

**JOINT ESTIMATION OF COVARIANCE
MATRIX VIA CHOLESKY DECOMPOSITION**

JIANG XIAOJUN

NATIONAL UNIVERSITY OF SINGAPORE

2012

**JOINT ESTIMATION OF COVARIANCE
MATRIX VIA CHOLESKY DECOMPOSITION**

JIANG XIAOJUN

(B.Sc. Peking University of China)

**A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF STATISTICS AND APPLIED
PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE**

2012

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my gratitude to my supervisor associate professor Leng Chenlei. He is such a nice mentor not only because of his brilliant ideas but also his kindness to his students. I can not finish this thesis without his kind guidance. It is my luck to have him as my supervisor. Special acknowledgement also goes to the faculties and staff of DSAP. Anytime I encountered difficulties and tried to seek help from them, I was always warmly welcomed. I also have to express my thanks to my colleges. You make my four years study in DSAP a pleasant time.

CONTENTS

Acknowledgements	ii
Summary	v
List of Notations	vii
List of Tables	viii
List of Figures	ix
Chapter 1 Introduction	1
1.1 Cholesky Decomposition	3
1.2 Penalized Method	6
1.3 Penalties with Group Effect	15

Chapter 2 Literature Review	21
2.1 Direct Thresholding Approaches	22
2.2 Penalized Approaches	26
2.3 Methods Based on Ordered Data	29
2.4 Motivation and Significance	31
Chapter 3 Model Description	37
3.1 Penalized Joint Normal Likelihood Function	38
3.2 IL-JMEC Method	44
3.3 GL-JMEC Method	46
3.4 Computation Issue	51
3.5 Main Results	56
Chapter 4 Simulation Results	60
4.1 Simulation Settings	60
4.2 Simulation with Respect to Different Data Sets	62
4.3 A Real Data Set Analysis	76
Chapter 5 Conclusion	82
Chapter A Appendix	86
A.1 Three Lemmas	86
A.2 Proof of Theorems	90
Bibliography	105

ABSTRACT

Covariance matrix estimation is a very important topic in statistics. The estimate is needed in various aspects of statistics. In this research, we focus on jointly estimating covariance matrix and precision matrix for grouped data with natural order via Cholesky decomposition. We treat autoregressive parameters at the same position in different groups as a set and impose penalty functions with group effect to these parameters together. A sparse l_∞ penalty and a sparse group LASSO penalty are used in our methods. Both penalties may produce common zeros in the autoregressive matrices for different groups which reveal the common relationships of variables between groups. When data structures in different groups are close, our approaches can do better than separate estimation approaches by providing more accurate covariance and precision matrix estimates and they are guaranteed to be positive definite. A coordinate decent algorithm is used in the optimization

procedure and convergence rates have been established in this study. We can prove that under some regularity conditions, our penalized estimators are consistent. In the simulation part, we show their good performance by comparing our methods with the separated estimation methods. An application to classify cattle from two treatment groups based on their weights is also included.

LIST OF NOTATIONS

$A \otimes B$	Kronecker product of two matrices A and B
$ A _1$	l_1 norm of matrix A
$\text{Vec}(A)$	The vectorization of matrix A
$\ A\ $	The singular value of matrix A which equals the square root of maximal eigenvalue of AA'
$\ A\ _F$	Frobenius Norm of matrix A which equals $\sqrt{\text{tr}AA'}$
$U(a, b)$	Uniform distribution on interval (a, b)
$I(A)$	Indicator function on event A
$\langle \alpha, \beta \rangle$	Inner product of vectors α and β

List of Tables

Table 4.1	Simulation result when sample size is growing.	69
Table 4.2	Simulation result when number of groups is growing while the autoregressive matrices are identity matrix.	71
Table 4.3	Simulation result when number of groups is growing while autoregressive matrices are randomly generated.	72
Table 4.4	Simulation result when data have different degrees of similarity.	74

Table 4.5	Simulation result when when autoregressive matrices have many non zero elements.	75
Table 4.6	Performance of discrimination study for the cattle weight data.	78

List of Figures

Figure 3.1	Minimizer of $\phi_1^2 - (a + b)\phi_1 + \phi_2^2 - a\phi_2 + \lambda \phi _1 + \beta\ \phi\ _\infty$. . .	45
Figure 3.2	Contour graph for sparse group LASSO (left) and sparse l_∞ LASSO (right).	50
Figure 4.1	Ratio of Frobenius loss and Operator loss in example 1. . . .	70
Figure 4.2	Ratio of Frobenius loss and Operator loss in example 3. . . .	73
Figure 4.3	Trend of weights for the two groups of cattle.	77

CHAPTER 1

Introduction

Covariance matrix and precision matrix estimation are very important in statistics. The covariance matrix and its inverse are widely used in statistics such as discrimination analysis and principle component analysis. In finance, the estimator of the covariance matrix of a collection of assets is required in order to achieve an optimal portfolio. In Gaussian graphical modeling, a sparse precision matrix is uniquely corresponding to an undirected graph that represents the conditional independent relationships of the target variables (see Pearl 2000).

Standard estimators of covariance matrix and precision matrix are the sample

covariance matrix and its inverse multiplies a scale parameter. These two estimators are proved to be unbiased and consistent. Moreover, they are very easy to calculate. Due to these properties, they are widely used in statistics. In recently years, alternative estimators of the covariance matrix and the precision matrix have been proposed due to high dimensional data requirement and also the requirements for special structures of the variables. These new methods aimed to eliminate the disadvantages of sample covariance matrix when the dimension is large (see Johnstone 2001 and Bai 1993) and to provide structured and interpretable estimators. Penalized estimation methods and thresholding methods (Ledoit and Wolf 2004; Huang et al. 2006; Lan and Fan 2009; Rothman 2008 and so on) have made great contributions to achieve these goals.

