 penalty uses the same weights for different parameters in the penalty term, which may be too restrictive. Intuitively, different parameters should be penalized differently according to their relative magnitude. Recent work on adaptive LASSO suggests the advantage of weighted L_1 penalty over the standard L_1 penalty using proper weights (Zou, 2006; Zhang and Lu, 2007). Using the spatial information of the observations, we consider distance-based weights for the L_1 penalty.

The effectiveness of the proposed procedure has been demonstrated using both simulated and real examples. Numerical results suggest that our ordering strategy and the distance-based weights for the L_1 penalty improve the efficiency of the estimation procedure. Asymptotic results are obtained, which are also applicable for the longitudinal setting. When the weights are properly chosen, the oracle properties for our estimator can be established.

The remaining sections are organized as follows. In Section 2, we describe the penalized likelihood method based on the modified Cholesky decomposition for estimating the precision matrix. Section 3 discusses the ordering issue and our algorithm for ordering the spatial points. Section 4 gives the details of the algorithm for estimating the precision matrix, and the asymptotic results are presented in Section 5. Simulation studies are carried out in Section 6 to illustrate our method, and the procedure is applied to a rainfall dataset in Section 7. We conclude the paper with some discussion in Section 8.

2 Methodology

Let $\{Z(s_i,t) : i = 1,\ldots,n;\ t = 1,\ldots,T\}$ be T copies of a non-stationary spatial GMRF observed at location $S = \{s_1, s_2, \ldots, s_n\}$, $\mathbf{Z}_i = (Z(s_i,1),\ldots,Z(s_i,T))'$ and $\mathbf{Z}(t) = (Z(s_1,t),\ldots,Z(s_n,t))'$. Typical data examples include disease counts, mortality rates, air pollutants, and meteorological variables observed in both space and time. In this paper we assume that $\mathbf{Z}(t) \sim i.i.d.N(\mu, \Sigma_S)$, i.e., no temporal correlation, and our objective is to estimate the spatial covariance matrix Σ_S or its inverse using $Z(s_i,t)$. Since our focus is on estimating the covariance matrix, we assume from now on that the dataset Z has been appropriately detrended so that $\mu = 0$. We will also use the notation Σ instead of Σ_S for the spatial covariance matrix when no confusion could arise.

If there is temporal dependence between $Z(s_i, t)$ and $Z(s_i, t+k)$ but no spatial dependence between $\mathbf{Z}_i$ and $\mathbf{Z}_j$, such that $\mathbf{Z}_i \sim i.i.d.N(\mu, \Sigma)$, then the data has the same structure as the longitudinal data. Pourahmadi (1999) studied estimating the temporal covariance Σ_T using the modified Cholesky decomposition $\Sigma_T^{-1} = LDL'$, where $L = \{l_{ij}\}$ is the unit lower triangular matrix and D is a diagonal matrix with diagonal elements d_t. It can be shown that estimating L and D is the same as fitting the following autoregressive time-series model

$$Z(s_i, t) = \sum_{j=1}^{t-1} \phi_{tj} Z(s_i, j) + \epsilon_{i,t},$$

with $\phi_{tj} = -l_{tj}$ and $\mathrm{Var}(\epsilon_{i,t}) = d_t$. Since the modified Cholesky decomposition is uniquely defined for any positive definite (P.D.) covariance matrix Σ, the problem of modelling Σ under the P.D. constraint is translated to an easier problem of modelling the autoregressive parameters l_{tj} and d_t, with $d_t > 0$; $t = 1, \ldots, T$, as the only constraints. Huang et al. (2006) proposed a nonparametric method to model L and D and estimate them using penalized likelihood methods.

A spatial analogue of the autoregressive time-series model is the simultaneously specified spatial autoregressive model (Whittle, 1954; Ripley, 1981)

$$\mathbf{Z} = B\mathbf{Z} + \boldsymbol{\epsilon}, \tag{1}$$

where $B = \{b_{ij}\}$ is a matrix describing the spatial dependence with $b_{ii} = 0$ and $(I - B)$ nonsingular, and $\boldsymbol{\epsilon} \sim N(0, D)$, where D is a diagonal matrix with diagonal elements d_i. Matrices B and D are related to Σ by $\Sigma^{-1} = (I - B)'D^{-1}(I - B)$. Note that B is in general not identifiable as the above decomposition of Σ^{-1} is not unique. One specific spatial autoregressive model that has received considerable attention in econometrics literature assumes $B = \rho W$ with W a given matrix (Ord, 1975; Smirnov and Anselin, 2001; Lee, 2004).

In this paper we assume that $\mathbf{Z}$ forms a GMRF. Since the zero elements in Σ^{-1}

indicate conditional independence (Speed and Kiiveri, 1986), the corresponding Σ^{-1} for $\mathbf{Z}$ will be a sparse matrix. We use model (1) to model Σ^{-1}, with the constraint that B is a lower triangular matrix, i.e., $b_{ij} = 0, \forall i \leq j$. As noted before, this constraint is not restrictive and makes B uniquely identifiable due to the fact that there is a unique modified Cholesky decomposition $\Sigma = LDL'$ for any positive-definite covariance matrix Σ. This constraint introduces an ordering in that the observations are assumed to be dependent on the observations entered before them. Unlike the longitudinal data considered in Pourahmadi (1999) and Huang et al. (2006), there is no natural ordering for spatial data, and one can not interpret the elements in L the same way as in the longitudinal case. We merely use L as a reparameterization for getting efficient estimators of Σ, and make inference about conditional independence structure only through Σ^{-1}. When Σ^{-1} is sparse, different orderings may affect performance of the procedure, and some are better than the others for estimating Σ^{-1}. We discuss the ordering issue in detail in Section 3.

Under model (1), the log likelihood function of $\mathbf{Z}$ is given by

$$\begin{aligned} l(B, D; \mathbf{Z}) &= \tfrac{T}{2} \log |(I - B')D^{-1}(I - B)| - \frac{1}{2} \sum_{t=1}^{T} \mathbf{Z}(t)'(I - B')D^{-1}(I - B)\mathbf{Z}(t) \\ &= -\tfrac{T}{2} \sum_{i=1}^{n} \log d_i - \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{n} \frac{1}{d_i} (Z(s_i, t) - \sum_{j<i} b_{ij} Z(s_j, t))^2. \end{aligned} \tag{2}$$

A direct maximization of (2) over B and D leads to an estimated Σ that is equivalent to the commonly used sample covariance matrix, which is known to be unstable for large covariance matrix. To explore the spatial structure in the data, we propose to estimate L and D by minimizing the LASSO type penalized negative likelihood function

$$-l(B, D; \mathbf{Z}) + \lambda \sum_{j<i} w_{ij} |b_{ij}|, \tag{3}$$

where λ is the tuning parameter to be determined by certain model selection tech-

niques (see Section 4), and $w_{ij} \geq 0$ is the weight for the L_1 penalty. The L_1 penalty is used to force the small elements in the estimated B to shrink to zero, which will lead to a sparse estimated precision matrix of Σ^{-1}. The weight w_{ij} can be chosen to be inversely proportional to the distance between s_i and s_j, or derived from the sample covariance matrix to achieve the oracle property (see Section 5).

