

**Estimating Spatial Covariance using Penalized Likelihood with
Weighted L_1 Penalty**

Yufeng Liu

Department of Statistics and Operations Research

Carolina Center for Genome Sciences

University of North Carolina at Chapel Hill

`http://www.unc.edu/~yfliu`

Joint work with Zhengyuan Zhu (UNC)

Examples of Spatial Data on regular and irregular grids

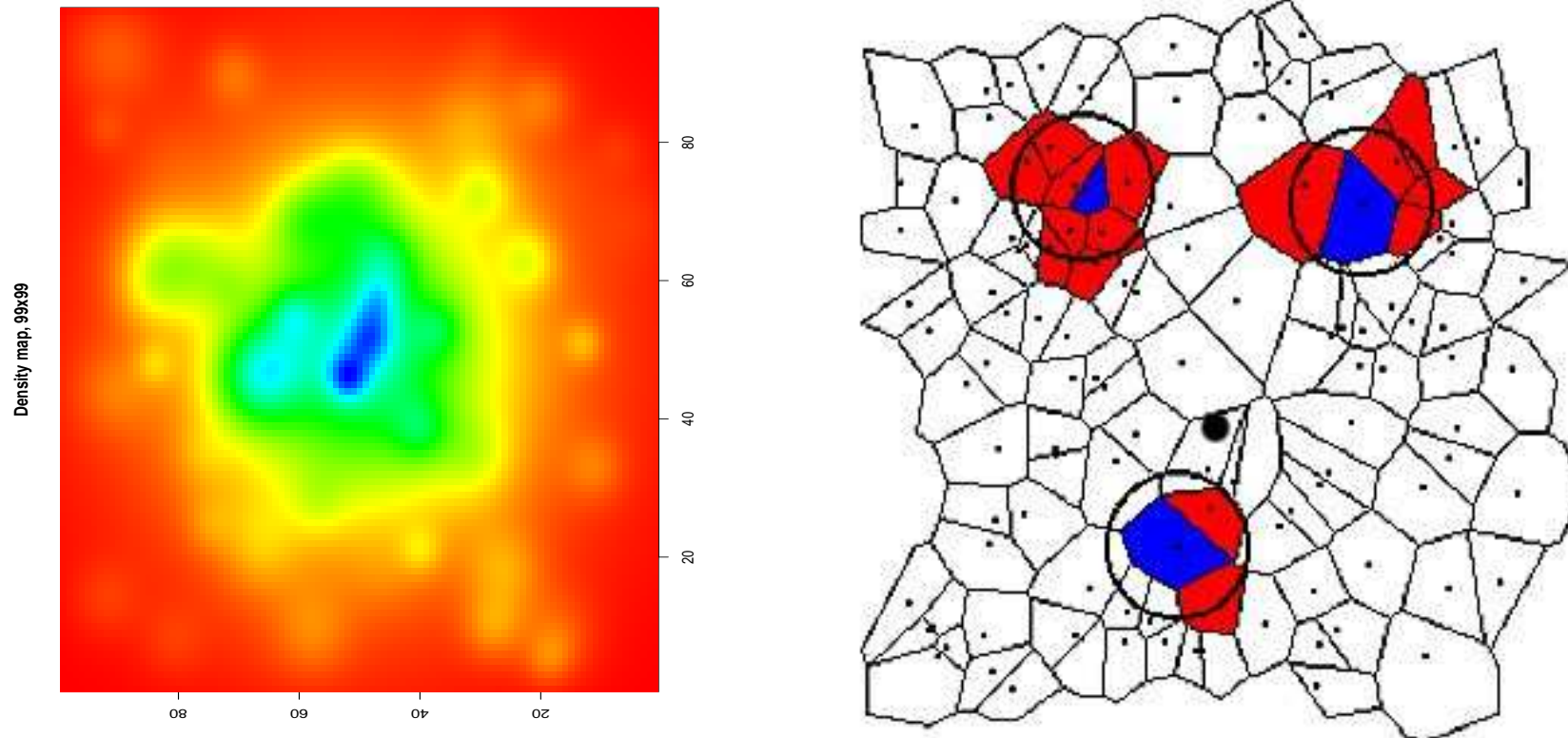
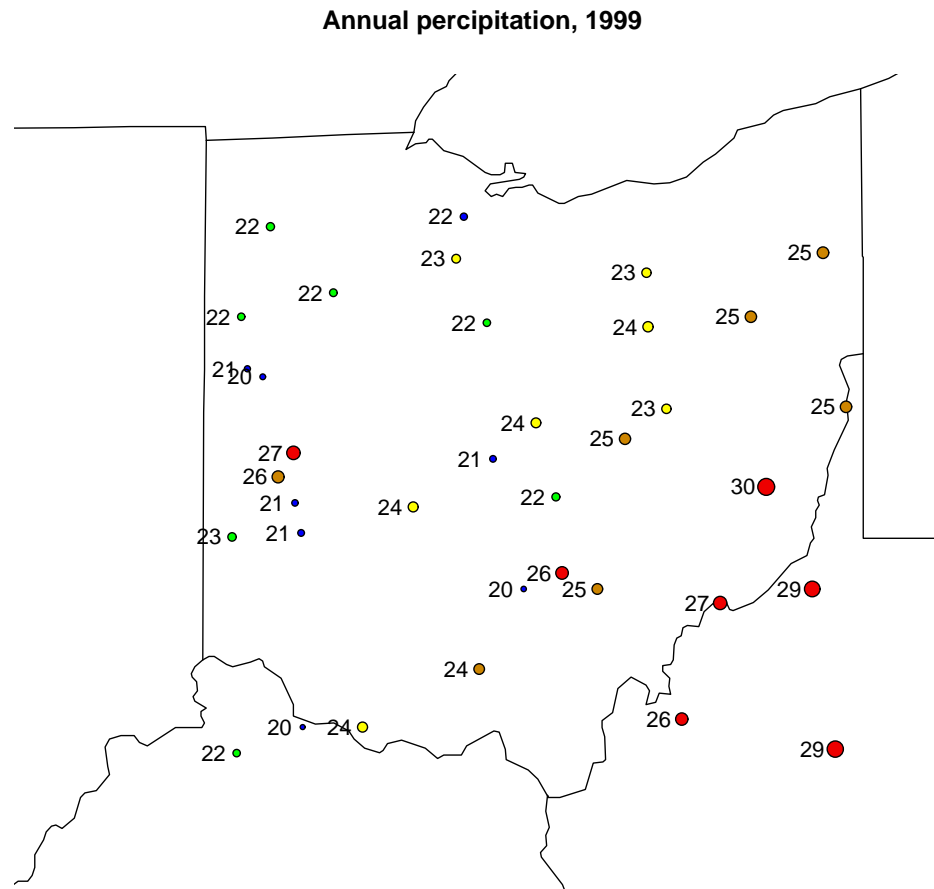


