Large Gaussian Covariance Matrix Estimation with Markov Structures[1]

Xinwei Deng and Ming Yuan

School of Industrial and Systems Engineering, Georgia Institute of Technology

(October 20, 2008)

Abstract

Covariance matrix estimation for a large number of Gaussian random variables is a challenging yet increasingly common problem. A fact neglected in practice is that the random variables are frequently observed with certain temporal or spatial structures. Such a problem arises naturally in many practical situations with time series and images as the most popular and important examples. Effectively accounting for such structures not only results in more accurate estimation but also leads to models that are more interpretable. In this paper, we propose shrinkage estimators of the covariance matrix specifically to address this issue. The proposed methods exploit sparsity in the inverse covariance matrix in a systematic fashion so that the estimate conforms with models of Markov structure and is amenable for subsequent stochastic modeling. The present approach complements the existing work in this direction that deals exclusively with temporal orders and provides a more general and flexible alternative to explore potential Markov properties. We show that the estimation procedure can be formulated as a semi-definite program and efficiently computed. We illustrate the merits of these methods through simulation and the analysis of a real data example.

1 Introduction

In the Gaussian covariance matrix estimation problem, one wishes to estimate the covariance matrix of a multivariate normal vector $X = (X^{(1)}, \ldots, X^{(p)})'$ given an independent and identically distributed sample $X_1, \ldots, X_n$ of X. Assuming that $X \sim \mathcal{N}(\mu, \Sigma)$, μ is typically

[1]Address for correspondence: Ming Yuan, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 (E-mail: myuan@isye.gatech.edu).

estimated by the sample mean $\bar{X} = (\bar{X}^{(1)}, \ldots, \bar{X}^{(p)})'$ where

$$\bar{X}^{(i)} = \frac{1}{n} \sum_{j=1}^{n} X_j^{(i)}, \tag{1}$$

and Σ by the sample covariance matrix. Increasingly common in practice, we need to estimate the covariance matrix when the dimension p is moderate or large. It is well known that the sample covariance matrix is not a stable estimate in such cases because of the large number of unknowns involved. Even worse, when $p \geq n$, the sample covariance matrix is not positive definite and therefore not a legitimate covariance matrix estimator for many purposes.

In recent years, a number of new methods have been developed to overcome these drawbacks of the sample covariance matrix. Earlier developments have focused on shrinking the eigenvalues of the sample covariance matrix (Stein, 1977; Haff, 1980; Dey and Srinivasan, 1985; Perron, 1992). Similar idea of perturbing the eigenvalues of the sample covariance matrix also appears in the approach of Ledoit and Wolf (2003) who considered a linear combination of the sample covariance matrix and the identity matrix. Bayesian treatment of covariance matrix estimation can also be found in Smith and Kohn (2002) and Wong, Carter and Kohn (2003) and references therein. Covariance matrix estimation is closely related to the covariance selection problem (Dempster, 1972) where the interest is in constructing a graphical model that can be used to describe the conditional independence structure among the variables. Yuan and Lin (2007) proposed penalized likelihood methods to simultaneously addressing both problems. Denote $C = (c_{ij}) = \Sigma^{-1}$. A zero entry $c_{ij} = 0$ indicates zero partial correlation between the two random variables $X^{(i)}$ and $X^{(j)}$ and therefore conditional independence given the other variables. The shrinkage estimators of Yuan and Lin (2007) encourage sparsity in the inverse covariance matrix and thus conduct estimation and selection at the same time. Such estimator has also been recently studied by d'Aspremont, Banerjee, and El Ghaoui (2008), Rothman et al. (2008) and Friedman, Hastie and Tibshirani (2008). Correspondence with a sparse graphical model makes these covariance matrix estimators more interpretable.

A fact neglected by these existing methods is that the random variables are often observed with certain temporal or spatial structures, which arises naturally in the analysis of time series or images. One exception is the approach pioneered by Pourahmadi (1999; 2000) who considered the case when the variables are temporally ordered. Pourahmadi suggested to

work on a modified Cholesky decomposition of the covariance matrix: $T\Sigma T' = D$ where

$$T = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \phi_{21} & 1 & 0 & \dots & 0 \\ \phi_{31} & \phi_{32} & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \phi_{p1} & \phi_{p2} & \phi_{p3} & \dots & 1 \end{pmatrix}$$

is a lower-triangular matrix with ones on its diagonal and D is a diagonal matrix. It can be shown that the sub-diagonal entries on the ith row of T, $(\phi_{i1}, \dots, \phi_{i,i-1})$ can be interpreted as the minus of the coefficients when regressing $X^{(i)}$ over $X^{(1)}, \dots, X^{(i-1)}$. This provides a natural reparametrization of the covariance matrix when $X^{(i)}$s are ordered temporally such as in time series. Various shrinkage methods have been proposed within this framework to encourage sparsity in T (Wu and Pourahmadi, 2003; Huang, Liu, Pourahmadi and Liu, 2006; Bickel and Levina, 2008; Levina, Rothman and Zhu, 2008). In particular, Levina et al. (2008) introduced a penalized likelihood estimate that encourages the sparsity of the inverse covariance matrix by forcing a particular pattern of sparsity on T. Note that $\Sigma^{-1} = T'D^{-1}T$. By requiring $\phi_{ij} = 0$ if $\phi_{i,j+1} = 0$ and $j < i-1$, some entries of the inverse covariance matrix that are far away from the diagonal can be shrunken to zeros and therefore the estimate can be interpreted as Markov chains. These approaches, however, only apply to temporal orders and may not be suitable if the $X^{(i)}$s are observed with more complicated structures such as spatial orders.

To elaborate, consider analyzing handwritten digits based on a training sample of images (LeCun et al., 1990) as shown in Figure 1. Covariance matrix estimation of the intensity values on the $256 = 16 \times 16$ pixels plays a critical role in various statistical analysis such as principal component analysis and linear discriminant analysis. The correlation between the intensity on two pixels is clearly related to the positions of the pixels. Furthermore, images of this sort can most often be adequately modeled as a Markov random field of a relatively small order since the intensity values on pixels far away from each other are generally independent conditional on intensities of the other pixels (Winkler, 2006). A covariance matrix estimate that conforms with such models not only reduces the dimensionality of the estimation problem but also is much more valuable in subsequent stochastic modeling. Unlike the Markov structure in the temporally ordered cases, the Markov random field can not be

Figure 1: Sample images of handwritten digits: each image is of size 16×16.

inferred from the sparsity pattern of matrix T of the modified Cholesky decomposition. To illustrate, consider four random variables that are observed from a 2×2 grid as shown below.