3 Ordering

Unlike time series/longitudinal data, there is no nature ordering for spatial data. Since our procedure is not permutation invariant, the ordering of the data could have a significant effect on the estimation. To exploit the location information, we would like to keep points close in space also close in our ordering, and maintain the banding structure in L. As we use the LASSO type penalty for parameter estimation, it is natural to expect our procedure to work better if there are more zeros in the L matrix so that the sparsity effect of the L_1 penalty will be more useful. In this section we propose several methods for finding good orderings of the data which can reduce the non-zero elements in the L matrix of the Cholesky decomposition while keeping the banded structure.

We begin by introducing some graph-theoretic notations and definitions. We denote an undirected graph by $G = (X, E)$, where X is the set of vertices of size n, and E is the set of edges, which are unordered pairs of vertices. For any $Y \subset X$, we denote the *adjacent* set of Y by $Adj(Y) = \{x \in X \backslash Y : \{x, y\} \in E \text{ for some } y \in Y\}$. When Y consists of a single vertex $\{y\}$, we will write $adj(y)$ and refer to it as the neighbor of vertex y. The degree of a vertex x in a set Y is defined as the number $|adj(x) \cap Y|$. An ordering α of G is a mapping of $\{1, 2, \ldots, n\}$ onto X. A graph G with order α will be denoted as $G^\alpha = (X^\alpha, E)$, with x_i^α as the ith element in X^α.

For any $n \times n$ symmetric matrix C with c_{ij} as its (i, j)th element, we can define

an ordered graph $G^\alpha = (X^\alpha, E)$ to represent its zero patterns, with $X^\alpha = \{x_i^\alpha, i = 1, 2, \ldots, n\}$, and $\{x_i^\alpha, x_j^\alpha\} \in E$ if and only if $c_{ij} = c_{ji} \neq 0$. The bandwidth of a matrix C can be defined as $\beta(C) = max\{|i-j| : a_{ij} \neq 0\}$, and the bandwidth of an ordered graph $\beta(G^\alpha)$ is the bandwidth of the matrix corresponding to G^α. We use $\beta(G)$ to denote the minimum bandwidth of $\beta(G^\alpha)$ among all ordering α.

Let C be the precision matrix Σ^{-1} of a spatial process Z observed at $S = (s_1, s_2, \ldots, s_n)$, and G^α the associated ordered graph describing the zero patterns of Σ^{-1}. The corresponding unordered graph G describes the conditional independence structure of $Z(s)$, with $Z(s_i)$ and $Z(s_j), (s_i, s_j) \notin E$ conditionally independent given $adj(s_i)$. A change of ordering in S corresponds to a permutation of the rows and columns in Σ^{-1} and a different ordering of G. The problem of finding good permutations of a sparse symmetric matrix to reduce extra non-zero elements (i.e., fill-in) in its Cholesky decomposition has been studied extensively in the numerical analysis literature for the purpose of minimizing storage and speeding up computation. See for example, George and Liu (1981); Duff et al. (1989). Our problem is similar to theirs, but we need to determine the ordering before knowing the structure of the sparse matrix Σ^{-1}, which we try to estimate. Moreover, our ordering also needs to preserve the distance information. Nevertheless these studies provide useful tools for us to find good orderings once we have a good preliminary estimator of Σ^{-1}.

To find a good ordering, we first define a natural spatial neighborhood structure based on the Vornoi tessellation (Green and Sibson, 1978). Let $S = (s_1, s_2, \ldots, s_n)$ be the set of locations where the data are collected, R be the region of interest, and $s_i \in R$ for all i. The Vornoi tessellation divides R into n subregions $R_i = \{s \in R : ||s - s_i|| \leq ||s - s_j||, \forall j \neq i\}$. We define the neighbors of s_i as all s_j's such that $R_i \cap R_j \neq \emptyset$. This neighborhood structure can be represented by an undirected graph $G = (X, E)$, with the set of vertices X the same as S, and the set of edges E

representing the neighborhood structure, with an edge between s_i and s_j if and only if they are neighbors.

Next we give an algorithm for ordering based on the natural spatial neighborhood structure G. It is similar to the reverse Cuthill-Mckee algorithm (Cuthill and McKee, 1969; George and Liu, 1981) used in matrix theory to reduce the bandwidth of sparse symmetric matrices, while we use the distance information in addition to the vertex degree to order the points.

The Ordering Algorithm:

1. Find s^* and s^{**} such that $||s^* - s^{**}|| = \max_{i,j} ||s_i - s_j||$.

2. Let s^* be the first element of an ordered set Q, repeat the following steps for $i = 1, 2, \ldots, n-1$:

   - Construct the set $A_i = Adj(q_i) \backslash Q$, the adjacency set of the ith element of Q excluding the vertices that are already in Q, where q_i denotes the i-th element of Q.
   - Sort elements in A_i first with ascending vertex degree in $G \backslash Q$. For elements with the same vertex degrees, sort with ascending distance to q_i.
   - Append A_i to the end of Q.

3. Repeat 2 with s^{**} as the initial element in Q to get a second ordered set Q'.

4. Select among Q, Q', and their reverse orderings the one that gives the least fill-in in the symbolic Cholesky decomposition of the symmetric matrix corresponding to G as the resulting ordering.

The intuition behind the ordering algorithm is that if the natural neighborhood structure is the same or close to the real neighborhood structure of the GMRF, then the ordering algorithm gives ordering which reduces the bandwidth of Σ^{-1}. This in

turn leads to more zeros in L. For one-dimensional spatial processes, the ordering obtained by the ordering algorithm is the same natural ordering used for time series data, which achieves the minimum bandwidth for Σ^{-1} and has the least fill-in for a large class of GMRF models, including $AR(k)$ processes.

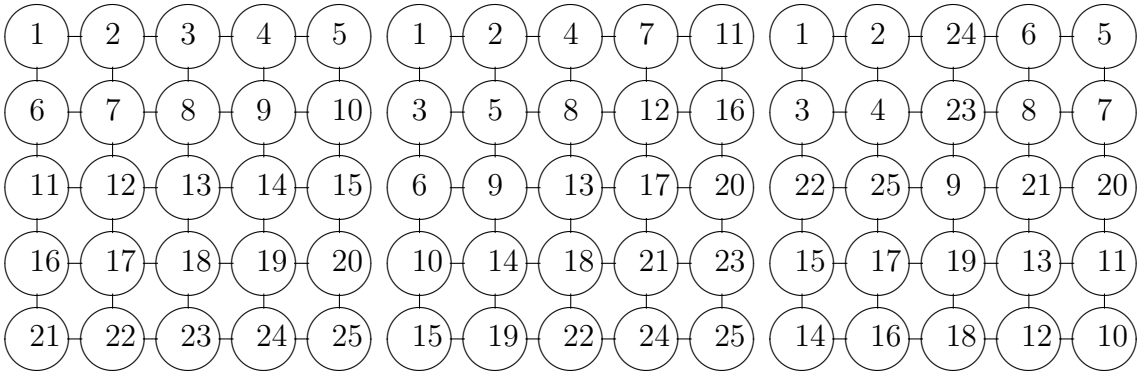

Figure 1: An example of the three orderings for 5×5 grids. Left: default ordering; Middle: natural ordering; Right: minimum degree ordering.