Figure 1: Left: image of the mass density of a galaxy cluster; Right: an irregular grid.

Example of Spatial Data on R^2



Gaussian Models

- $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)' \sim N(\mu, \Sigma)$
- Estimation of Σ important for estimating μ and spatial prediction.
- For large spatial data
 - Computational difficulty in estimation and prediction.
 - Non-stationarity.
 - Positive definite constraint on Σ .

Gaussian Markov Random Fields

- $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)' \sim N(\mu, \Sigma)$
- Conditional dependence structure can be described by an undirected graph $G = (X, E)$.
- X is the set of vertices of size n , with x_i corresponding to Z_i .
- 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 $Adj(Y) = \{x \in X \setminus Y : \{x, y\} \in E \text{ for some } y \in Y\}$.
- For Y a single vertex $\{y\}$, we write $adj(y)$ and refer to it as the neighbor of vertex y .
- Markov property: $Z_i \perp \mathbf{Z}_{-\{i, adj(i)\}} \mid \mathbf{Z}_{adj(i)}$.

Gaussian Markov Random Fields

- Speed and Kiiveri (1986):

$$Z_i \perp Z_j | \mathbf{Z}_{-ij} \Leftrightarrow x_i \notin \text{adj}(x_j) \Leftrightarrow \Sigma_{ij}^{-1} = 0.$$

- For small $\text{adj}(i)$, Σ^{-1} is a sparse matrix.
- Spatial prediction only depends on neighbor.
- Good approximation to GRF (Stein (2002), Rue and Tjelmeland (2002))

Simultaneous Autoregression Models

- Spatial SAR: Whittle (1954), Ripley (1981), Cressie (1993).

$$\mathbf{Z} = B\mathbf{Z} + \boldsymbol{\epsilon} \quad (1)$$

- $B = \{b_{ij}\}$ describes the spatial dependence with $b_{ii} = 0$ and $(I - B)$ nonsingular.
- $\boldsymbol{\epsilon} \sim N(0, D)$, D is a diagonal matrix with diagonal elements d_i .
- B and D are related to Σ by $\Sigma^{-1} = (I - B)'D^{-1}(I - B)$.
- B is in general not identifiable as the above decomposition of Σ^{-1} is not unique.
- One special mode assuming $B = \rho W$ with W a given matrix received some attention in econometrics literature (Ord 1975, Smirnov and Anselin 2001, Lee 2004).

Conditional Autoregression Models

- Joint density specified through full conditionals (Besag 1974, 1975).
- $G = (X, E)$ determined by spatial closeness.

Our Model and Notation

- Let $\{Z(s_i, t) : i = 1, \dots, n; t = 1, \dots, T\}$ be T independent copies of a non-stationary spatial GMRF observed at location $S = \{s_1, s_2, \dots, s_n\}$.
- We model $Z(s_i, t)$ using (1) with the constraint that B is a lower triangular matrix, i.e., $b_{ij} = 0, \forall i \leq j$.
- Equivalent to the modified Cholesky decomposition $\Sigma^{-1} = L'D^{-1}L = (I - B)'D^{-1}(I - B)$.
- No assumption on neighbourhood structure: we estimate $G = (X, E)$ as well as the coefficients.

LASSO type estimator

- The log likelihood function of \mathbf{Z} is given by

$$l(B, D; \mathbf{Z}) = -\frac{T}{2} \sum_{i=1}^n \log d_i - \frac{1}{2} \sum_{t=1}^T \sum_{i=1}^n \frac{1}{d_i} \left(Z(s_i, t) - \sum_{j < i} b_{ij} Z(s_j, t) \right)^2. \quad (2)$$

- Direct maximization of (2) gives sample covariance matrix, known to be unstable for large covariance matrix and not sparse.
- We estimate B and D by minimizing

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

- The L_1 penalty forces the small elements in the estimated B to shrink to zero, which leads to a sparse estimated precision matrix of Σ^{-1} .