Most researches so far focused on estimating single covariance matrix or precision matrix. However, in some cases, it is much more valuable to jointly estimate them if grouped data were observed from similar categories. For instance, we consider gene data that describes different types of the same diseases or observations of patients from different treatment groups. It is reasonable to assume that data from different groups share similar structures, and it is obviously a waste of information if we estimate the covariance matrices separately because the similarity of data is simply ignored. Meanwhile, it is not feasible to combine the data all together and estimate a single covariance matrix while treating them as a single

group. A possible way to employ the information of similarity between different groups is to jointly estimate the matrices. We can expect that estimation accuracy may be increased if the joint estimation method is employed. In this research, in order to achieve the joint estimation objective and keep our estimates positive definite, grouped penalization approaches based on Cholesky decomposition are investigated.

In the subsequent sections, background knowledge about Cholesky decomposition and penalty approaches will be reviewed. These are the key tools in our new methods. In chapter 2, the development of matrix estimation approaches will be reviewed.

1.1 Cholesky Decomposition

Using Cholesky decomposition to estimate the covariance and the precision matrix was firstly introduced by Pourahmadi (1999). A joint mean-covariance model has been proposed to estimate the autoregressive parameters of the covariance matrix in that approach. After that, this decomposition was widely used in longitudinal study and matrix estimation (see Pourahmadi 2000, Huang 2006, Rothman 2008, Shojaie and Michailidis 2010, Rothman et al. 2010, Leng et al. 2010).

The Cholesky decomposition illustrates that for every positive definite matrix Σ , there exists a unique lower triangular matrix R , such that

$$\Sigma = RR', \quad (1.1)$$

where the diagonal entries of R are all nonnegative. The elements $r_{11}, r_{21}, \dots, r_{p1}, r_{22}, r_{32}, \dots, r_{pp}$ of matrix R can be obtained consequently. Assume the diagonal entries of matrix R are $\sigma_1, \sigma_2, \dots, \sigma_p$ and matrix D is a diagonal matrix with diagonal entries $\sigma_1^2, \sigma_2^2, \dots, \sigma_p^2$ and matrix $T = D^{1/2}R^{-1}$, then the above decomposition can be reorganized as the following modified version

$$T\Sigma T' = D. \quad (1.2)$$

In this modified decomposition, matrix T is a lower triangular matrix with ones on its diagonal while matrix D is a diagonal matrix. A charming advantage of the Cholesky decomposition is that the parameters in matrix T is free to constraints and the only requirement for matrix D is that its diagonal elements are all positive. Moreover, The modified Cholesky decomposition has a natural statistical explanation (see Pourahmadi 1999).

Following the argument in Pourahmadi (1999), the elements in matrix T can be expressed as the successive regression coefficients of variables regressed on their predecessors and the elements in matrix D can be expressed as the regression error variances. If we further assume the variables have a multivariate normal

distribution, the elements in matrix T can be explained as the coefficients of the best prediction of one variable based on its predecessors.

To be more precise, assume that we have a random vector $Y = (y_1, y_2, \dots, y_p)'$. Here Y follows a multivariate normal distribution $N(\mu, \Sigma)$ where we assume $\mu = 0$ for simplicity. We consider the distribution of variable y_k conditional on its predecessors y_1, y_2, \dots, y_{k-1} . This distribution can be easily found in the book of Anderson (2003)

$$\mathcal{F}(y_k|Y_{(k)}) \sim N(\sigma'_{(k)}\Sigma_{(k)}^{-1}Y_{(k)}, \sigma_{kk} - \sigma'_{(k)}\Sigma_{(k)}^{-1}\sigma_{(k)}). \quad (1.3)$$

Here, we denote $Y_{(k)} = (y_1, y_2, \dots, y_{k-1})'$, $\Sigma_{(k)}$ the $k-1$ dimensional main submatrix of Σ , $\sigma_{(k)}$ as a vector that contains the first $k-1$ elements of the k th column of matrix Σ and σ_{kk} as the kk th element of matrix Σ .

Denote $\tilde{y}_k = E(y_k|Y_{(k)}) = \sigma'_{(k)}\Sigma_{(k)}^{-1}Y_{(k)}$ and $\tilde{\epsilon}_k$ as the residual term $y_k - \tilde{y}_k$. Obviously, $\tilde{y}_1 = 0$ and $\tilde{\epsilon}_1 = y_1$. Since $E(y_k|Y_{(k)})$ can be treated as the projection of y_k on the σ -field $\sigma(y_1, y_2, \dots, y_{k-1})$, it is straightforward to conclude that $\tilde{\epsilon}_k$ is independent with $\tilde{\epsilon}_1, \tilde{\epsilon}_2, \dots, \tilde{\epsilon}_{k-1}$.