Let $G_{m,n} = (X_{m,n}, E_{m,n})$ be the graph of a regular $m \times n$ grids in $\mathbb{R}^2$ such that $X_{m,n} = \{(i,j) : i = 1,2,\ldots,m; j = 1,2,\ldots,n\}$ and $E_{m,n} = \{\{(i,j),(i,j+1)\},\{(i,j),(i,j-1)\},\{(i,j),(i,j+1)\},\{(i,j),(i,j+1)\} : i = 1,2,\ldots,m; j = 1,2,\ldots,n\}$. We define the default ordering α_d for such a grid as the one which orders (i,j) as the $(i \times n + j)$th element. It is easy to check that one of the eight equivalent orderings given by the ordering algorithm is $\{(1,1),(1,2),(2,1),(1,3),(2,2),(3,1),\ldots\}$, which we will refer to as the natural ordering α_n. Figure 1 (left and middle plots) show examples of the default ordering and the natural ordering for $m = n = 5$. It is easy to show that the bandwidths for the default and natural ordering of $G_{m,n}$ are m and $\min\{m,n\}$ respectively.

The following lemma and propositions reveal some properties of the default and natural ordering. The proofs of the propositions involve standard induction arguments and are omitted here.

Lemma 1 $\beta(G_{m,n}) = \min\{m, n\}$.

Proposition 1 *There are at most* $m^2n + mn - m^2 + m - 1$ *non-zero elements in the symbolic Cholesky decomposition of matrix corresponding to* $G^{\alpha_d}_{m,n}$.

Proposition 2 *For* $m \leq n$, *there are at most* $m^2n + mn - \frac{1}{3}m^3 + \frac{1}{2}m^2 - \frac{7}{6}m$ *non-zero elements in the symbolic Cholesky decomposition of matrix corresponding to* $G^{\alpha_n}_{m,n}$.

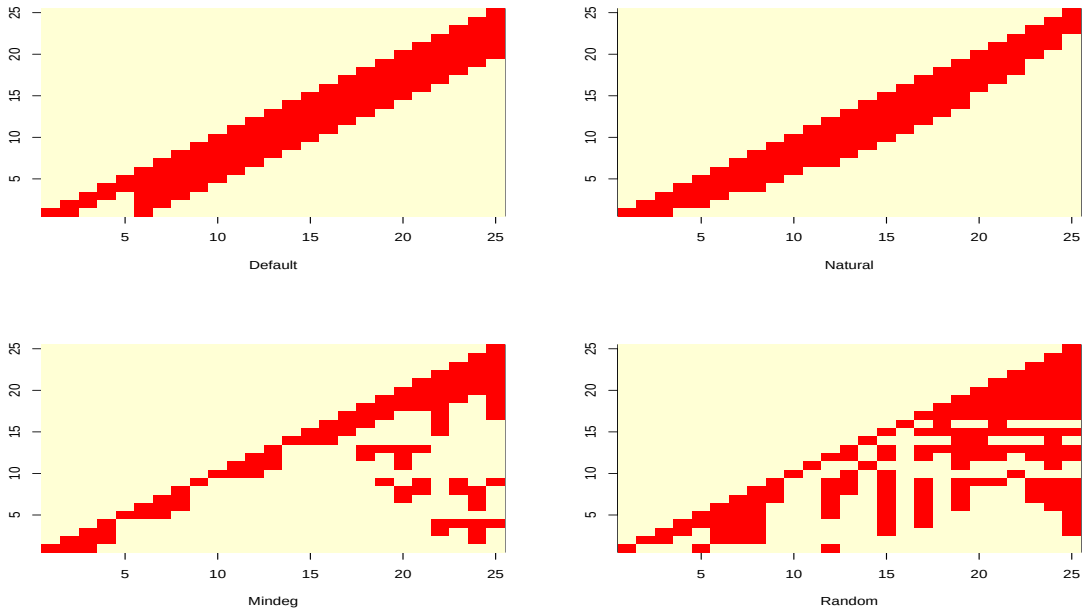

Figure 2: Illustration of the L matrix zero patterns of the stationary example in Section 6 for the four orderings: default, natural, mindeg, and random.

From Lemma 1 we know that the natural ordering achieves the minimum bandwidth. The default order also achieves the minimum bandwidth when $m \leq n$. The two propositions show that even though both orderings achieve the minimum bandwidth, the natural ordering has smaller number of non-zero elements for large m and n. For $m = n$ large, natural ordering has approximately 1/3 less non-zero elements. Numerical results show that both orderings give significantly smaller numbers of non-zero elements compared with random ordering. For example, for $G_{5,5}$, the L matrix

from natural ordering has 115 non-zero elements, while for default ordering there are 129 non-zero elements, and the average non-zero elements for random ordering is 215.

We note that the general problem of finding the optimal ordering which minimizes the bandwidth or the fill-in in the Cholesky decomposition is NP-complete (Papadimitriou, 1976; Yannakakis, 1981), and most available algorithms, such as various minimum degree and minimum local fill-in algorithms, are greedy algorithms based on heuristics. Although these algorithms are typically more efficient in reducing the fill-in than our algorithm, they do not use the location information, and usually result in orderings that put spatially close points far away. The right panel of Figure 1 gives an example of minimum degree ordering.

Figure 2 shows the L matrix zero patterns for the simulated stationary example in Section 6, for the four orderings: default, natural, minimum degree, and random. Random ordering generates significantly more nonzero elements in the L matrix. Although minimum degree ordering produces marginally smaller number of zeros in the L matrix compared to the default and natural ordering, it does not have the banded structure and leads to inferior estimates of Σ^{-1} with the distance based weighting (see more discussions on weighting in Sections 5 and 6). Thus we do not pursuit it further in this paper.