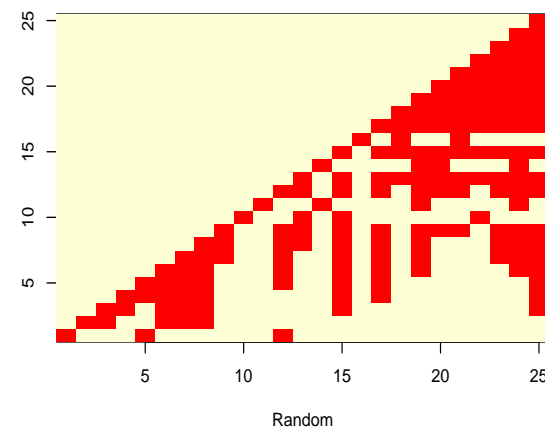
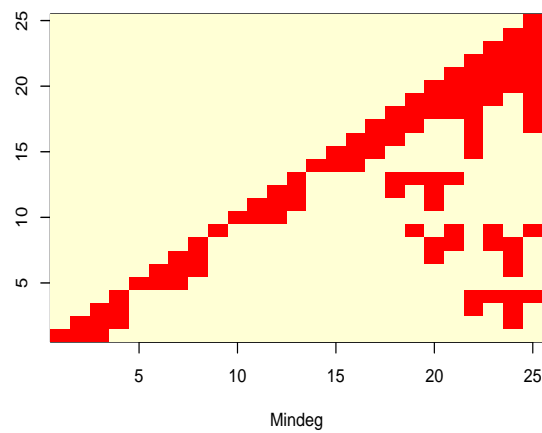
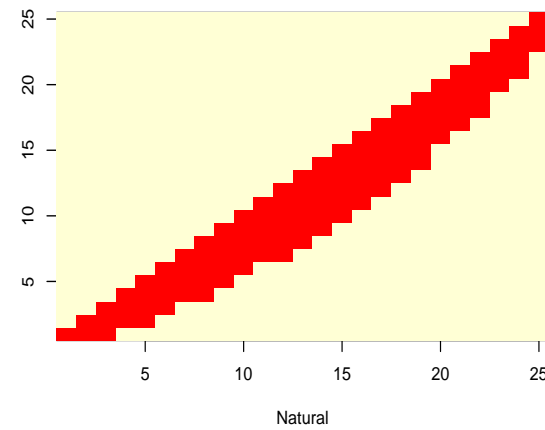
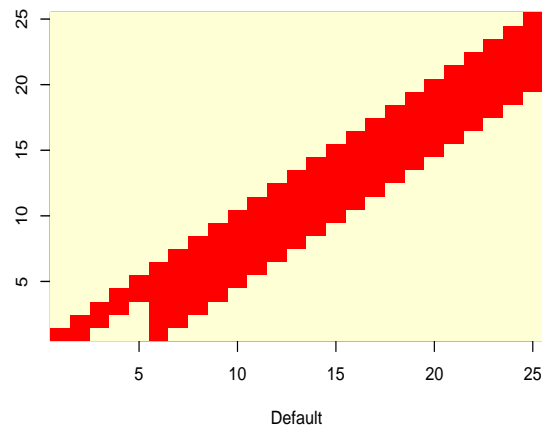
Modified Cholesky Decomposition for Longitudinal Data

- Pourahmadi (1999), Huang et. al. (2006) used such decomposition to model covariance matrix of longitudinal data.
- The key idea: decompose $\Sigma^{-1} = L'D^{-1}L$. Estimate L and D instead of Σ to get rid of the P.D. constraint.
- Longitudinal data has a natural time order. Such decomposition leads to the well known AR model, with b_{ij} the autoregression coefficients.

Modified Cholesky Decomposition for Spatial Data

- No natural ordering for spatial data.
- Estimation not invariant to permutation of the data.
- Does the ordering matter? Which ordering should one use?

Cholesky Decomposition with Different Ordering



Cholesky Decomposition with Different Ordering

- 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).
- The zero pattern of the sparse matrix Σ^{-1} is known.
- NP-hard, various heuristic algorithms available to minimize bandwidth or fill-in (Cuthill-McKee, minimum degree, etc.)

Ordering for estimating Σ^{-1}

- Want sparsity in L .
- Want to preserve spatial closeness in the ordering for appropriate weighting.
 - Minimum bandwidth preferred over minimum fill-in
- Ordering algorithm for $G = (X, E)$.
 - For data on R^2 , can define a natural spatial neighborhood structure based on the Vornoi tessellation and derive initial G .
 - G can be constructed from preliminary estimator of Σ^{-1} .

Natural Ordering Algorithm

Adapted from the reverse Cuthill-McKee algorithm to use the distance info.

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 $i = 1, 2, \dots, n - 1$:
 - Construct the set $A_i = Adj(q_i) \setminus Q$, the adjacency set of the i th 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 $G \setminus Q$. 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.

Properties of the Natural Ordering

For graph on a regular $m \times n$ grids in R^2

- Both natural ordering and default ordering achieves minimum bandwidth.
- For $m = n$ large, Natural ordering has approximately 1/3 less non-zero elements.

Optimization Routine

Goal: minimize the following objective function with respect to (B, D) :

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

- For any given B , the minimizer of (4) can be obtained as

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

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

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

Standard Optimization Routine

- Introduce slack variables $\xi_{ij} \geq 0$ and simplify problem (6) to a quadratic programming (QP) problem with linear constraint:

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

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

- Minimize (4) via an iterative procedure. Initialize B and D using the modified Cholesky decomposition of the sample covariance matrix.