Denote the k th row of matrix L by L'_k in which the first $k-1$ elements satisfy $(\phi_{k1}, \phi_{k2}, \dots, \phi_{k(k-1)}) = -\sigma'_{(k)}\Sigma_{(k)}^{-1}$, $\phi_{kk} = 1$ and $\phi_{kl} = 0$ for $l > k$. This implies $L'_k Y = \hat{\epsilon}_k$ ($k = 1, \dots, p$). Write these p equations into matrix form, we have

$$LY = \hat{\epsilon}, \quad (1.4)$$

where $\hat{\epsilon} = (\hat{\epsilon}_1, \dots, \hat{\epsilon}_p)'$. We denote $\sigma_k^2 = \sigma_{kk} - \sigma'_{(k)} \Sigma_{(k)}^{-1} \sigma_{(k)}$. Because $\hat{\epsilon}_k$ are independent with each other, taking variance on both side of (1.4), we have

$$L\Sigma L' = D.$$

The best prediction of y_k based on $Y_{(k)}$ is $\sigma'_{(k)} \Sigma_{(k)}^{-1} Y_{(k)}$, thus elements in the k th row of T can be explained as the coefficients in the best prediction of y_k based on variables y_1, \dots, y_{k-1} with opposite signs.

If we relax the assumption of multivariate normality of Y , $\sigma'_{(k)} \Sigma_{(k)}^{-1} Y_{(k)}$ is still the best linear prediction of y_k with least square error. Thus the elements in the k th row of T are the least square regression coefficients of variable y_k regressed on variables y_1, \dots, y_{k-1} . This explanation makes the autoregressive parameters meaningful. It reminds us that we may impose some special structures on the parameters if we have prior information about the data.

1.2 Penalized Method

In the traditional methods, parameters are estimated based on some meaningful loss functions $L(\theta)$, mostly by minimizing these target loss functions. Likelihood functions constitute a widely used collection of the loss functions, for instance, the popular negative log likelihood function of multivariate normal $\text{tr}(\Sigma^{-1}S) +$

$\log |\Sigma|$. Minimizing this negative log likelihood function will lead to the maximum likelihood estimator of the covariance matrix. In some other applications, the loss functions can be also chosen as norms, for example, l_1 or l_2 norm. The widely used linear regression is an application of the l_2 norm loss. Minimizing the squared l_2 norm of $Y - X\beta$ leads to a linear model of variable y based on variables x_1, x_2, \dots, x_p with smallest squared fitting error. Here Y is the vector of observations of variable y and X is the design matrix for the explanatory variables x_1, x_2, \dots, x_p .

In these classical estimation approaches, the parameters or covariates are all included in the model. For example, the standard linear regression model always contains all the explanatory variables that we have observed. However, as we know, including many covariates leads to low bias but high prediction variance or say over fitting. This over fitting phenomenon can be explained in linear regression problems as follows. The coefficient of determination R^2 is always decreasing when one adds more and more explanatory variables to the model. However, the prediction variance can be very high. In order to reduce the prediction variance, one can sacrifice a little bias so as to decrease the prediction variance by making a tradeoff between them.

A natural idea is to make some special assumption on the data. For instance, there are a lot of small coefficients or there are a lot of unimportant explanation variables. Based on this kind of prior information, more adaptive models can be

proposed to investigate the data.

Penalization methods were introduced as a simple and straightforward way to achieve this objective. The idea is similar to the AIC and BIC methods. A trade-off is made between the fitness and the prediction accuracy by adding a penalty function $p_\lambda(\theta)$ to the loss function L_θ and minimizing the new objective function

$$L_\theta + p_\lambda(\theta), \tag{1.5}$$

instead of the original loss function L_θ . In this new method, the loss function L_θ controls the fitness of the model and $p_\lambda(\theta)$ can be used to set a constraint to the complexity (the number of nonzero parameters included in the model or say sparsity) or structure of the model.

Penalization approaches also have close relationship with Bayesian method (see Zhao et al. 2009). Particularly, if we assume the loss function in (1.5) is a log likelihood function and the penalty function $p_\lambda(\theta)$ is a log prior density function of parameters θ , then the objective function (1.5) can be explained as the log posterior density function of θ conditional on the observations, for example, the ridge regression when we choose $p_\lambda(\theta) = \lambda \|\theta\|_2^2$ is the same as the Bayesian approach where a multivariate normal prior $N(0, \frac{1}{2\lambda} I_p)$ is imposed to the parameters θ .

One great advantage of the penalized approaches comparing to Bayesian approaches is that a more flexible function $p_\lambda(\theta)$ can be used to constrain the parameters in penalized approaches. However, in Bayesian approaches, a proper prior density function is needed. A carefully chosen penalty function can make a tradeoff between the bias and the prediction accuracy. It may also introduce some desired properties or structures to the model. In order to introduce the penalized methods, we use the squared l_2 norm loss function $\|Y - X\beta\|_2^2$ as an example and denote the ordinary least square estimator of the coefficients by θ_{ols} . For simplicity, we assume $X^T X = I_p$.

Frank and Friedman (1993) introduced the bridge regression method in which they add a penalty term $p_\lambda(\theta) = \lambda\|\theta\|_q^q$ to the loss function $\|Y - X\theta\|_2^2$. Instead of using the ordinary least square method, they minimized the following function

$$\theta_{bridge} = \operatorname{argmin}_\theta [L(\theta) + p_\lambda(\theta)] = \operatorname{argmin}_\theta [\|Y - X\theta\|_2^2 + \lambda\|\theta\|_q^q].$$

Here q is a positive constant and λ is the threshold parameter. When $q > 1$, bridge regression method shrinks the parameters θ and reduces variability. When $q \leq 1$, bridge regression method provides sparse estimates of the parameter θ . Particularly, bridge regression is also called LASSO when the constant q is set to 1 (see Tibshirani 1996). When q is set to 2, bridge regression is also known as ridge regression. Overall, bridge regression provides a more stable model compared to the ordinary regression method while bias is increased.