4 Computational Algorithm

4.1 The Optimization Routine

To implement our method, we need to minimize the following objective function with respect to (B, D):

$$\frac{T}{2}\sum_{i=1}^{n}\log d_i + \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{1}{d_i}(Z(s_i,t) - \sum_{j<i} b_{ij}Z(s_j,t))^2 + \lambda\sum_{j<i} w_{ij}|b_{ij}|. \qquad (4)$$

Note that for any give B, the minimizer of (4) can be obtained as

$$d_i = \frac{1}{T}\sum_{t=1}^{T}(Z(s_i,t) - \sum_{j<i} b_{ij}Z(s_j,t))^2. \tag{5}$$

For any give D, the function in (4) can be minimized by solving

$$\min_{\{b_{ij}\}} \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{1}{d_i}(Z(s_i,t) - \sum_{j<i} b_{ij}Z(s_j,t))^2 + \lambda\sum_{j<i} w_{ij}|b_{ij}|. \tag{6}$$

Notice that the first term in (6) is a quadratic function in b_{ij}. To deal with the absolute function in the second term of (6), we introduce slack variables $\xi_{ij} \geq 0$ and simplify problem (6) as

$$\min_{\{b_{ij}\}} \quad \frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{1}{d_i}(Z(s_i,t) - \sum_{j<i} b_{ij}Z(s_j,t))^2 + \lambda\sum_{j<i} w_{ij}\xi_{ij}, \tag{7}$$

$$\text{subject to} \quad \xi_{ij} \geq b_{ij}, \xi_{ij} \geq -b_{ij}; \forall j < i. \tag{8}$$

Problem (7) involves minimization of a quadratic function subject to linear constraints. Thus it is a quadratic programming (QP) problem and can be solved using many routine optimization softwares. In this article, all examples were computed using the commercial optimization software CPLEX with the AMPL interface (Fourer et al., 2003).

We minimize (4) via an iterative procedure. We first initialize B and D using the solution of the unpenalized maximum likelihood estimator, i.e., the Cholesky decomposition of the sample covariance matrix. Then we solve D in (5) and B in (6) iteratively until convergence.

4.2 The Tuning Procedure

The minimizer of (4) depends on the value of λ. The bigger λ is, the more shrinkage is performed on B. If $\lambda = 0$, it reduces to the regular maximum likelihood procedure. If $\lambda = \infty$, it implies $b_{ij} = 0$. In general, there is an optimal $\lambda \geq 0$ which gives the best balance between data fitting in terms of likelihood and shrinkage.

For a given λ, we denote the corresponding covariance estimator as $\hat{\Sigma}_\lambda$. In the ideal case, if we knew the true Σ, we can compare $\hat{\Sigma}_\lambda$ with Σ and find the λ whose corresponding estimator is the closest to Σ. There are many different measures to qualify difference of two matrices. In this paper, we use the commonly used entropy criterion

$$Ent(\Sigma, \hat{\Sigma}_\lambda) = tr(\Sigma^{-1}\hat{\Sigma}_\lambda) - \log(\Sigma^{-1}\hat{\Sigma}_\lambda) - n. \tag{9}$$

Clearly, Σ is not available for evaluation and we need to utilize model selection techniques such as AIC, BIC, and validation/cross validation. To use validation, for a given λ, we evaluate the corresponding covariance estimator $\hat{\Sigma}_\lambda$ on a separate dataset. Suppose we have a training set and a validation set. Once we get $\hat{\Sigma}_\lambda$ using the training set, we can compare it with the estimator $\hat{\Sigma}_v$ (such as the sample covariance matrix) from the validation set and find the λ so that $\hat{\Sigma}_\lambda$ and $\hat{\Sigma}_v$ are the closest. In this way, we hope to get $\hat{\Sigma}_\lambda$ that can generalize well for new data. For cases where a validation set is not available such as many real applications, we propose to use K-fold cross validation (CV) with $K = 5$ or 10.

One can also use other existing model selection methods including AIC and BIC to choose λ. More discussions can be found in Section 6.

5 Asymptotic Properties

In this section, we study the asymptotic behavior of our proposed estimators. In particular, we derive asymptotic theories analogous to that of Knight and Fu (2000) in the least squares regression setting. This type of asymptotic theories was also studied in Yuan and Lin (2007); Zou (2006).

Consider the following objective function for minimization:

$$L_T(B, D; \mathbf{Z}) = \log|D| + \frac{1}{T}\sum_{t=1}^{T} \mathbf{Z}(t)'(I - B')D^{-1}(I - B)\mathbf{Z}(t) + \lambda \sum_{j<i} w_{ij}|b_{ij}|. \tag{10}$$

Note that $\frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}(t)'(I-B')D^{-1}(I-B)\mathbf{Z}(t) = tr((I-B')D^{-1}(I-B)\frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}(t)\mathbf{Z}(t)') = tr((I-B')D^{-1}(I-B)\bar{A})$, where $\bar{A} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{Z}(t)\mathbf{Z}(t)'$. Then (10) can be reduced to

$$L_T(B, D; \mathbf{Z}) = \log|D| + tr((I-B')D^{-1}(I-B)\bar{A}) + \lambda\sum_{j<i} w_{ij}|b_{ij}|, \tag{11}$$

where B is a lower triangular matrix and D is a diagonal matrix with positive diagonal elements. Denote $(\hat{B}, \hat{D})$, having the same form as that of (B, D), as the minimizer of (11).

The following theorem characterizes the asymptotic behavior of the estimator $(\hat{B}, \hat{D})$ with equal weights $w_{ij} = 1$.

Theorem 1 *If $\sqrt{T}\lambda \to \lambda_0 \geq 0$ as $T \to \infty$ and $w_{ij} = 1$ for $j < i$, the estimator $(\hat{B}, \hat{D})$ of (10) satisfies that*

$$(\sqrt{T}(\hat{B} - B), \sqrt{T}(\hat{D} - D)) \to_d argmin_{\{U_B, U_D\}} V(U_B, U_D), \tag{12}$$

where U_B is a lower triangular matrix and U_D is a diagonal matrix with positive diagonal elements, and $V(U_B, U_D) =$

$$-tr(U_D D^{-1} U_D D^{-1}) + tr(UW) + \lambda_0 \sum_{j<i}(U_B(i,j)sign(b_{ij})I(b_{ij} \neq 0) + |U_B(i,j)|I(b_{ij} = 0)),$$

with $U = -(I-B)' D^{-1} U_D D^{-1}(I-B) - U_B' D^{-1}(I-B) - (I-B)' D^{-1} U_B$, W a random symmetric $n \times n$ matrix satisfying that $vec(W) \sim N(0, \Lambda)$, and $\Lambda = Cov(vec(W))$ such that $Cov(w_{ij}, w_{kl}) = Cov(Z(s_i)Z(s_j), Z(s_k)Z(s_l))$.

We now consider the asymptotic behavior of weighted L_1 penalty. Consider $w_{ij} = 1/\tilde{b}_{ij}$, where $\tilde{b}_{ij}$ satisfies $\sqrt{T}(\tilde{b}_{ij} - b_{ij}) \to 0$ in probability.

Theorem 2 *Denote $(\hat{B}, \hat{D})$ as the minimizer of (11). Assume $\sqrt{T}\lambda \to 0$ and $T\lambda \to \infty$ as $T \to \infty$ and $w_{ij} = 1/\tilde{b}_{ij}$ for $j < i$ with $\sqrt{T}(\tilde{b}_{ij} - b_{ij}) \to 0$ in probability, then we can conclude that*

(1). $P(\hat{b}_{ij} = 0) \to 1$ *if* $b_{ij} = 0$*;*

(2). $\hat{D}$ *and the nonzero* $\hat{b}_{ij}$*'s have the same asymptotic distribution as the maximum likelihood estimators.*

Theorem 2 implies that if the weights in the penalty term are properly chosen, the minimizers have the so called oracle property (Fan and Li, 2001). Specifically, $\hat{b}_{ij} = 0$ is zero asymptotically with probability one if the true b_{ij} is zero. For other parameters D and nonzero b_{ij}, they can be estimated as well as the usual maximum likelihood estimators. Therefore, asymptotically, the weighted penalty helps to recover the sparsity pattern in B without sacrificing estimation accuracy of other non-zero parameters.