Optimization using the Regularized Solution Path Algorithm

- Regularized solution path algorithm (Efron et al., 2004; Rosset and Zhu 2007): gives the entire solution path for the LASSO problem

$$\min_{\{\beta_j; j=1, \dots, i-1\}} \sum_{t=1}^T (y_t - \sum_{j<i} \beta_j x_{jt})^2 + \lambda^* \sum_{j<i} |\beta_j|. \quad (7)$$

- For any given $d_i > 0$, (4) can be simplified to n small problem

$$\min_{\{b_{ij}\}} f(\lambda, d_i) = \sum_{t=1}^T (Z(s_i, t) - \sum_{j<i} b_{ij} Z(s_j, t))^2 + 2\lambda d_i \sum_{j<i} w_{ij} |b_{ij}|, \quad (8)$$

which is equivalent to the LASSO problem.

Optimization

- The solution path of (7) yields the $f(\lambda_0, d_i)$ as a function of d_i with $d_i = \lambda^* / (2\lambda_0)$.
- Solution of d_i is the minimizer of

$$g(d_i) = \frac{T}{2} \log d_i + \frac{f(\lambda_0, d_i)}{2d_i}. \quad (9)$$

- Can solve D and B for different λ with very little additional costs.

The Tuning Procedure

- We use entropy criterion

$$Ent(\Sigma, \hat{\Sigma}_\lambda) = tr(\Sigma^{-1}\hat{\Sigma}_\lambda) - \log(\Sigma^{-1}\hat{\Sigma}_\lambda) - n. \quad (10)$$

and cross-validation.

- Other existing model selection methods including AIC and BIC can also be used to choose λ .

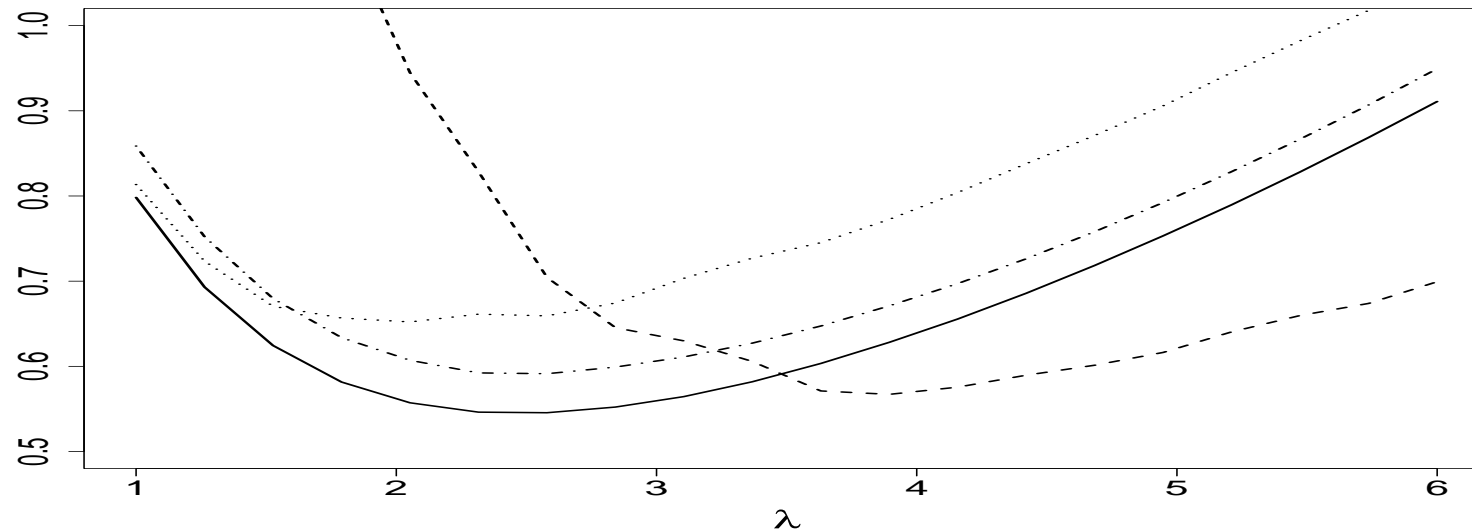


Figure 2: Comparison of different methods for selecting the tuning parameter λ . Solid line: entropy with true Σ ; dotted line: AIC; dashed line: BIC; dash-dotted line: entropy with $\hat{\Sigma}$ from the validating data.