$X^{(1)}$	$X^{(2)}$
$X^{(3)}$	$X^{(4)}$

A Markov random field of order one is equivalent to $c_{41} = c_{32} = 0$, which can only imply that $\phi_{41} = 0$ as illustrated by (2).

$$\Sigma^{-1} = \begin{pmatrix} 1 & 0.4 & 0.4 & 0 \\ 0.4 & 1 & 0 & 0.4 \\ 0.4 & 0 & 1 & 0.4 \\ 0 & 0.4 & 0.4 & 1 \end{pmatrix} \Longrightarrow T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.59 & 1 & 0 & 0 \\ 0.48 & -0.19 & 1 & 0 \\ 0 & 0.4 & 0.4 & 1 \end{pmatrix}. \tag{2}$$

But on the other hand, a Markov random field of order one can not be inferred from $\phi_{41} = 0$. A counterexample can be given by simply altering the $(3, 2)$ entry of T while keeping the

same D:

$$\Sigma^{-1} = \begin{pmatrix} 1 & 0.55 & 0.4 & 0 \\ 0.55 & 1 & 0.32 & 0.4 \\ 0.4 & 0.32 & 1 & 0.4 \\ 0 & 0.4 & 0.4 & 1 \end{pmatrix} \Longrightarrow T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.59 & 1 & 0 & 0 \\ 0.48 & 0.19 & 1 & 0 \\ 0 & 0.4 & 0.4 & 1 \end{pmatrix}. \tag{3}$$

This simple example shows that the modified Cholesky decomposition may no longer be suitable for exploring Markov structures when the variables are observed with structures more general than temporal orders.

The lack of a method that can handle general Markov structures among the random variables motivates the present work. In this paper, we propose a more direct strategy to explore conditional independence relationships among variables when they are observed with temporal and spatial structures. We suggest to exploit sparsity directly on the inverse covariance matrix. We consider constrained maximum likelihood methods with constraints that encourage sparsity in a systematic fashion so that estimates that conform with models of Markov structure are favored. We shall introduce the proposed methods in the next section, followed by examples in Sections 3 and 4. We conclude with some discussions in the last section.

2 Methodology

The log likelihood for μ and $C = \Sigma^{-1}$ based on a random sample $X_1, \ldots, X_n$ of X is

$$\ln |C| - \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)' \, C \, (X_i - \mu) = \ln |C| - \text{trace}(C\bar{A}) \tag{4}$$

up to a constant not depending on μ and C, where

$$\bar{A} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right) \left(X_i - \bar{X}\right)' \tag{5}$$

is the maximum likelihood estimate of Σ. To estimate C, we consider a shrunken version of $\tilde{C} = \bar{A}^{-1}$: $C = (\theta_{ij}\tilde{c}_{ij})$ where θ'_{ij}s are shrinkage coefficients. Other choices of $\tilde{C}$ are also possible; we focus on $\bar{A}^{-1}$ in this paper to fix ideas. Given that $\tilde{C}$ is a reasonably good initial estimate of the inverse covariance matrix it is appropriate to require that the shrinkage

coefficients be nonnegative, $\theta_{ij} \geq 0$. To achieve sparse graph structure and encourage sparsity in C, one can maximize the log likelihood subject to the constraint that

$$\sum_{i \neq j} \theta_{ij} \leq M \tag{6}$$

for some tuning parameter $M \geq 0$. This is the so-called graphGarrote estimator proposed by Yuan and Lin (2007). Clearly when $M = +\infty$, the constraint becomes inactive and the resulting estimate reduces to $\bar{A}^{-1}$ and no shrinkage takes place. On the other hand when $M = 0$, all the off-diagonal entries of the inverse covariance matrix will be shrunken to zero and the estimate becomes diagonal which implies mutual independence among $X^{(i)}$s. A choice of tuning parameter M between these two extremes will result in covariance matrix estimates with varying degrees of sparsity. The procedure is similar in spirit to the nonnegative garrote estimator proposed by Breiman (1995) for linear regression.

We now consider the situation when the random variables are observed in a space with a certain distance measure defined. Assume that $X^{(i)}$ is observed at location t_i. For example, t_i is a point in a two dimensional lattice in the case of images. Most often dependence between two variables dwindles as the distance between them increases. To incorporate this prior information into the estimation of the covariance matrix, we impose the following constraints on the shrinkage coefficients.

$$\theta_{ij} \leq \theta_{ik} \quad \text{if} \quad d_{ij} \geq d_{ik} \tag{7}$$

where $d_{ij} = \text{dist}(t_i, t_j)$ is the pairwise distance. Because the entries of $\tilde{C}$ are generally nonzero, constraint (7) implies that $c_{ij} = 0$ if $c_{ik} = 0$. It is worth pointing out that this constraint only encourages more shrinkage towards zero for entries that are farther away from the diagonal to reflect our preference towards Markov models; it does not force $c_{ij} \leq c_{ik}$. In

summary, we propose to estimate C by $\widehat{C} = (\widehat{\theta}_{ij}\tilde{c}_{ij})$ where $\widehat{\Theta} = (\widehat{\theta}_{ij})$ is the solution to

$$\begin{aligned} \min & -\left[\ln|C| - \operatorname{trace}(C\bar{A})\right] \\ \text{subject to} \quad & C \text{ is positive definite} \\ & c_{ij} = \theta_{ij}\tilde{c}_{ij} \\ & \theta_{ij} \geq 0 \\ & \sum_{i \neq j} \theta_{ij} \leq M \\ & \theta_{ij} \leq \theta_{ik} \quad \text{if} \quad d_{ij} \geq d_{ik} \text{ and } j, k \neq i. \end{aligned} \tag{8}$$

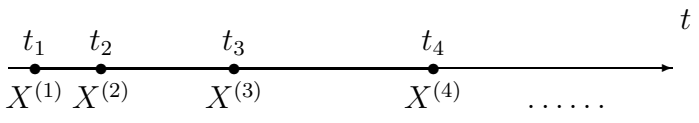
The problem is a semi-definite program and can be easily solved using standard software packages such as `SDPT3` (Tütüncü, Toh and Todd, 2003).

Thus far we have assumed that the tuning parameter M is fixed. In practice, it also needs to be estimated. A commonly used approach is the multi-fold cross-validation which can be computationally demanding. A much more efficient alternative is the BIC criterion introduced by Yuan and Lin (2007):

$$\mathrm{BIC}(M) = -\ln|\hat{C}(M)| + \operatorname{trace}(\hat{C}(M)\bar{A}) + \frac{\ln(n)}{n}\sum_{i \leq j}\hat{e}_{ij}(M), \tag{9}$$

where $\hat{e}_{ij} = 0$ if $\hat{c}_{ij} = 0$, and $\hat{e}_{ij} = 1$ otherwise. We shall adopt this criterion in our implementation and it works very well in practice according to our experience.