6 Simulations

In this section, we use simulation to evaluate the performance of our shrinkage estimator of the precision matrix. Four orderings and two weightings for the shrinkage estimator are considered. They are compared with the maximum likelihood estimator of the precision matrix, as well as estimators based on the parametric spatial model. We use the entropy criterion in (9) to measure how close the true and the estimated precision matrices are.

Two dependence structures are considered in the simulation. For the first example we assume a stationary Gaussian Markov random field on a 5×5 grid, with each observation (not on the boundary) conditionally dependent on only the four nearest neighbors. For the second example, we consider a non-stationary GMRF on a 5×5 grid, with the size of the neighborhood varying spatially. The zero patterns of the two precision matrices are given in Figure 3.

We first compare the performance of validation with AIC and BIC for selecting the tuning parameter λ. Figure 4 shows the validation approach selects a λ which

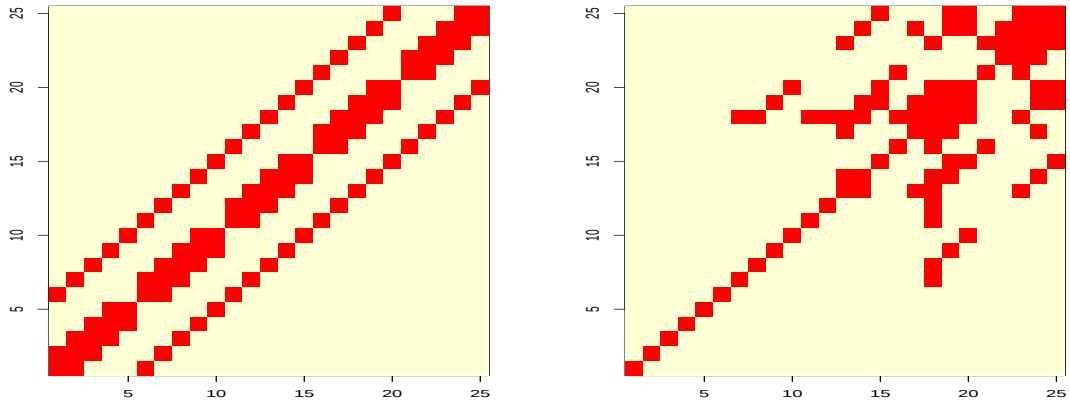

Figure 3: Plots of zero patterns of the two precision matrices corresponding to the stationary and nonstationary simulated examples in Section 6.

is very close to the optimal λ. AIC selects a slightly smaller λ than the true one. In contrast, BIC selects a larger λ, resulting in a sparser estimated covariance matrix. This is consistent with the common wisdom that BIC tends to choose more parsimonious models (Hastie et al., 2001). Since validation works the best here for selecting tuning parameters, we used separate validation datasets to choose λ for our simulation studies.

Next we simulate data under the stationary GMRF model, and compare the performance of penalized likelihood (PLIK) approach with that of sample covariance (SC) and model-based geostatistical approach assuming a stationary exponential (EXP) model. For comparison, we use both the entropy criterion as well as the percentage of correctly identified zeros and non-zeros in the precision matrix. For the PLIK approach, we compare the performance of four orderings (default, natural, minimum degree, and random) and two weightings (constant and distance weighting). Sample sizes 50 and 100 are considered. The results are summarized in Table 1 and Table 2, and we have the following observations:

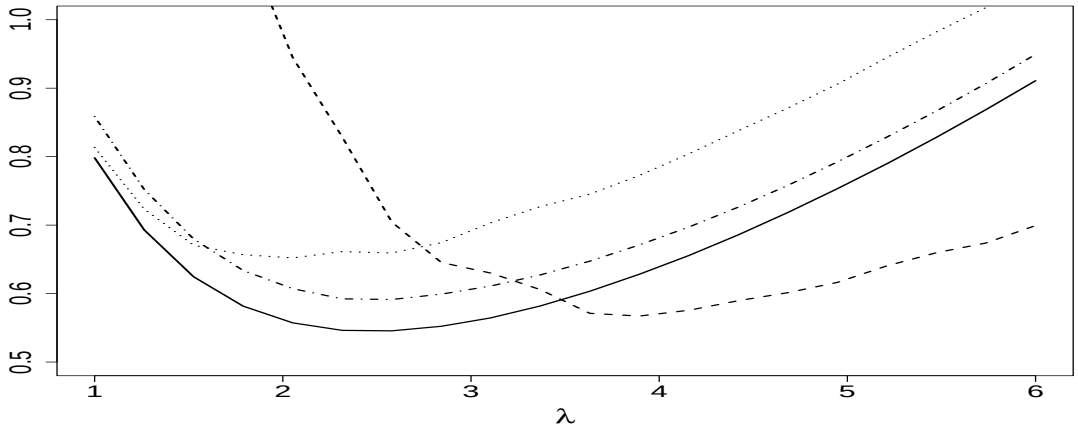

Figure 4: Comparison of different methods for selecting the tuning parameter λ. Solid line: entropy with true Σ; dotted line: AIC; dashed line: BIC; dash-dotted line: entropy with $\hat{\Sigma}$ from the validating data.

1. In terms of entropy criterion, the PLIK approach gives covariance estimators which are much better than the SC. The model-based geostatistical approach gives overall the best results. This is not surprising as the data are simulated from a stationary model. It is worth noting that both SC and the model-based geostatistical approach can not correctly identify any zeros in the precision matrix.

2. The distance based weighting is better than the constant weighting by a large margin. The same pattern also holds for simulations under the nonstationary model (not reported in the paper). For spatial problems we recommend the use of distance based weighting unless other reliable estimators are available to derive the weights.

3. For distance based weighting, the default and natural orderings give better results than the random or minimum degree orderings in terms of both the entropy criterion and the percentage of correct zero/non-zero entries. There

		PLIK, ordering					
		Default	Natural	Mindeg	Random	SC	EXP
n=50	CW	3.33(.07)	3.34(.07)	3.41(.08)	3.44(.07)	19.35 (.45)	0.10 (.01)
	DW	2.24(.06)	2.23(.06)	2.35(.06)	2.35(.06)	-	-
n=100	CW	1.72(.04)	1.72(.04)	1.76(.04)	1.79(.04)	5.21 (.09)	0.08 (.00)
	DW	1.10(.03)	1.15(.03)	1.25(.03)	1.21(.03)	-	-

Table 1: Entropy measure for different estimation methods. Results are based on 100 replications simulated from the stationary GMRF model. Default, natural, mindeg, and random represent PLIK estimators based on the four corresponding orderings described in Section 3. CW and DW represent constant weighting and distance weighting respectively. The numbers in parentheses are the standard errors.