Asymptotic Behavior for Equal Weights $w_{ij} = 1$

Theorem 1 *If $\sqrt{T}\lambda \rightarrow \lambda_0 \geq 0$ as $T \rightarrow \infty$ and $w_{ij} = 1$ for $j < i$, the estimator (\hat{B}, \hat{D}) of (4) satisfies that*

$$(\sqrt{T}(\hat{B} - B), \sqrt{T}(\hat{D} - D)) \rightarrow_d \operatorname{argmin}_{\{U_B, U_D\}} V(U_B, U_D), \quad (11)$$

where U_B is a lower triangular matrix and U_D is a diagonal matrix with positive diagonal elements, and $V(U_B, U_D) = -\operatorname{tr}(U_D D^{-1} U_D D^{-1})$

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

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 $\operatorname{vec}(W) \sim N(0, \Lambda)$, and

$\Lambda = \operatorname{Cov}(\operatorname{vec}(W))$ such that

$$\operatorname{Cov}(w_{ij}, w_{kl}) = \operatorname{Cov}(Z(s_i)Z(s_j), Z(s_k)Z(s_l)).$$

Asymptotic Behavior for Consistent Weights

Theorem 2 Denote (\hat{B}, \hat{D}) as the minimizer of (4). Assume $\sqrt{T}\lambda \rightarrow 0$ and $T\lambda \rightarrow \infty$ as $T \rightarrow \infty$ and $w_{ij} = 1/\tilde{b}_{ij}$ for $j < i$ with $\sqrt{T}(\tilde{b}_{ij} - b_{ij}) \rightarrow 0$ in probability, then we can conclude that

(1). $P(\hat{b}_{ij} = 0) \rightarrow 1$ 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).

Simulation Studies

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.