Note that without the last constraint in (8), our estimate becomes the graphGarrote of Yuan and Lin (2007). The last constraint takes the temporal or spatial structure of the observation into consideration. Consider, for example, the case where the observations are temporally ordered.



Because d_{ij} is a monotone increasing transformation of $|i-j|$, the last constraint can be simplified to

$$\theta_{i,j+1} \leq \theta_{ij} \quad \text{if } j > i, \text{ and } \theta_{i,j+1} \geq \theta_{ij} \quad \text{if } j < i-1. \tag{10}$$

This encourages more shrinkage to the partial correlation between $X^{(i)}$ and $X^{(j)}$ if the two observations are farther away from each other. Together with the constraint on the sum of the shrinkage coefficients, it induces a sparse estimate of the inverse covariance matrix that follows a non-stationary Markov chain in that there exist $h_1, h_2, \ldots, h_p > 0$ such that

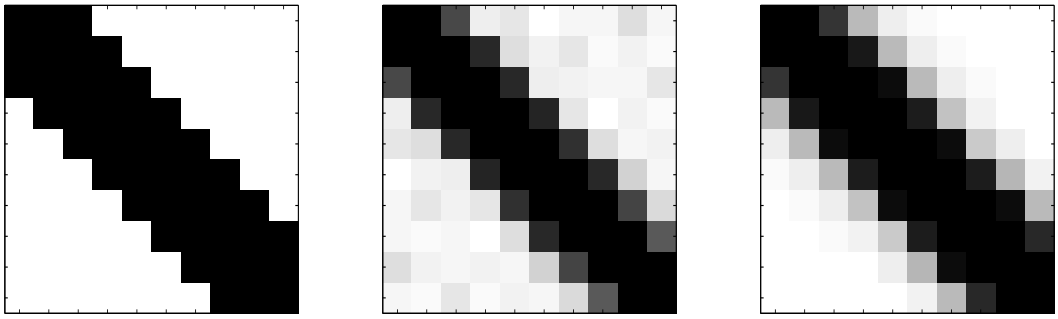
$$X^{(i)} \perp \{X^{(j)} : d_{ij} > h_i\} \big| \{X^{(j)} : 0 < d_{ij} \leq h_i\}.$$

To demonstrate its effect, we apply both graphGarrote and the proposed estimate, hereafter we refer to as the structured graphGarrote, to data sets that are simulated from a AR(2) model with $c_{ij} = 1$ if $i = j$, 0.5 if $|i - j| = 1$, 0.25 if $|i - j| = 2$ and 0 otherwise. We consider sample size $n = 100$ and dimension $p = 10$. For both estimates, the tuning parameter M is chosen by the BIC criterion defined by (9). Panel (a) of Figure 2 shows the nonzero pattern of the true inverse covariance matrix. A black block indicates that the coefficient is not zero and a white block corresponds to a zero entry. Panels (b) and (c) give the heatmap representing the frequency that each entry of the inverse covariance is estimated as nonzero over 100 simulations. A darker block indicates higher frequency. It is evident that by taking advantage of the temporal order, the proposed method is more suitable to exploit the Markov structure of the true data generating mechanism. Another interesting observation from this experiment is that BIC tends to select similar values of tuning parameter M for graphGarrote and structured graphGarrote. With the structural constraints, the structured graphGarrote generally delivers less shrinkage (therefore larger θ value) to entries that are more likely to be nonzero under Markov structure.

In the case of images, the random variables are observed on a two dimensional lattice. Let $X^{(i)}$ be observed at the pixel located on the r_ith row and c_ith column. A natural distance defined on a two dimensional lattice is the so-called city block distance or Manhattan distance:

$$\text{dist}(i, j) = |r_i - r_j| + |c_i - c_j|. \tag{11}$$

It is noteworthy that the pairwise distances between observations on a regular lattice take only a relatively few distinct values. It is natural to expect that similar degrees of shrinkage is needed for entries of the inverse covariance matrix that correspond to similar pairwise distances. In other words, it is reasonable to have $\theta_{ij} = \theta_{i'j'}$ if $d_{ij} = d_{i'j'}$. For convenience, we shall refer to this modification as the homogeneous structured graphGarrote estimate. It

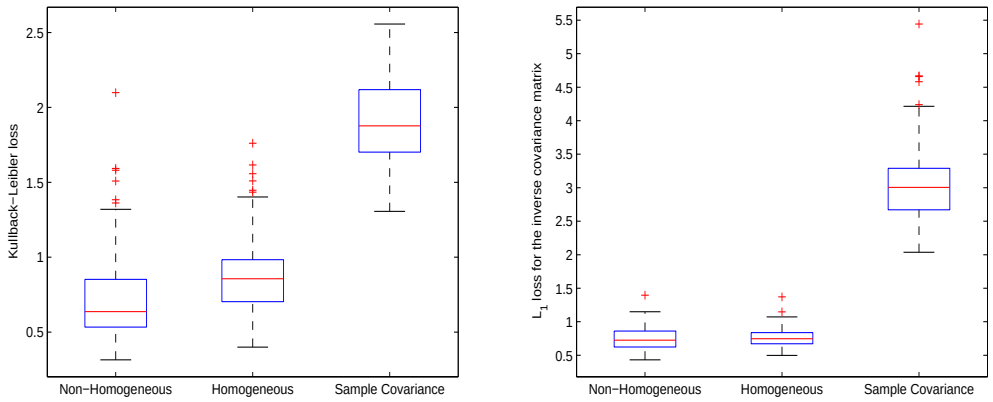


(a) True (b) graphGarrote (c) Structured graphGarrote

Figure 2: Effect of the structure constraint: Panel (a) represents the true nonzero pattern of the inverse covariance matrix with a block at the ith row and jth column indicating $c_{ij} \neq 0$; Panels (b) and (c) give the frequency that each entry of the inverse covariance matrix is estimated by a nonzero value. A darker block indicates higher frequency. Panel (b) correspond to graphGarrote and (c) corresponds to the structured graphGarrote.