		Default	Natural	Mindeg	Random
n=50	CZ	0.82(.01)	0.81(.00)	0.70(.01)	0.73(.01)
	CN	0.74(.00)	0.53(.01)	0.45(.01)	0.46(.00)
n=100	CZ	0.77(.01)	0.79(.00)	0.49(.01)	0.62(.01)
	CN	0.77(.00)	0.56(.00)	0.58(.00)	0.52(.00)

Table 2: Comparison of zero patterns for the four orderings. Results are based on 100 replications simulated from the stationary GMRF model. Default, natural, mindeg, and random represent PLIK estimators based on the four corresponding orderings described in Section 3. CZ and CN represent percentages of correctly identified zeros and non-zeros in the precision matrix respectively by each method. The numbers in parentheses are the standard errors.

is no significant difference between default and natural orderings in terms of the entropy criterion and correct zeros, while the default ordering gives higher percentage of correct non-zeros. For regularly spaced data we recommend the default ordering when using the PLIK approach. When the data are observed from irregularly spaced locations and no default ordering is available, the natural ordering is recommended.

Last we present simulation results for the non-stationary GMRF model in Table 3. Sample sizes 50, 100, and 200 are considered, and we again compare the proposed PLIK method with the SC and the model-based geostatistical approach (under the stationarity assumption) using the entropy criterion. The percentage of correct

		PLIK	SC	EXP
n=50	Entropy	1.47(0.04)	18.98(0.36)	3.26 (0.01)
	CZ	0.93(0.00)	0.00(0.00)	0.00 (0.00)
	CN	0.60(0.00)	1.00(0.00)	1.00 (0.00)
n=100	Entropy	0.84(0.01)	5.17(0.06)	3.22 (0.00)
	CZ	0.85(0.00)	0.00(0.00)	0.00 (0.00)
	CN	0.72(0.00)	1.00(0.00)	1.00 (0.00)
n=200	Entropy	0.46(0.01)	2.06(0.02)	3.21 (0.00)
	CZ	0.80(0.00)	0.00(0.00)	0.00 (0.00)
	CN	0.77(0.00)	1.00(0.00)	1.00 (0.00)

Table 3: Simulation results for non-stationary GMRF. Results are based on 100 replications simulated from the non-stationary GMRF model. PLIK represents the penalized likelihood estimator based on the default ordering. CZ and CN represent percentages of correctly identified zeros and non-zeros in the precision matrix respectively by each method. The numbers in parentheses are the standard errors.

zero/nonzero entries are also reported. For all sample sizes considered, the entropy criterion for our PLIK approach are much smaller than that of the other two approaches. The PLIK approach also correctly identifies most zero elements in the precision matrix, which is useful for identifying the neighborhood structure for the GMRF model. The comparison between different orderings and weightings for the PLIK approach is omitted, since it is similar to that of the stationary case.

7 Real Application

In this section we apply our methodology to an annual rainfall dataset in the eastern US. We first estimate the inverse covariance matrix and examine the estimated conditional independence structure. Next we compare the prediction performance using our estimated covariance matrix with several other methods to demonstrate the advantage of our method.

We chose the study region to be between latitude 38.5 and 41.5 and longitude -84.8 and -80.5, which includes 36 stations and cover the state of Ohio and some

surrounding area. The time period of the data is between 1960 and 1999. Detailed documentation of the complete dataset can be found in Groisman (2000). For each station data we subtracted the mean over the 40 years and model the difference as a Gaussian random process. A look at the normal quantile plots shows that the data are mostly consistent with the Gaussian assumption, with less than 0.5% possible outliers. We use robust measures to compare different prediction methods, which reduces the influence of the outliers. An examination of the autocorrelation and cross correlation of the observations at different stations reveals no significant time dependence, thus we model the observations from each year as independent realizations of the same process with a different mean. We further assume that the annual rainfall at any stations are conditionally independent with annual rainfall at other stations conditional on the rainfall at neighboring stations, i.e., the precision matrix Σ^{-1} is sparse. We estimate Σ^{-1} using the method described in Section 2, the zero pattern of which gives an estimated neighborhood structure. Both BIC criterion and cross-validation indicate that the best λ is between 15 and 30, and we use $\lambda = 18.3$ for estimating Σ, which minimizes the BIC criterion.

Figure 5 shows three examples of estimated neighborhood structures. It is worth noting that the neighborhood structure is neither homogenous across space nor isotropic. An independent estimate of the variance at each location also indicates strong nonstationarity in space. Thus a stationary spatial AR model would not fit the data well. The fact that most neighborhood structures are wider in the north-south direction can be partially explained by the local geology.

One important objective of estimating the spatial covariance matrix is for spatial prediction. Thus one can compare different estimates of covariance matrix by comparing their corresponding prediction performance. Here we use spatial prediction to compare three methods for estimating the covariance matrix: our PLIK estimator,

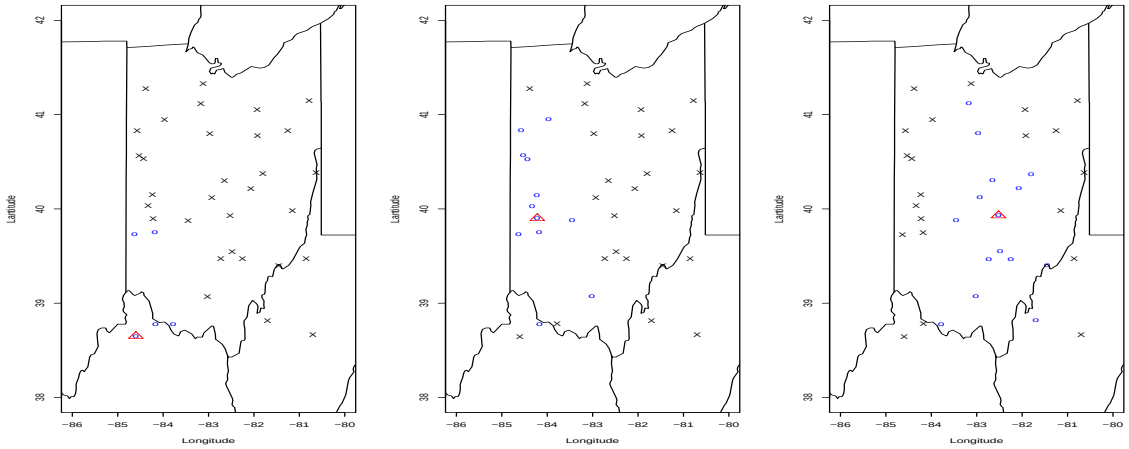

Figure 5: Examples of neighborhood structure. Conditional on observations at the blue dots, the observation at the red triangle is independent of the observations at the black crosses.