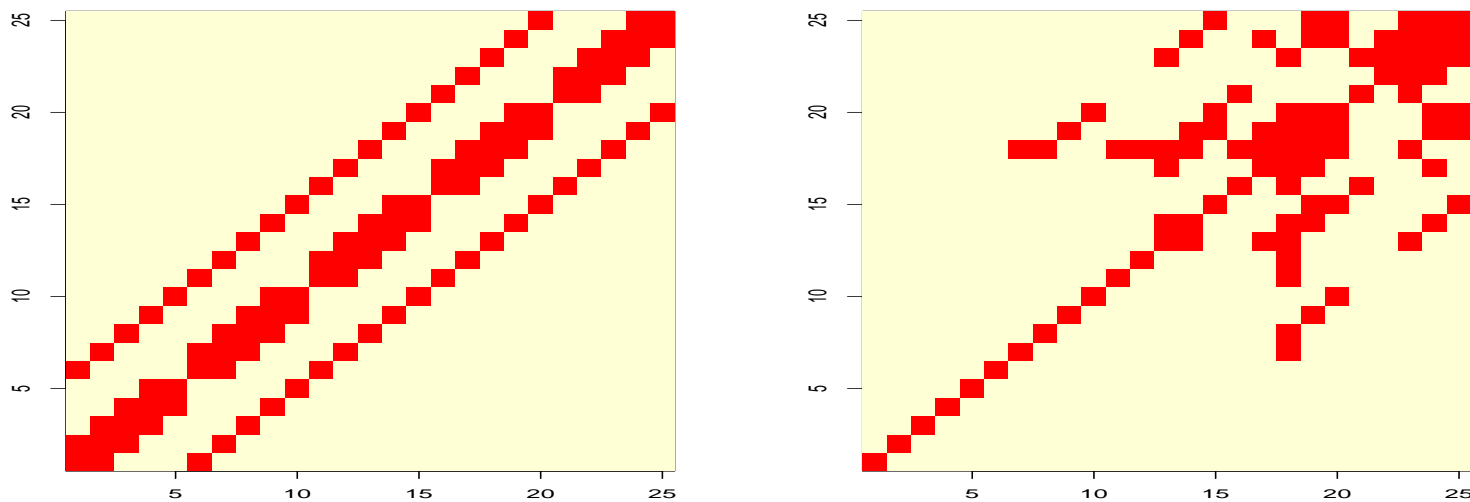


Figure 3: Plots of zero patterns of the two precision matrices corresponding to the stationary and nonstationary simulated examples in Section 6.

Comparison for Stationary GMRF: Entropy Measure

		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. CW and DW represent constant weighting and distance weighting respectively.

Comparison for Stationary GMRFs: Zero Pattern

		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. CZ and CN represent percentages of correctly identified zeros and non-zeros in the precision matrix respectively by each method.

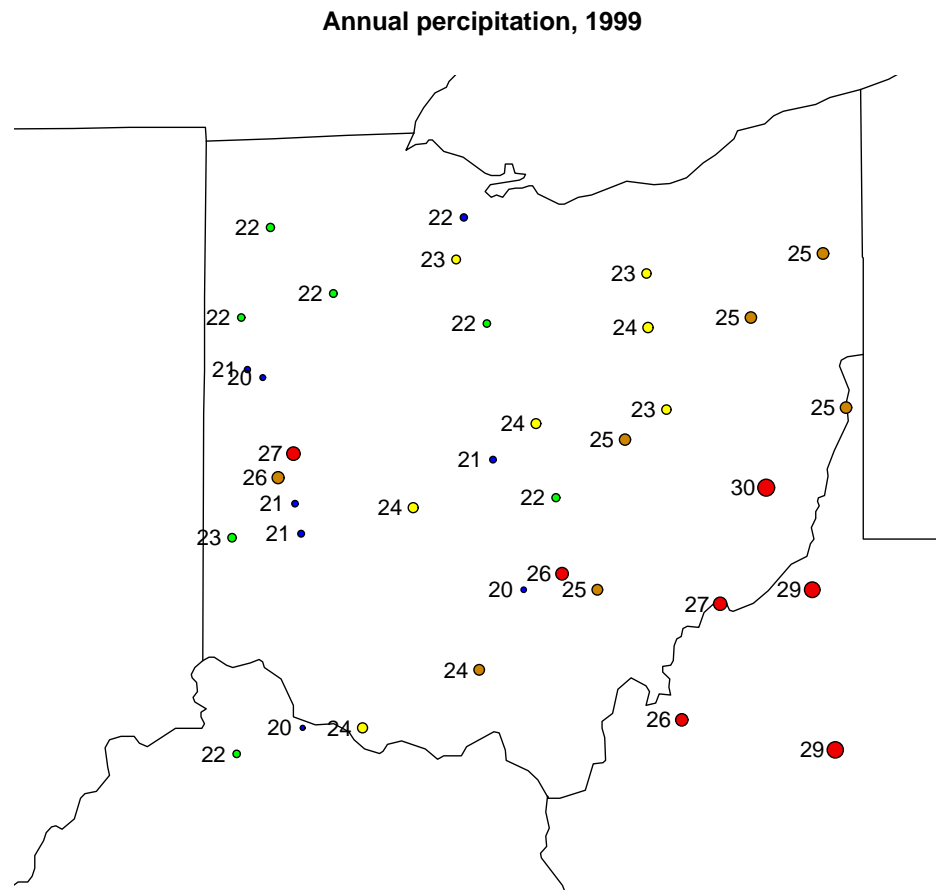
Comparison for Non-stationary GMRFs

		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.

Annual Rainfall Data at Ohio

- The study region covers the state of Ohio and some surrounding area.
- The time period between 1960 and 1999.
- No significant deviation from the Gaussian assumption.
- No significant time dependence
- Estimate the inverse covariance matrix and examine the estimated conditional independence structure.
- Compare the prediction performance using our estimated covariance matrix with several other methods.