is worth pointing out that the homogeneous estimate does not impose stationarity by forcing $c_{ij} = c_{i'j'}$. It, however, greatly reduces the dimensionality of the optimization problem (8), which brings about great computational efficiency. More specifically, the nonhomogeneous version involves $p(p-1)/2$ shrinkage parameters and generally linear constraints of the order $O(p^2)$, whereas the homogeneous version uses only $p-1$ shrinkage parameters and has $O(p)$ linear constraints. The difference in estimation accuracy between the structured graphGarrote and its homogeneous version is generally marginal. To illustrate this, consider $p = 4 \times 4$ random variables that are observed on a two dimensional lattice. We generate $n = 100$ observations from a multivariate normal distribution with the inverse covariance matrix generated in the following fashion. First we generate $c_{ij} \sim U(0,1)$ if $d_{ij} = 1$, and set $c_{ij} = 1$ if $d_{ij} = 0$ and 0 if $d_{ij} > 1$. Here d_{ij} represents the city block distance between i and j, and $U(0,1)$ denotes the uniform distribution from 0 to 1. Next for all i, we normalize c_{ij} so that $\sum_{i \neq j} c_{ij} = 0.9$ to ensure positive definiteness. We apply both the structured graphGarrote and its homogeneous version to the simulated data. We also include the sample covariance matrix in the comparison to serve as the baseline. Figure 3 shows the boxplot of the estimation accuracy measured by both the Kullback Leiber loss and the matrix ℓ_1 loss for

the three methods, summarized over 100 simulated data sets. Both criteria will be defined in the next section. We observe from Figure 3 that even if the true data generating mechanism is non-stationary as in this example, the structured graphGarrote and its homogeneous version behave very similarly. Because of the similarity in estimation accuracy and the great computational advantage of the homogeneous version, we shall use it throughout the paper unless otherwise indicated.



(a) Kullback-Leibler loss (b) L_1 loss for the inverse covariance matrix

Figure 3: Comparison of estimation accuracy between the structured graphGarrote and its homogeneous versions. Panels (a) and (b) correspond to loss functions given by (12) and (15) respectively.

3 Simulations

To illustrate the merits of the proposed methods, we conducted several sets of simulation studies.

3.1 Temporal Structures

We first compare through simulations the proposed method with several popular alternative shrinkage estimators of the covariance matrix when the variables are temporally ordered. We include the estimators of Huang et al. (2006), Bickel and Levina (2008), Levina et al. (2008)

as well as the sample covariance matrix for comparison. We compare these methods on the basis of the number of false positives (FP; incorrectly identified nonzero entries of Σ^{-1}), the number of false negatives (FN; incorrectly missed nonzero entries), and the Kullback-Leibler loss defined as

$$\mathrm{KL} = -\log|\hat{C}| + \mathrm{tr}(\hat{C}\Sigma) - (-\log|\Sigma^{-1}| + p), \tag{12}$$

In the approach of Bickel and Levina (2008), the Cholesky factor T is banded to estimate the inverse covariance matrix, i.e.,

$$\phi_{ij} = 0, \qquad \forall |i-j| > h$$

for some $h > 0$. The banding parameter h is chosen by cross validation using the matrix ℓ_1 loss. Huang et al. (2006) suggested adding ℓ_1 or ℓ_2 penalty

$$\lambda \sum_{i=1}^{p} \sum_{j=1}^{i-1} |\phi_{ij}|^{\gamma}, \qquad \gamma = 1 \text{ or } 2$$

on the elements of T to the normal likelihood (4), which leads to Lasso or ridge type shrinkage of the ϕ_{ij}s. The authors also suggested to choose the tuning parameter $\lambda > 0$ by cross validation. The ℓ_2 penalty ($\gamma = 2$) was used in our simulation study. Instead of ℓ_1 penalty in Huang et al. (2006), Levina et al. (2008) introduced a nested Lasso penalty on ϕ_{ij}s. The so-called J_2 nested Lasso penalty is given by $\sum_j J_{2j}$ where

$$J_{2j} = \lambda_1 \sum_{k=1}^{j-1} |\phi_{jk}| + \lambda_2 \sum_{k=1}^{j-2} \frac{|\phi_{jk}|}{|\phi_{j,k+1}|}, \tag{13}$$

and $\lambda_1, \lambda_2 > 0$ are tuning parameters. The nested Lasso penalty forces a random variable to be conditionally dependent only on its nearest neighbors. Different from banding, the number of nearest neighbors selected with the nested Lasso penalty is allowed to vary across variables. As suggested by the authors, the tuning parameters are selected with a validation set which is set aside from the original training data set.

The following three models were considered.

Model 1. $\Sigma = I_p$.

Model 2. (AR(1) model) $c_{ij} = 1$ if $i = j$, 0.45 if $|i-j| = 1$ and 0 otherwise.

Model 3. (AR(2) model) $c_{ij} = 1$ if $i = j$, 0.5 if $|i-j| = 1$, 0.25 if $|i-j| = 2$ and 0 otherwise.

For each model, we simulated data sets with sample size $n = 100$ and dimension $p = 10$, $n = 100$ and $p = 30$, or $n = 400$ and $p = 100$. Table 1 documents the means and standard errors (in parentheses), summarized from 100 runs for each combination.

As shown in Table 1, all shrinkage methods improve upon the sample covariance matrix. The improvement is particularly significant for high dimensional problems. Among the shrinkage methods, the structured graphGarrote enjoys the best performance overall in terms of estimation accuracy. It also dominates the other methods overwhelmingly in recovering the nonzero patterns of the inverse covariance matrix or equivalently the Markov structure among the variables. We have also compared the methods using several other commonly used estimation accuracy measures, namely the quadratic loss

$$\text{QL} = \text{tr}(\Sigma^{-1}\hat{\Sigma} - I)^2, \tag{14}$$

where $\hat{\Sigma} = \hat{C}^{-1}$, and the matrix ℓ_1 loss

$$L_1 = \|\Sigma - \hat{\Sigma}\|_{\ell_1}, \tag{15}$$

where $\|M\|_{\ell_1} = \sup\{\|Mx\|_{\ell_1} : \|x\|_{\ell_1} = 1\}$ and $\|x\|_{\ell_1}$ is ℓ_1 norm of vector x. The results are similar to that of the Kullback-Leibler loss and therefore omitted here for brevity.

We have also conducted simulation on a couple of other models considered by Huang et al. (2006) and Bickel and Levina (2008) respectively, which are

Model 4. Covariance matrix such that $\sigma_{ij} = \rho^{|i-j|},\ 1 \le i, j \le p$ with $\rho = 0.5$.

Model 5. $C = T'D^{-1}T$, where $D = 0.01 \times I$, and $T = -(\phi_{ij})$, with $\phi_{ii} = 1$, $\phi_{i+1,i} = 0.8$, and $\phi_{ij} = 0$ otherwise.

Both are AR(1) and the results are very similar to those of Model 1 and therefore omitted here.