SC, and EXP. For each year, we estimate the covariance matrix using the rest of the data, and predict at each location using the rest of the data at that year and the estimated covariance matrix. Kriging is used as the prediction method. As a benchmark we also include the prediction performance of a model-free method using inverse distance weighted average (IDWA). The mean square prediction error (MSPE) is computed for each location across forty years, and the median MSPE across all locations are reported in Table 4. When we are estimating parameters for the stationary exponential model, it is such a bad fit for the data that the maximum likelihood method often runs into numerical problems, and we reported the result with the parameter which minimizes the median MSPE. Among all the methods, the prediction using covariance matrix estimated by our PLIK method is the best, and it is substantially better than the SC.

	PLIK	SC	EXP	IDWA
median MSPE	2.31	8.08	2.43	2.64

Table 4: Comparison of prediction performance

8 Discussion

In this paper, we propose a nonparametric approach to estimate covariance matrix for GMRF models. A LASSO type penalized likelihood method is considered. Since the method is not permutation invariant, several ordering schemes are considered using the spatial locations of observations to maximize the estimation accuracy of the method. In order to recover the structure of the decomposed triangular matrix, a weighted L_1 penalty with weights chosen using the location information of the observations has been proposed. Both theoretical and numerical studies show that the proposed method performs competitively.

In our algorithm, estimating B involves a QP problem. Although standard QP solvers can be used for solving moderate size problems, more efficient algorithms are needed to handle large-scale spatial problems. The recent LARS algorithm (Efron et al., 2004) may be extended here and deserves further exploration.

Our proposed method is developed under the assumption of GMRF models. Even when the spatial processes are not exactly GMRFs, it is natural to expect the conditional dependence between sites that are far away to be very small, and the LASSO-type estimator could be more efficient than the common likelihood estimator with a small sample size, and leads to better prediction as is demonstrated in the data example in Section 7. Rue and Tjelmeland (2002) showed empirically that GMRFs approximation to general Gaussian random fields can be surprisingly good. The zeros in the estimated Σ^{-1} can also be used to identify the neighborhood structure of the process and help build a more structured spatial covariance model.

The current setting assumes no temporal correlation and estimates the spatial covariance only. In some situations, modelling both temporal and spatial correlations can be beneficial. As future research work, we will generalize the proposed method for estimating temporal and spatial covariance matrices simultaneously.

Appendix

Proof of Lemma 1: WLOG, assume $m \leq n$. Since $\beta(G^{\alpha n}) = m$, we only need to show $\beta(G) \geq m$ for any α. This can be proved by contradiction. Since $\beta(G) \geq \max_k \min_{|Y|=k, Y \subset X} |\partial Y|$ (Chinn et al. (1982), Theorem 2), if $\beta(G) < m$, then for any k, there exists $Y \subset X$ s.t. $|Y| = k$, $|\partial Y| < m$. However, for $|\partial Y| = m' < m$, $|Y| \leq m'(m'-1)/2$ or $|Y| \geq nm - m'(m'-1)/2$. Thus for k between $m(m-1)/2+1$ and $mn - m(m-1)/2 - 1$ there is no Y such that $|Y| = k$ and $|\partial Y| < m$. Therefore $\beta(G) \geq m$.

Proof of Theorem 1: Define $V_T(U_B, U_D)$ by

$$L_T(B + U_B/\sqrt{T}, D + U_D/\sqrt{T}; \mathbf{Z}) - L_T(B, D; \mathbf{Z}) = E_1 + E_2 + E_3, \tag{13}$$

where $E_1 = \log|D + U_D/\sqrt{T}| - \log|D|$, $E_2 = tr((I - (B + U_B/\sqrt{T})')(D + U_D/\sqrt{T})^{-1}(I - (B + U_B/\sqrt{T}))\bar{A}) - tr((I - B')D^{-1}(I - B)\bar{A})$, and $E_3 = \lambda \sum_{j<i}(|b_{ij} + U_B(i,j)/\sqrt{T}| - |b_{ij}|)$. Then it is easy to see that $V_T(U_B, U_D)$ is minimized at $(\sqrt{T}(\hat{B} - B), \sqrt{T}(\hat{D} - D))$.

Note that $E_1 = \log|I + U_D D^{-1}/\sqrt{T}| = \frac{tr(U_D D^{-1})}{\sqrt{T}} - \frac{tr(U_D D^{-1} U_D D^{-1})}{T} + o(\frac{1}{T})$, where $tr(\cdot)$ denotes the trace of a matrix.

To simplify E_2, we first note that

$$\begin{aligned}(D + U_D/\sqrt{T})^{-1} &= D^{-1}(I + U_D D^{-1}/\sqrt{T})^{-1} \\ &= D^{-1} - D^{-1}U_D D^{-1}/\sqrt{T} + o(1/T).\end{aligned}$$

Consequently, we have

$$\begin{aligned}
& (I-(B+U_B/\sqrt{T})')(D+U_D/\sqrt{T})^{-1}(I-(B+U_B/\sqrt{T}) \\
= \; & ((I-B')-U_B'/\sqrt{T})(D^{-1}-D^{-1}U_DD^{-1}/\sqrt{T}+o(1/T))((I-B)-U_B/\sqrt{T}) \\
= \; & (I-B)'D^{-1}(I-B)+\frac{1}{\sqrt{T}}U+o(\frac{1}{T}),
\end{aligned}$$

where $U=-(I-B)'D^{-1}U_DD^{-1}(I-B)-U_B'D^{-1}(I-B)-(I-B)'D^{-1}U_B$. Thus, E_2 can be simplified as $\frac{1}{\sqrt{T}}tr(U\bar{A})+o(\frac{1}{T})=\frac{1}{\sqrt{T}}tr(U(\bar{A}-\Sigma))+\frac{1}{\sqrt{T}}tr(U\Sigma)+o(\frac{1}{T})$.

Since $\Sigma=(I-B)^{-1}D(I-B')^{-1}$, we have

$$tr(U\Sigma)=tr(-D^{-1}U_D)-tr(U_B'(I-B')^{-1})-tr(U_B(I-B)^{-1}).$$

Note that both B and U_B are lower triangular matrices, thus we have $tr(U_B'(I-B')^{-1})=tr(U_B(I-B)^{-1})=0$. This implies that $tr(U\Sigma)=-tr(U_DD^{-1})$.

For E_3, notice that the sign of $b_{ij}+U_B(i,j)/\sqrt{T}$ is determined by that of b_{ij} as T becomes sufficiently large. Consequently, E_3 can be simplified as

$$E_3=\frac{\lambda}{\sqrt{T}}\sum_{j<i}(U_B(i,j)\text{sign}(b_{ij})I(b_{ij}\neq 0)+|U_B(i,j)|I(b_{ij}=0)).$$

Using the simplifications of E_1, E_2, and E_3, $TV_T(U_B,U_D)$ can be simplified as

$$-tr(U_DD^{-1}U_DD^{-1})+tr(UW_T)+\sqrt{T}\lambda\sum_{j<i}(U_B(i,j)\text{sign}(b_{ij})I(b_{ij}\neq 0)+|U_B(i,j)|I(b_{ij}=0))+o(1),$$

where $W_T=\sqrt{T}(\bar{A}-\Sigma)\to N(0,\Lambda)$ as $T\to\infty$. Since $\sqrt{T}\lambda\to\lambda_0$, $TV_T(U_B,U_D)\to V(U_B,U_D)$ in distribution. The desired result then follows.