In recent years, especially along with the development of large dimensional data set, penalty approaches which lead to sparse estimators gained more and more attention. This is mainly because models from traditional approaches are relatively complicated since nearly all of the parameters are nonzero. For instance, the sample covariance matrix which is obtained from gaussian likelihood function has no zero element in it, which means all variables must be marginally correlated. All the coefficients in ordinary least square regression method are nonzero means all variables are important in predicting the target variable. This is confusing and hard to interpret.

Consequently, investigating simple models that only include the important variables becomes a popular research area in statistics. Penalizing methods which is capable to provide sparse estimators have been extensively investigated. They can efficiently reduce the complexity of the underling model. In Fan and Li (2001), they listed three desired properties of an ideal penalty function which are

- 1: Unbiasedness: The resulting estimator should be nearly unbiased and the large coefficients should be only slightly shrunk in order to guarantee the accuracy.

- 2: Sparsity: The solution must be sparse that provides a more interpretable model.

- 3: Continuity: The solution is continuous with respect to the data in order to

avoid instability.

It has to be noted that most of the penalty functions can not satisfy all these three requirements especially convex penalty functions. Convex penalty functions always shrink small coefficients as well as large coefficients. Nonconvex penalty functions may meet all these three requirements. However, nonconvex penalties always lead to computation difficulties and it is hard to find the global minimizer.

A very natural shrinkage approach called hard thresholding method was mentioned in the Antoniadis (1997). The solution for the least square linear regression problem with a hard threshold penalty term is

$$\hat{\theta}_{hard} = \hat{\theta}_{ols} I(|\hat{\theta}_{ols}| > \lambda). \quad (1.6)$$

The corresponding penalty function is

$$p_{\lambda}(\theta) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda). \quad (1.7)$$

Threshold parameter λ is a positive constant which was chosen by carefully balancing sparsity and bias. It directly shrinks small coefficients to zero and keep the large coefficients. However, the solution is not continuous with respect to data. This makes the resulting model sensitive to the observations.

In Tibshirani (1996), the so called LASSO has been proposed. This method gives a simple and straightforward way to achieve sparse models in regression

problems. The penalty function is

$$p_\lambda(\theta) = \lambda|\theta|_1, \quad (1.8)$$

which is a special case of the bridge regression. When $q = 1$, the bridge regression method is the same as LASSO.

The LASSO approach is to estimate the coefficients by minimizing the objective function

$$\|Y - X\theta\|_2^2 + \lambda|\theta|_1.$$

Recall that we assumed the design matrix X is orthogonal, then the LASSO estimate has following formula

$$\hat{\theta}_{lasso} = \text{sign}(\hat{\theta}_{ols})(|\hat{\theta}_{ols}| - \lambda)_+. \quad (1.9)$$

This formula for orthogonal design case shows some insights of the LASSO method that this method can shrink small coefficients to zero and provide a sparse solution. The solution is continuous with respect to data and also continuous with respect to the threshold parameter λ . The LASSO method performs well when the coefficients are sparse while the ridge regression method is well performed when there are a lot of small coefficients. That's because the ridge regression only shrinks the coefficients towards zero while the LASSO algorithm is a thresholding approach which shrinks some coefficients to exact zero.

Efron et al. (2004) proposed the so called LARS algorithm which is a very important work and it can cover both LASSO algorithm and Forward Stagewise selection method. The solution path of LASSO can be obtained efficiently by a simple modification of LARS. Moreover, The LARS algorithm gives a geometrical explanation and provides researchers with further understanding of LASSO. In the paper of Rosset et al. (2008), they also proposed the solution path for the l_1 penalized approaches but with more general loss functions. The loss functions were extended to the class of differentiable and piecewise quadratic functions with respect to the response variable y and the term $x_i^T \theta$. These researches made important contributions to LASSO since one can efficiently calculate the whole solution path for different λ .

However, as a convex penalty function, there is also a problem with LASSO. The LASSO solution for the orthogonal design case which is presented in (1.9) reminds us that the LASSO algorithm also shrinks the large coefficients. This effect leads to bias and affects the prediction accuracy.

In order to eliminate the disadvantage of LASSO and try to satisfy the three conditions mentioned in Fan and Li (2001), in that paper, the authors proposed the SCAD penalty function which is a nonconvex function. The solution for SCAD is continuous with respect to data and retains the large coefficients. When the design

matrix is orthogonal, the SCAD penalized solution is

$$\begin{cases} \text{sign}(\hat{\theta}_{ols})(|\hat{\theta}_{ols}| - \lambda)_+, & |\hat{\theta}_{ols}| \leq 2\lambda, \\ \{(a-1)\hat{\theta}_{ols} - \text{sign}(\hat{\theta}_{ols})a\lambda\}/(a-2), & 2\lambda \leq |\hat{\theta}_{ols}| \leq a\lambda, \\ \hat{\theta}_{ols}, & |\hat{\theta}_{ols}| \geq a\lambda. \end{cases}$$

The corresponding penalty function is relatively complex, but the first order derivative of SCAD penalty function has an explicit form

$$p'_\lambda(\theta) = \lambda\{I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda}I(|\theta| > \lambda)\}. \quad (1.10)$$

This penalty function can provide sparse estimators of the coefficients by shrinking the small coefficients while the large coefficients remain the same. It can be treated as a combination of LASSO and hard thresholding method. Fan and Li (2001) showed that this penalty function satisfies the three requirements which are previously mentioned and also showed that this penalty has the so called oracle property which means this penalized method can perform as good as the zero coefficients are already known. However, the SCAD penalty function is not convex. This may lead to computation problem.