When the observations follow Markov chains of varying lengths from variable to variable, the number of nonzero elements differs among the rows of the Cholesky factor matrix T, i.e.,

$$\phi_{ij} = 0 \text{ if and only if } i - j > h_i$$

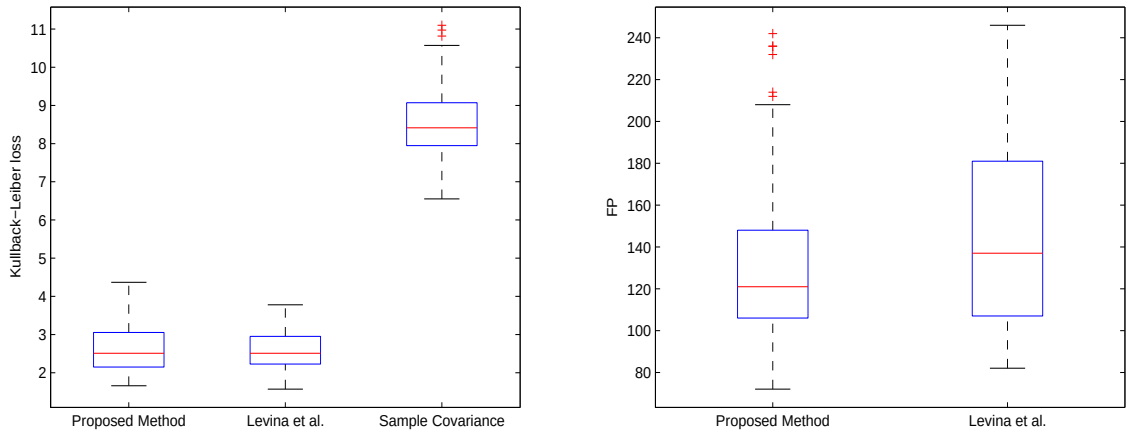
Table 1: Simulation results for the three models with temporal orders. Averages and standard errors are calculated from 100 runs.

		Structured graphGarrote			Bickel and Levina			Levina et al.			Huang et al.	Sample
p	Model	KL	FP	FN	KL	FP	FN	KL	FP	FN	KL	KL
	1	0.10	0.00	0.00	0.11	0.18	0.00	0.11	1.84	0	0.12	0.69
		(0.05)	(0.00)	(0.00)	(0.05)	(1.79)	(0.00)	(0.05)	(5.39)	(0.00)	(0.06)	(0.17)
10	2	0.26	0.16	0.00	0.63	64.98	0.00	0.26	0.72	0	0.60	0.68
		(0.10)	(1.59)	(0.00)	(0.17)	(8.28)	(0.00)	(0.10)	(7.16)	(0.00)	(0.13)	(0.14)
	3	0.36	2.76	0.00	0.71	6.90	7.04	0.66	53.76	0.64	0.62	0.70
		(0.11)	(6.06)	(0.00)	(0.38)	(7.95)	(7.94)	(0.21)	(10.97)	(3.13)	(0.16)	(0.18)
	1	0.31	0.00	0.00	0.32	0.58	0.00	0.32	45.52	0.00	0.34	8.37
		(0.09)	(0.00)	(0.00)	(0.11)	(5.77)	(0.00)	(0.08)	(57.72)	(0.00)	(0.10)	(0.90)
30	2	1.03	0.06	0.00	7.26	734.60	0.00	0.88	125.76	0.00	5.46	8.58
		(0.23)	(0.60)	(0.00)	(1.44)	(83.69)	(0.00)	(0.17)	(75.16)	(0.00)	(0.63)	(0.94)
	3	1.43	0.54	0.00	2.05	29.54	15.68	4.59	168.1	40.86	4.92	8.76
		(0.29)	(5.37)	(0.00)	(1.28)	(31.48)	(25.14)	(0.26)	(82.18)	(6.71)	(0.36)	(0.92)
	1	0.26	0.00	0.00	0.26	5.94	0.00	0.25	147.48	0.00	0.35	20.14
		(0.04)	(0.00)	(0.00)	(0.06)	(33.78)	(0.00)	(0.04)	(103.8)	(0)	(0.05)	(0.74)
100	2	0.81	0.00	0.00	18.38	9227.58	0.00	1.13	6930.58	0.00	13.27	20.11
		(0.10)	(0.00)	(0.00)	(1.37)	(393.22)	(0.00)	(0.16)	(731.462)	(0)	(0.35)	(0.65)
	3	1.26	0.00	0.00	1.31	400.20	0.00	8.08	7501.46	16.8	12.81	20.17
		(0.14)	(0.00)	(0.00)	(0.13)	(73.76)	(0.00)	(0.51)	(511.37)	(7.90)	(0.31)	(0.67)

for different bandwidth h_is. The banding method in Bickel and Levina (2008) may no longer be appropriate since it assumes that $h_1 = \ldots = h_p$. Levina et al. (2008) addressed this by allowing different bandwidths for different rows of T. The non-homogeneous structured graphGarrote can also overcome this problem. To illustrate, consider a model similar to that of Levina et al. (2008).

Model 6. $C = (I - \Phi)' D^{-1}(I - \Phi)$, with $D = 0.01 \times I$ and $\Phi = (\phi_{i,j})$ where $\forall j \geq 2$, $k_j \sim U(\lceil j/2 \rceil, j-1)$; $\phi_{j,j'} = 0.5$, $k_j \leq j' \leq j-1$; $\phi_{i,j} = 0$, $j' < k_j$.

Here $U(k_1, k_2)$ denotes an integer selected randomly from integer k_1 to k_2. In this simulation we take $p = 30$ as an example. To avoid poorly conditioned covariance matrix, we divided the 30 variables into two independent blocks with 15 variables each, and generated a random structure from Model 6 for each block, i.e., $(X^{(1)}, \ldots, X^{(15)})$ and $(X^{(16)}, \ldots, X^{(30)})$ each follows Model 6 but are mutually independent. We simulated samples with size $n = 100$ and compared the structured graphGarrote, the method used in Levina et al. (2008), and the sample covariance matrix. Figure 4 reports the boxplot for the Kullback-Leibler loss (KL) and the number of false positive of the inverse covariance matrix (FP) from 100 runs.



(a) Kullback-Leibler loss (KL)

(b) Number of false positives (FP)

Figure 4: Estimation comparison for Model 6.

From Figure 4, we observe that the performance of our proposed method is very similar to that of Levina et al. (2008). By capturing the structure of the true model, the structured

graphGarrote committed relatively fewer false positives. We note that the optimization problem involved in the approach of Levina et al. (2008) is not convex. An iterative procedure was developed by Levina et al. (2008) to tackle the computational challenge. Although efficient, it can still be sensitive to the choice of tuning parameters and initial values. In contrast, the proposed method is strictly convex and more stable in computation.

3.2 Spatial Structures

Next, we consider the situation when the random variables are observed on a two dimensional lattice. Three different models were used in our simulation.

Model 7 $\Sigma = I_p$.

Model 8 (Markov random field of order one) $c_{i,j} = 1$ if $d_{ij} = 0$, 0.25 if $d_{ij} = 1$, and 0 otherwise.