Proof of Theorem 2: The proof is similar to that of Theorem 1. Define $TV_T(U_B,U_D)$ by

$$T(L_T(B+U_B/\sqrt{T},D+U_D/\sqrt{T};\mathbf{Z})-L_T(B,D;\mathbf{Z}))=T(E_1+E_2+E_3), \quad (14)$$

where E_1, E_2 are defined as in the proof of Theorem 1 and $TE_3=T\lambda\sum_{j<i}\tilde{b}_{ij}^{-1}(|b_{ij}+U_B(i,j)/\sqrt{T}|-|b_{ij}|)$.

We only need to simplify TE_3. First consider the case that $b_{ij} \neq 0$. Since the sign of $b_{ij} + U_B(i,j)/\sqrt{T}$ is determined by that of b_{ij} as T becomes sufficiently large, then $\sqrt{T}(|b_{ij} + U_B(i,j)/\sqrt{T}| - |b_{ij}|) = U_B(i,j)\text{sign}(b_{ij})$. This, together with the fact that $\sqrt{T}\lambda \to 0$ and $\sqrt{T}(\tilde{b}_{ij} - b_{ij}) \to_P 0$ as $T \to \infty$, we have $TE_3 = \sqrt{T}\lambda \sum_{j<i} \tilde{b}_{ij}^{-1} U_B(i,j)\text{sign}(b_{ij}) \to_P 0$.

When $b_{ij} = 0$, $\sqrt{T}(|b_{ij} + U_B(i,j)/\sqrt{T}| - |b_{ij}|) = |U_B(i,j)|$ and $\sqrt{T}\tilde{b}_{ij} = O_P(1)$. Moreover, $T\lambda \to \infty$ as $T \to \infty$. Thus, $TE_3 = T\lambda \sum_{j<i} \frac{1}{\sqrt{T}\tilde{b}_{ij}} |U_B(i,j)| \to_d \infty$ as $T \to \infty$.

Therefore, the minimizer of $TV_T(U_B, U_D)$ satisfies that $U_B(i,j) = 0$ if $b_{ij} = 0$ with probability tending to 1. For the nonzero b_{ij}'s and D,

$$(\sqrt{T}(\hat{B} - B), \sqrt{T}(\hat{D} - D)) \to_d \text{argmin}_{\{U_B, U_D\}} - tr(U_D D^{-1} U_D D^{-1}) + tr(UW_T), \quad (15)$$

where the minimum is taken over all lower triangular matrices U_B's with the ij-th element being 0 if $b_{ij} = 0$. The desired results then follow.

References

Banerjee, S., Gelfand, A. E., and Carlin, B. P. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC.

Chinn, P. Z., Chvatalova, J., Dewdney, A. K., and Gibbs, N. E. (1982), "The bandwidth problem for graphs and matrices - a survey," *J. Graph Theory*, 16, 223–254.

Cressie, N. (1993), *Statistics for Spatial Data*, Wiley.

Cuthill, E. and McKee, J. (1969), "Reducing the bandwidth of sparse symmetric matrices," *In Proc. 24th Nat. Conf. ACM*, 157–172.

Duff, I. S., Erisman, A. M., and Reid, J. K. (1989), *Direct Methods for Sparse Matrices*, Oxford University Press, New York.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics (with discussion)*, 407–499.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fourer, R., Gay, D. M., and Kernighan, B. W. (2003), *AMPL: A Modeling Language for Mathematical Programming*, Duxbury Press.

Fuentes, M. and Smith, R. (2001), "A new class of nonstationary models," *Tech. report at North Carolina State University.*

George, A. and Liu, J. W. H. (1981), *Computer Solution of Large Sparse Positive Definite Systems*, Prentice Haa Inc., Englewood Cliffs.

Green, P. J. and Sibson, R. (1978), "Computing Dirichlet Tessellations in the Plane," *The Computer Journal*, 21, 168–173.

Groisman, P. (2000), *Data Documentation for TD-3721: Gridded US Daily Precipitation and Snowfall Time Series.*, National Climatic Data Center, Asheville, N.C.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag: New York.

Higdon, D., Swall, J., and Kern, J. (1999), *Non-stationary spatial modeling*, In Bayesian Statistics 6, eds. J.M. Bernardo et al., Oxford University Press, pp. 761–768.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance selection and estimation via penalised normal likelihood," *Biometrika*, 93, 85–98.

Knight, K. and Fu, W. J. (2000), "Asymptotics for lasso-type estimators," *Annals of Statistics*, 28, 1356–1378.

Lee, L.-F. (2004), "Asymptotic Distributions of Quasi-maximum Likelihood Estimators for Spatial Autoregressive Models," *Econometrica*, 72, 1899–1925.

Levina, L. and Zhu, J. (2006), "Sparse estimation of large covariance matrices via a hierarchical lasso penalty," *Submitted*.

Meinshusen and Buhlmann, P. (2006), "High-dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, 34, 1436–1462.

Ord, K. (1975), "Estimation Methods for Models of Spatial Interaction," *Journal of the American Statistical Association*, 70, 120–126.

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

Papadimitriou, C. H. (1976), *The NP-Completeness of the bandwidth minimization problem*, Springer.

Pintore, A. and Holmes, C. C. (2003), "Constructing localized non-stationary covariance functions through the frequency domain," *Technical Report. Imperial College, London*.

Pourahmadi, M. (1999), "Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation," *Biometrika*, 86, 677–690.

Ripley, B. D. (1981), *Spatial Statistics*, John Wiley & Sons.

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

Sampson, P. D. and Guttorp, P. (1992), "Nonparametric Estimation of Nonstationary Spatial Covariance Structure," *Journal of the American Statistical Association*, 87, 108–119.

Smirnov, O. and Anselin, L. (2001), "Fast Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: A Characteristic Polynomial Approach," *Computational Statistics & Data Analysis*, 35, 301–319.

Speed, T. P. and Kiiveri, H. T. (1986), "Gaussian Markov Distributions over Finite Graphs," *The Annals of Statistics*, 14, 138–150.

Stein, M. L. (2005), "Nonstationary spatial covariance functions," *Unpublished technical report.*

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

Whittle, P. (1954), "On Stationary Processes in the Plane," *Biometrika*, 41, 434–449.

Yannakakis, M. (1981), "Computing the Minimum Fill-In is NP-Complete," *SIAM Journal on Algebraic and Discrete Methods*, 2, 77–79.

Yuan, M. and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35.

Zhang, H. H. and Lu, W. (2007), "Adaptive lasso for Cox's proportional hazards model," *Biometrika*, to appear.

Zou, H. (2006), "The adaptive Lasso and its oracle properties." *Journal of the American Statistical Association*, 101, 1418–1429.