Zou (2006) proved that the LASSO algorithm does not satisfy the oracle property. Alternatively, he proposed an adaptive LASSO method. Instead of penalizing

each coefficient equally, the adaptive method penalizes each coefficient with a particular weight. The penalty function is

$$p_\lambda(\theta) = \sum_{k=1}^p \lambda w_k |\theta_k|. \quad (1.11)$$

Suppose $\tilde{\theta}$ is a root n consistent estimator of the coefficients, for example, the ordinary least square estimator $\hat{\theta}_{ols}$, then we can choose the weight vector ω as $1/\tilde{\theta}^\gamma$. In that paper, it was proved that this adaptive penalized estimator is consistent and enjoys oracle property. But it has to be noted that the weights must be based on a consistent estimator of the coefficients θ . In the regression case, the natural root n consistent estimator of θ is not available when the dimension is large. One have to find another root n consistent estimator to implement the adaptive LASSO.

1.3 Penalties with Group Effect

In Section 1.2, the penalty functions penalize parameters individually. However, in some applications, one may be interested in penalty functions that have group effect which penalize the parameters together. With the group effect of the penalty functions, one may achieve desired structure of the variables, for instance, making the variables close or shrinking them towards zero together.

Tibshirani et al. (2005) proposed a fusion LASSO method. This fusion LASSO not only penalizes the coefficients themselves, it also penalizes the successive differences of the coefficients. The fused LASSO penalty function is

$$p_{\lambda}(\theta) = \lambda_1 \sum_{j=1}^p |\theta_j| + \lambda_2 \sum_{j=2}^p |\theta_j - \theta_{j-1}|. \quad (1.12)$$

The fused LASSO estimates are obtained as

$$\hat{\beta}_{fusion} = \operatorname{argmin} \|Y - X\theta\|_2^2 + \lambda_1 \sum_{j=1}^p |\theta_j| + \lambda_2 \sum_{j=2}^p |\theta_j - \theta_{j-1}|. \quad (1.13)$$

This penalization method shrinks some of the coefficients towards zero and also shrinks the differences of successive variables towards zero. It provides sparse estimates for the coefficients and make some of the successive coefficients to be exactly the same. This property is interesting when the effects of the explanatory variables can be divided into several levels. A similar pairwise fusion LASSO approach was proposed in Petry et al. (2011).

In Bondell and Reich (2008), they proposed another penalization method which is called OSCAR. The penalty function was chosen as a combination of the l_1 norm and a pairwise l_{∞} norm. The objective function can be presented as

$$\hat{\beta}_{oscar} = \operatorname{argmin}_{\beta} \|Y - X\theta\|_2^2 + \lambda \sum_{k=1}^p |\theta_k| + \beta \sum_{l < k} \max(|\theta_l|, |\theta_k|).$$

The effect of this penalty function is very similar to the above fusion LASSO, but the order assumption for the variables is not required.

The so called elastic net was proposed by Zou and Hastie (2005). The elastic net estimator β_{en} is defined as

$$\hat{\beta}_{en} = \operatorname{argmin}_{\beta} \|Y - X\beta\|_2^2, \quad \text{subject to } (1 - \alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \leq t.$$

The penalty term is a convex combination of LASSO penalty and ridge penalty. It enjoys the sparsity property of LASSO. Moreover, it also benefits from the good property of ridge regression that the bias is low comparing to LASSO approach when the true model has many small coefficients. The combination of these two also tends to have a group effect on variable selection that highly correlated variables tend to be in or out of the model together.

Some other penalty functions with group effects have been investigated in order to meet some special requirements in multi-ANOVA problems. In the multi-ANOVA problems, factors can be a combination of measures and may have several levels. The main goal of multi-ANOVA is often to select the important factors and to identify the level of importance of variables within the factor. Suppose there are J factors and the j th factor has coefficients θ_j which is a p_j dimensional vector. The corresponding design matrix for the j th factor is X_j and the response is Y . In order to find the estimates of coefficients θ_j ($j=1, 2, \dots, J$), one can fit a linear regression model and minimize the objective function

$$\hat{\beta} = \operatorname{argmin}_{\beta} \|Y - \sum_{j=1}^J X_j \theta_j\|_2^2. \quad (1.14)$$

It is reasonable to assume that some factors are not important in the model which means that some of the coefficient vectors θ_j must be 0. Meanwhile, for the important factors, the variables in the same group may perform differently. Under this concern, Yuan and Lin (2004) introduced the group LASSO algorithm. They imposed a penalty term

$$p_\lambda(\beta) = \lambda \sum_{j=1}^J \|\theta_j\|_{(K_j)}$$

to the objective function. Here $\|\theta_j\|_{(K_j)} = (\theta_j^T K_j \theta_j)^{1/2}$ where K_j is a kernel matrix which was set to $p_j I_j$ in their paper. One important feature of this group LASSO is that it can select important factors and set coefficients in unimportant factors to be all zero. A group LARS algorithm is also investigated in their paper. However, different from the relationship of LASSO and LARS, group LARS can not reveal the solution path of group LASSO (the solution path of group LASSO is not piecewise linear).