Model 9 (Markov random field of order two) $c_{i,j} = 1$ if $d_{ij} = 0$, 0.4 if $d_{ij} = 1$, 0.15 if $d_{ij} = 2$, and 0 otherwise.

For each model, we simulated samples of size $n = 100$ and dimension $p = 4 \times 4$, $n = 200$ and $p = 8 \times 8$, or $n = 600$ and $p = 16 \times 16$. Although in principle, all the methods described previously can be applied in these settings, none of them except for the structured graphGarrote is devised to take advantage of the spatial structure explicitly. In particular, the methods from Huang et al. (2006), Bickel and Levina (2008), and Levina et al. (2008) are all very sensitive to the ordering of the variables. To demonstrate the merits of the structured graphGarrote, we compare it with the graphLasso and graphGarrote from Yuan and Lin (2007) which do not rely on the particular ordering among variables. In particular, the graphLasso estimate is given as

$$\hat{C}^{\text{graphLasso}} = \arg\min \left[-\ln|C| + \text{trace}(C\bar{A}) + \lambda \sum_{i \neq j} |c_{ij}| \right],$$

where with the minimization is taken over all symmetric and positive definite matrices C, and $\lambda > 0$ is a tuning parameter. It is also worth pointing out that unlike the structured graphGarrote, neither graphLasso nor graphGarrote utilizes the information of the spatial

structure. Table 2 shows that being able to account for the spatial structures, the structured graphGarrote enjoys considerably improved performance over the other methods.

Table 2: Simulation results with spatial structure. Averages and standard errors are calculated from 100 runs. The proposed method compared with GraphGarrote and GraphLasso

		Structured graphGarrote			GraphGarrote			GraphLasso		
p	Model	KL	FP	FN	KL	FP	FN	KL	FP	FN
	7	0.17	0.00	0.00	0.19	0.94	0.00	0.17	0.02	0.00
		(0.06)	(0.00)	(0.00)	(0.07)	(2.44)	(0.00)	(0.06)	(0.20)	(0.00)
16	8	0.55	0.02	0.00	0.93	19.24	5.96	0.97	38.00	3.18
		(0.14)	(0.20)	(0.00)	(0.21)	(7.63)	(3.92)	(0.33)	(14.11)	(4.29)
	9	0.96	1.90	0.02	1.31	20.50	38.08	1.58	44.86	44.00
		(0.23)	(10.80)	(0.30)	(0.20)	(8.99)	(9.24)	(0.37)	(16.07)	(11.21)
	7	0.34	0.00	0.00	0.33	0.26	0.00	0.33	0.00	0.00
		(0.06)	(0.00)	(0.00)	(0.06)	(1.01)	(0.00)	(0.06)	(0.00)	(0.00)
64	8	1.76	0.08	0.00	2.45	173.97	6.56	4.43	231.93	3.50
		(0.24)	(0.80)	(0.00)	(0.45)	(51.77)	(4.36)	(0.35)	(25.93)	(2.37)
	9	4.37	0.00	0.12	6.48	204.44	190.56	9.00	519.30	246.20
		(0.53)	(0.00)	(0.47)	(0.36)	(23.94)	(15.96)	(1.61)	(155.88)	(25.49)
	7	0.64	0.00	0.00	0.43	0.22	0.00	0.43	0.00	0.00
		(0.06)	(0.00)	(0.00)	(0.03)	(1.30)	(0.00)	(0.04)	(0.00)	(0.00)
256	8	3.14	0.00	0.00	4.68	132.87	0.50	13.30	1497.28	0.26
		(0.18)	(0.00)	(0.00)	(0.22)	(20.27)	(0.87)	(0.31)	(65.15)	(0.79)
	9	11.34	0.10	0.12	31.04	3.50	69.60	20.70	1583.97	8.90
		(0.75)	(0.99)	(0.47)	(1.13)	(2.88)	(10.48)	(4.03)	(1127.71)	(4.46)

3.3 When $p > n$

The proposed estimates of the inverse covariance matrix are shrunken version of $\bar{A}^{-1}$. As we pointed out earlier, other initial estimate of Σ can also be employed. For example, we can consider the MLE of C with $c_{ij} = 0$ for $d_{ij} > H$ and a pre-specified bandwidth H that is large but much smaller than p. This is particularly appealing when $p \geq n$ and the inverse

of $\bar{A}$ does not exist. To illustrate, we re-examine Models 7, 8 and 9, but with the following combinations of sample size and dimensionality: $(n, p) = (10, 4 \times 4)$, $(n, p) = (50, 8 \times 8)$, and $(n, p) = (200, 16 \times 16)$. We used MLE with $H = 3$ as the initial estimator and the performance of the structured graphGarrote is summarized in Table 3. Compared with GraphGarrote and GraphLasso, the structured graphGarrote enjoys better performance.

Table 3: Simulation results with spatial structure in the case of $p > n$. Averages and standard errors are calculated from 100 runs. The proposed method compared with GraphGarrote and GraphLasso

		Structured graphGarrote			GraphGarrote			GraphLasso		
p	Model	KL	FP	FN	KL	FP	FN	KL	FP	FN
	7	3.43	0.00	0.00	5.20	6.80	0.00	3.24	2.42	0.00
		(1.87)	(0.00)	(0.00)	(3.42)	(5.82)	(0.00)	(2.60)	(11.12)	(0.00)
16	8	5.75	7.32	0.06	14.58	16.40	39.72	9.50	25.26	38.94
		(2.10)	(20.88)	(0.34)	(12.02)	(17.41)	(8.91)	(6.55)	(39.83)	(12.99)
	9	6.84	0.00	57.90	17.87	6.94	95.58	15.54	27.04	84.53
		(2.37)	(0.00)	(24.10)	(17.14)	(7.24)	(20.71)	(10.96)	(32.33)	(34.16)
	7	1.46	0.00	0.00	1.53	2.02	0.00	1.44	0.12	0.00
		(0.30)	(0.00)	(0.00)	(0.31)	(4.89)	(0.00)	(0.28)	(0.84)	(0.00)
64	8	8.24	0.00	0.06	13.35	48.22	126.32	13.79	150.00	96.58
		(1.01)	(0.00)	(0.34)	(0.68)	(12.56)	(5.29)	(2.91)	(79.72)	(35.71)
	9	11.94	0.00	19.88	17.57	36.32	425.16	25.48	5.24	598.30
		(1.21)	(0.00)	(84.46)	(0.72)	(8.18)	(6.22)	(0.85)	(16.95)	(29.39)
	7	1.32	0.00	0.00	1.34	0.36	0.00	1.30	0.06	0.00
		(0.14)	(0.00)	(0.00)	(0.14)	(1.42)	(0.00)	(0.12)	(0.34)	(0.00)
256	8	25.78	0.00	0.00	49.73	97.18	480.36	44.33	1015.73	134.06
		(0.82)	(0.00)	(0.00)	(0.27)	(18.18)	(0.82)	(1.85)	(101.21)	(13.81)
	9	29.82	0.00	0.06	62.11	0.00	1828.04	55.81	792.20	201.94
		(0.38)	(0.00)	(0.34)	(0.26)	(0.00)	(0.28)	(2.24)	(98.49)	(14.61)

It is worth pointing out that the success of this strategy hinges on the assumption that the true covariance structure follows a Markov model of order less than H. To this end, it is beneficial to take a relatively large H. It is also worth noting that one needs not to seek

an optimal choice of H, which can be quite challenging (see e.g., Bickel and Levina, 2008), since it is only used to construct the initial estimator. The final estimate can be much more sparse. Alternative choices of the initial estimate of C also include the inverse of a linear combination of $\bar{A}$ and the identity matrix. The performance of these choices are however generally unclear.