Examples of Neighborhood Structure

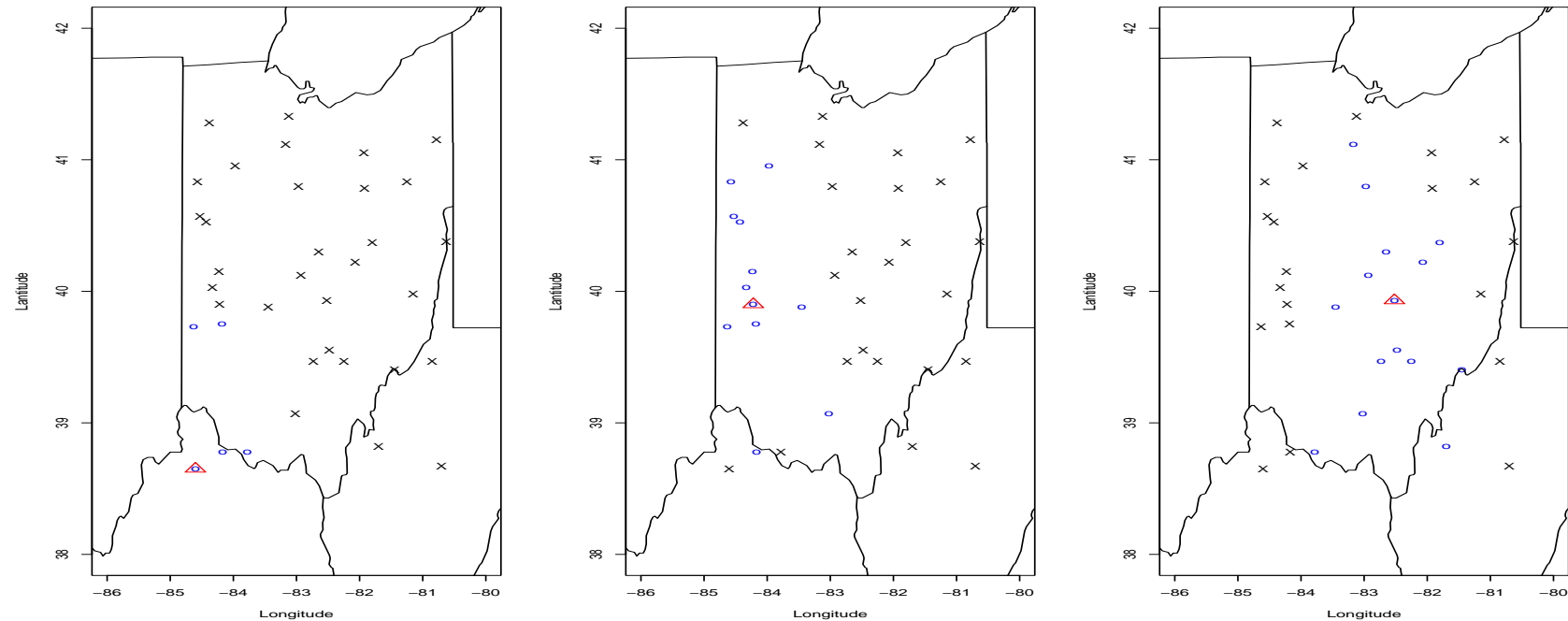


Figure 4: Examples of neighborhood structure. Conditional on observations at the blue dots, the observation at the red triangle is independent of the observations at the black crosses.

Comparison of Prediction Performance

- For each year, 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.
- The mean square prediction error (MSPE) is computed for each location across forty years, and the median MSPE across all locations are reported.

	PLIK	SC	EXP	IDWA
median MSPE	2.31	8.08	2.43	2.64

Table 4: Comparison of prediction performance

Discussion

- Gaussian graphical model is permutation invariant (Yuan and Lin, 2007, etc) and can be generalized for spatial cov. estimation.
- Efficient R package *glasso* can be used (Friedman et al., 2007).
- Comparisons of the two approaches