In group LASSO, the coefficients within a group will either estimated to be all zero or non of them is zero. This is not reasonable especially when the variables have different levels within a group. Bondell and Reich (2009) used a weighted fusion penalty method to solve this multi-factor ANOVA problem and considered the levels of the variables within a group. The penalty term is

$$p_\lambda(\theta) = \lambda \sum_{j=1}^J \sum_{1 \leq l < k \leq q_j} w_j^{kl} |\theta_{jk} - \theta_{jl}|.$$

They try to minimize the penalized objective function

$$\|Y - \sum_{j=1}^J X_j \theta_j\|_2^2 + \lambda \sum_{j=1}^J \sum_{1 \leq l < k \leq q_j} w_j^{kl} |\theta_{jk} - \theta_{jl}|,$$

where the weight w_j^{kl} was set to $(n_k + n_l)/(p_j + 1)$. The advantage of this penalty is that it can collapse levels within a group by setting the coefficients to be equal.

Zhao et al. (2008) introduced the so called composite absolute penalty. In their method, parameters are divided into several groups $\mathcal{G}_1, \mathcal{G}_2, \dots, \mathcal{G}_K$ using some prior knowledge. For each group, they penalize the parameters within the group with a l_{γ_k} norm. For the resulting K dimensional vector, it was penalized by an overall l_{γ_0} norm with power γ_0 . Their method can be presented by the following minimization problem

$$\hat{\beta}_{cap} = \operatorname{argmin}_{\theta} L(Y, X, \theta) + p_{\lambda}(\theta),$$

where the penalty term $p_{\lambda}(\theta)$ equals

$$\sum_{k=1}^K (\|\theta_{\mathcal{G}_k}\|_{\gamma_k})^{\gamma_0}.$$

In their setting, the overall parameter γ_0 was set to 1, and the inner parameters γ_k ($k=1, 2, \dots, K$) were chosen according to the requirement of the model. The overall l_{γ_0} norm will penalize some group norms to exact 0 which performs a group selection effect and the inner l_{γ_k} norm will construct some desired structures of the parameters within the group.

Zhou et al. (2010) proposed a hierarchical penalty function using a reparameterizing technique to construct the common zeros across different groups. In their approach, the parameters θ_{kj} in group k were reparameterized by $d_k\alpha_{kj}$. That is $\theta_{kj} = d_k\alpha_{kj}$. The parameters d_k and α_{kj} were both penalized by a LASSO type penalty. The estimates were obtained by minimizing

$$\operatorname{argmin}_{d,\alpha} \left\{ \left\| Y - \sum_{k=1}^K d_k X_k \alpha_k \right\|^2 + \lambda \sum_{k=1}^K d_k + \beta \sum_{k=1}^K \sum_{j=1}^{j_k} |\alpha_{kj}| \right\}.$$

The linking parameters d_k can be shrunk to zero which makes all the coefficients $\theta_{k1}, \theta_{k2}, \dots, \theta_{kp_k}$ in the k th group equal zero all together. This will perform a group selection property. Meanwhile, even if the linking parameter d_k is not zero, the parameter α_{kj} may be shrunk to zero. This also makes $\theta_{kj} = 0$. So an unique zero was obtained in the k th group. The consistent property and also the sparsity property were given in their paper.

CHAPTER 2

Literature Review

Covariance matrix and precision matrix are very important statistical tools which have been studied extensively. Standard estimators such as the sample covariance matrix performs well when the dimension is low. However, as we have mentioned in Chapter 1, in order to fulfill some special needs, for example, the large dimensional data requirement and the sparsity requirement, alternative methods have to be investigated.

2.1 Direct Thresholding Approaches

The sample covariance matrix estimator S is asymptotically unbiased. Nevertheless, according to the research of Yin (1988) and Bai (1993), the eigenvalues of sample covariance matrix S tend to be more dispersing than the population eigenvalues. This leads to shrinkage estimation methods that shrink the eigenvalues of sample covariance matrix. Dey and Srinivasan (1985) proposed an orthogonal invariant minimax estimator under the Stein's loss function. According to their setting, the estimator was chosen as $R\phi(L)R'$, where R is a matrix constructed by the eigenvectors of the sample covariance matrices and $\phi(L)$ is a diagonal matrix. Each entry of matrix $\phi(L)$ was chosen as a function of the eigenvalues of the sample covariance matrix. The eigenvectors of this estimator are the same as the sample covariance matrix but the eigenvalues are shrunken.

Ledoit and Wolf (2003a, 2003b, 2004) have developed a series of work that focused on combining the sample covariance matrix with a well structured matrix. Let Σ denote the true covariance matrix and S is the sample covariance matrix. The idea of their approach is to find an estimator $\hat{\Sigma} = \delta F + (1 - \delta)S$ that minimizes the following risk function

$$\min_{\delta} E\|\delta F + (1 - \delta)S - \Sigma\|_F. \quad (2.1)$$

Here δ ranges from 0 to 1 and F is a matrix that has special structure. This method shrinks the sample covariance matrix S towards the structured matrix F and makes a tradeoff between estimating bias and prediction variance.

The first work has been done by Ledoit and Wolf (2003a), where they chose F as a matrix that was computed from a single index model for the stock return data. In another work of Ledoit and Wolf (2003b), F was chosen as a matrix with equal off diagonal elements.

The matrix F was chosen to be vI in Ledoit and Wolf (2004). Under this setting, the resulting estimator is named as Ledoit-Wolf estimator. Because the minimizer of (2.1) depends on the underlying true covariance matrix Σ , the authors proposed asymptotic estimators of v and δ based on the sample covariance matrix. This work is considered as a benchmark due to the simplicity and convenience of calculation.