4 Handwritten Digit Data

To further illustrate the merits of the proposed method, we apply the proposed structured covariance matrix estimation to a real data example. The handwritten digit data (LeCun et al., 1990) come from automatic reading of handwritten zip codes appeared on envelopes by the United States Postal Service. Each handwritten digit is converted into a 16 by 16 grayscale image after some processing. The intensity values lie in the range from -1 and 1. Images as such can often be modeled as a Markov random field of a relatively small order. The proposed methods exploit the sparsity in the inverse covariance matrix so that the estimate conforms with models of such Markov structure.

A common goal of analyzing the handwritten digits is to distinguish images representing different digits. To this end, we consider applying linear discriminant analysis (LDA) with covariance matrix estimated using the proposed method as well as the sample covariance matrix to the data. The main purpose of this exercise is to demonstrate how the structured graphGarrote can lead to improved classification performance of LDA. For illustrative purpose, we focus on digits 6 and 9 which include a total of 1308 images in the data set. 600 images were randomly selected as the training set, and the rest were used as the test set. We repeated the experiment 100 times. The boxplot of the testing error is given in Figure 5. It shows that the covariance matrix estimated from the proposed method indeed leads to lower misclassification error.

To gain further insights, we also examine for a given pixel, how often its partial correlation with other pixels is estimated by a nonzero value. The (i, j) panel of Figure 6 corresponds to the partial correlation between the intensity at the (i, j)th pixel and other pixels. A darker cell indicating higher frequency. A few pixels around the four corners are removed from our analysis because their intensity values remain constant in the data set. A more detailed look

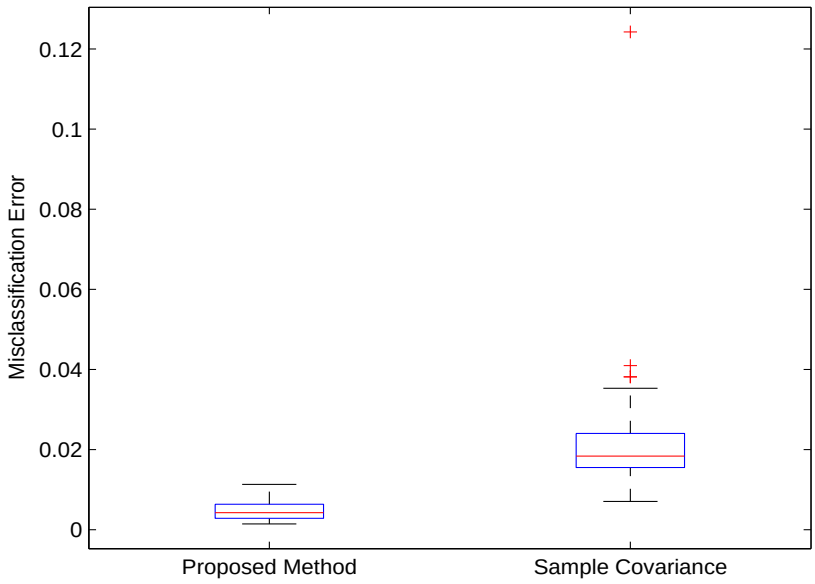


Figure 5: The boxplot of misclassification error on the test set for 100 replications.

at several selected pixels is given in Figure 7. From Figures 6 and 7, we observe that the handwritten digit images may be modeled by a Markov random field of order 4 or 5.

5 Discussions

In this paper, we have developed methods for estimating high dimensional Gaussian covariance matrix when the random variables are observed with temporal or spatial structures. By directly exploiting sparsity of the inverse covariance matrix, the estimate obeys certain Markov models. The proposed method can be formulated as a semi-definite program and efficiently computed using standard software.

Although we focused on the temporal and spatial structures, the method can be easily extended to more complicated situations such as spatial-temporal structures. More generally, our method can be applied in situations where a similarity/dissimilarity measure of the domain from which the variables are observed is available.

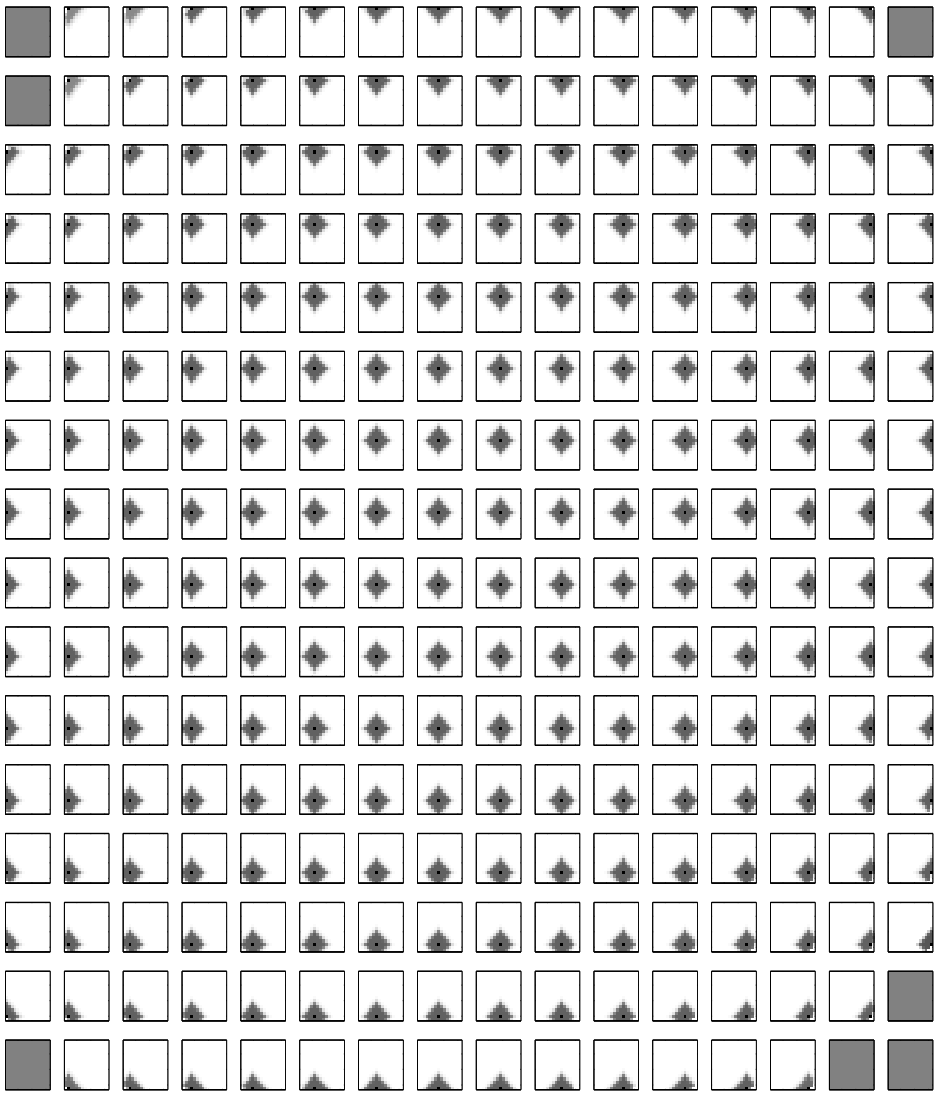


Figure 6: Heatmap plots of percentage of the nonzeros at each location in the estimated inverse covariance matrix from handwritten digit data. Black represents 100%, white 0%.

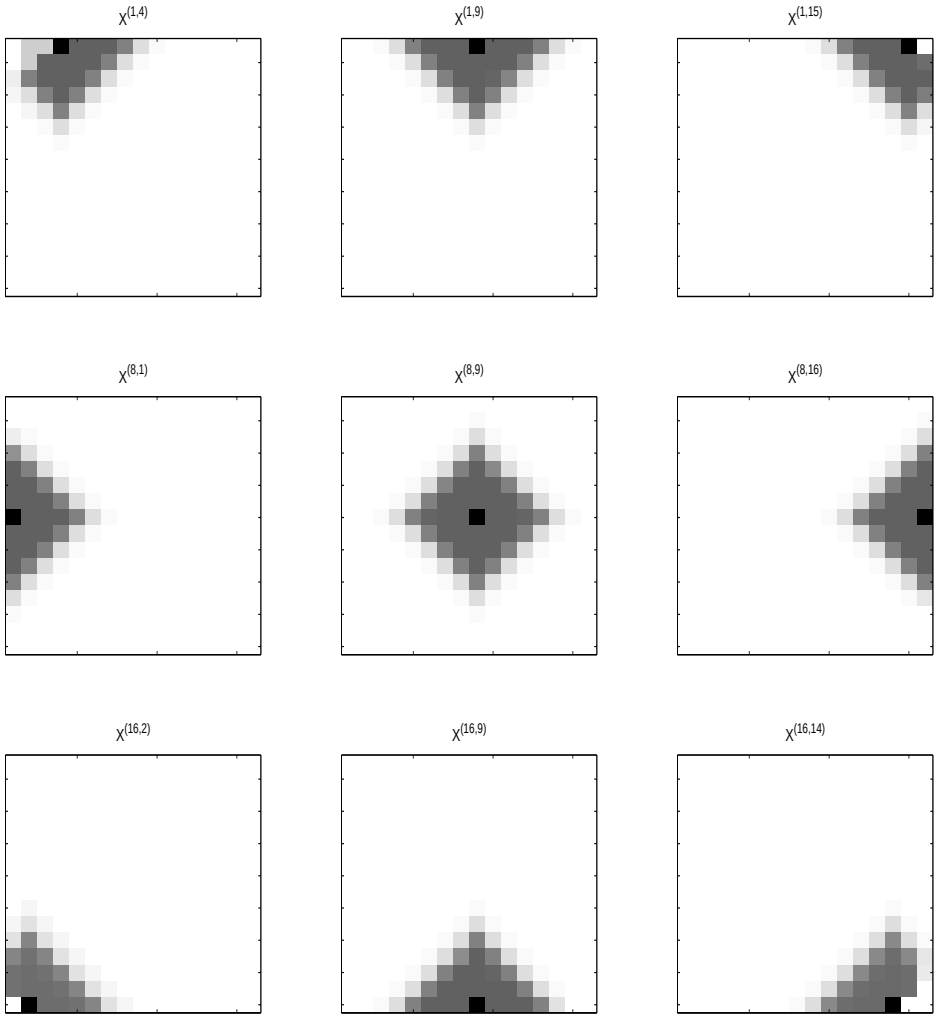


Figure 7: Some of heatmap plots of percentage of the nonzeros at each location in the estimated inverse covariance matrix from handwritten digit data. Black represents 100%, white 0%.

References

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199-227.

Breiman, L. (1995), Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373-384.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008), First-order methods for sparse covariance selection, *SIAM Journal on Matrix Analysis and it Applications*, 30, 56-66.

Dempster, A. (1972). Covariance selection. *Biometrics*, 28:157–75.

Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Steins loss. *Ann. Statist.*, 13(4):1581–1591.

Friedman, J., Hastie, T. and Tibshirani, T. (2008), Sparse inverse covariance estimation with the graphical lasso, *Biostatistics*, **9**, 432-441.

Haff, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.*, 8(3):586–597.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. *in* Touretzky, D. (ed.), *Advances in Neural Information Processing Systems*, Vol. 2, Morgan Kaufman, Denver, CO.

Ledoit, O. and Wolf, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88:365–411.

Levina, E., Rothman, A. J., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Annals of Applied Statistics*, 2(1):245-263.

Perron, F. (1992). Minimax estimators of a covariance matrix. *Journal of Multivariate Analysis*, 43:16–28.

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

Pourahmadi, M. (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika*, 87:425–435.

Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008), Sparse permutation invariant covariance estimation, *Electronic Journal of Statistics*, **2**, 494-515.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.*, 97(460):1141–1153.

Stein, C. (1977). Lectures on the theory of estimation of many parameters. In *Studies in the Statistical Theory of Estimation, Part I* (I. A. Ibragrniov and M. S. Nikulin, eds.). *Proc. Scientific Seminars Steklov Institute, Leningrad Division*, 74:4–65. (In Russian.)

Tütüncü, R. H., Toh, K. C., and Todd, M. J. (2003). Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming*, 95(2):189–217.

Winkler, W. E. (2006), *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*, New York: Springer.

Wong, F., Carter, C., and Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, 90:809–830.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90:831–844.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.