Besides shrinking the eigenvalues, nowadays more and more researchers focused on estimating sparse covariance matrices that the parameters in the covariance matrix were shrunk. This is because sparse covariance and precision matrices provide more interpretable structures of the variables. A zero element in the covariance matrix represents that the corresponding variables are marginally independent and a zero element in the precision matrix represents the corresponding two variables

are independent conditional on all the remaining variables. Both independent relationships will simplify the whole structure of the variables. Special interest was gained by the sparse precision matrix because a sparse precision matrix is uniquely corresponding to an undirected graph of the variables if the variables have a multivariate normal distribution.

Bickel and Levina (2008b) proposed a direct hard thresholding method for estimating the covariance matrix. The estimator can be simply obtained as

$$\hat{\Sigma}_\lambda \quad (\hat{\sigma}_{kl} = s_{kl}I_{|s_{kl}|>\lambda}, k \neq l),$$

where s_{kl} denotes the kl th element of the sample covariance matrix S . This method simply shrinks the small elements in the sample covariance matrix to zero and achieves a sparse estimator of the covariance matrix. The convergence rate under operator norm was given on a large class of matrices. El Karoui (2008) independently proposed a similar direct thresholding approach and the consistent property under operator norm was also given.

This direct thresholding method was further investigated by Rothman et al. (2009). They extended the hard thresholding method to more general methods. Instead of choosing the kl th element $\hat{\sigma}_{kl}$ as $s_{kl}I_{|s_{kl}|>\lambda}$, they chose

$$\hat{\sigma}_{kl} = p_\lambda(s_{kl}) \quad (k \neq l).$$

The threshold function p_λ can be extended from hard threshold function to a

more generalized thresholding operators that satisfy several requirements. The convergence rate is also given in their paper. These direct thresholding methods are attractive since there is nearly no computation burden except the computation of the threshold parameter using cross validation.

These two thresholding methods both employ universal threshold functions and an adaptive version of the direct thresholding methods was proposed by Cai and Wu (2011). They argued that the adaptive thresholding estimator $\hat{\Sigma}$ of Σ with kl th element $\hat{\sigma}_{kl} = p_{\lambda_{kl}}(s_{kl})$ would outperform the estimator from the universal thresholding methods because the sample covariances would have a wide range of variability. Here, $p_{\lambda_{kl}}$ is a threshold function with parameter λ_{kl} which is closely related to the sample correlation coefficients. An optimal rate of $s_0(p) \log(p/n)^{(1-q)/2}$ is achieved by the adaptive estimator.

These thresholding methods have sounding convergence properties which hold when $\log(p)/n = o(1)$. Nevertheless, it has to be noted that these methods can not guarantee the positive definiteness property of the estimators which is a fundamental requirement for covariance matrices.

2.2 Penalized Approaches

Most of the shrinkage methods were based on covariance matrices. One explanation is that the sample covariance matrix is always available. Shrinking the sample covariance matrix is easy and straightforward. However, shrinking the precision matrix is not easy. First of all, the inverse of sample covariance matrix may not exist at all which will occur when $p > n$. Even if the dimension p is less than n , it was shown that the inverse of sample covariance matrix may not be a good estimator for precision matrix because the estimator is ill-conditioned which means the estimation error will significantly increase when inverting the sample covariance matrix (see Ledoit and Wolf 2004).

Although directly shrinking the precision matrix may not be a good choice, alternative methods may also achieve the shrinkage objective, for example, penalized methods. By carefully choosing the loss function and penalty function, one can also achieve sparse estimates of covariance matrix and precision matrix.

The first approach that employed the penalized method in estimating a sparse precision matrix was done by Meinshausen and Bühlmann (2006), where they regressed each variable on all the rest variables using a LASSO method. The regression coefficients can be penalized to zero by the l_1 penalty term. The ij th and

j th components of the precision matrix were estimated to be zero if the coefficient of variable i regressed on variable j or the coefficient of variable j regressed on variable i equals zero, or both of them are zero. It has to be noted that this method only focuses on finding the positions of the zero entries in the precision matrix which reveals the underlying gaussian graphical model of the variables but does not provide an estimator of the precision matrix.

Most of the penalized approaches for estimating matrices are based on the normal likelihood function. In the work of d'Aspremont et al. (2008), they suggested a penalized method that imposes a penalty function on the number of nonzero elements of the precision matrix based on the negative log normal likelihood function, which made a tradeoff between the complexity of the target matrix and the estimation bias. This method is similar to the AIC method.

Instead of penalizing the number of nonzero elements in precision matrix, Friedman et al. (2008) and Rothman et al. (2008) both proposed a penalized method that directly penalizes the off diagonal elements of the precision matrix by adding a l_1 penalty to the elements of precision matrix based on negative log normal likelihood loss function. A very fast computation algorithm called GLASSO was developed in Friedman et al. (2008) which is based on the work of Friedman et al. (2007). The convergence rate of the estimator under the Frobenius norm was firstly given in Rothman et al. (2008).

Lam and Fan (2009) extended the penalized methods by replacing the l_1 penalty with more general penalties such as SCAD. Besides the estimator of precision matrix, the penalized estimators of covariance matrix, correlation coefficients matrix were also given in that paper. Explicit convergence rate of these estimators under Frobenius norm were also investigated.

Another interesting approach was done by Cai et al. (2011). In their approach, a sparse precision matrix is obtained by minimizing the elementwise l_1 norm of the matrix Ω under the constraint

$$\|S\Omega - I\|_\infty < \lambda.$$

In their paper, the l_1 norm of matrix A (a $n \times p$ matrix) is defined as $\sum_{i=1}^n \sum_{j=1}^p |a_{ij}|$, the l_∞ norm is defined as $\max_{i,j} |a_{ij}|$. The resulting estimator $\hat{\Omega}$ has elements

$$\hat{\omega}_{ij} = \hat{\omega}_{ji} = \hat{\omega}_{ij}^1 I\{\hat{\omega}_{ij} \leq \hat{\omega}_{ji}\} + \hat{\omega}_{ji}^1 I\{\hat{\omega}_{ij} > \hat{\omega}_{ji}\}$$

where $\hat{\omega}_{ij}$ is the ij th element of the estimator from the above minimization problem.

This work is interesting since it provided a penalized method without likelihood function. The method can be implemented by linear programming which is relatively simple to compute.

2.3 Methods Based on Ordered Data

Thresholding methods and direct penalization methods mentioned above are all invariant with respect to the order of variables. Nevertheless, in some applications, prior information about the order of variables are available. This drives researchers to investigate new methods that use the prior information.

In some applications, it is reasonable to assume variables that are far away may be not correlated to each other. Thus the corresponding covariances are zero. Based on this assumption, a direct banding method was proposed by Bickel and Levina (2008a). In that paper, the kl th element in the sample covariance matrix was shrunk to zero if and only if $|k - l| > M_n$. Here, M_n is an integer that was chosen by cross validation. The convergence rate was given for a large class of covariance matrices.

In some cases, the variables have a natural order, which means one can fit them with an autoregressive model. This property reminds us that the modified Cholesky decomposition can be implemented in estimating covariance matrix. The modified Cholesky decomposition of a given matrix Σ can be written as $\Sigma = L^{-1}DL^{-1'}$ and the elements in the lower triangle matrix L can be interpreted as the regression coefficients that one variable regressed on its predecessors. Wu and

Pourahmadi (2003) proposed an estimator which was obtained by regularizing the first K sub diagonals of the autoregressive matrix L through a smoothing technique and shrinks the rest elements to zero. Here, K was chosen by an AIC method. This method provides a K banded estimate for precision matrix.

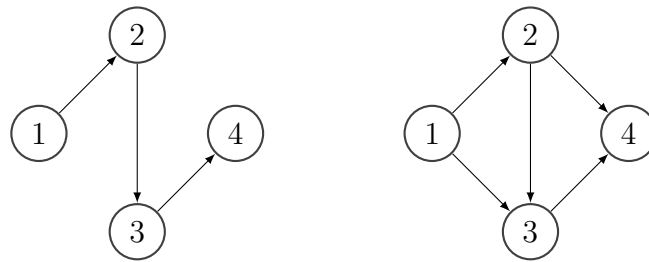
Huang et al. (2006) proposed a penalizing method that imposes a L_1 or L_2 penalty to the elements of autoregressive matrix T based on negative log normal likelihood function and achieved sparse estimate of T . Assumption of this research is more flexible than the banding approaches. However, it can not guarantee a sparse estimate of precision matrix or covariance matrix even though a sparse autoregressive matrix was obtained.

An adaptive banding approach has been proposed by Levina et al. (2006). A nested LASSO penalty was imposed to the Cholesky factors which leads to a sparse autoregressive matrix estimate that the banding length for each row of the autoregressive matrix can be different. This adaptive banding approach can efficiently introduce sparse precision matrix estimate and enjoy positive definite property by using Cholesky decomposition. In Rothman et al. (2010), a new interpretation of modified Cholesky decomposition was proposed, and an adaptive regularizing approach which is similar to Levina et al. (2006) was proposed. This method leads to banded inverse autoregressive matrix estimate and also banded covariance matrix estimate.

2.4 Motivation and Significance

As described in the review part, numerous methods for estimating covariance matrix and precision matrix have been proposed along with the development of penalizing methods and parsimony approaches. All these methods focused on estimating single covariance matrix or precision matrix. Nevertheless, In some applications, it is much more valuable to fit a collection of them if multiple groups of data are observed. This may be especially attractive when data is observed from different groups that have homologous structures. For example, we can consider the salary level of men and women in Singapore in the past ten years. The salary levels for these two groups may be different. Meanwhile, it is possible that both groups may share some common relationships. For instance, salary level for women at 2008 may not increase at all while it is possible that the salary level will slightly increased for men. But in 2000, both salaries of men and women would increase because the economic condition became better. If the covariance matrices of the salary level for men and women are estimated separately, then the information of similarity between both groups may be ignored and accuracy of the estimators is reduced.

Here is another example. The two graphs below show the relationships of four variables in two groups.



In the first group, x_1 affects x_2 , x_2 affects x_3 and x_3 affects x_4 . In the second group, the same relationships also exist. Furthermore, in the second graph, some other unique relationships exist that x_1 directly affects x_3 and x_2 directly affects x_4 . In both groups, x_1 and x_4 are independent condition on x_2 and x_3 . These independence relationships indicate that the corresponding parameters in both covariance matrices are all 0. The marginally independence relationships in both groups may be attractive in inference and model explanation. Revealing this kind of common zeros is one of the motivations of our work.

Besides the motivation of constructing common relationship between different groups, more importantly, we are also trying to improve the prediction accuracy. If prior information indicates that the data from different groups share similar structures, one can imagine that the estimation accuracy may be improved if we estimate the covariance matrix and precision matrix by borrowing information from other groups. Jointly estimating the covariance matrix and it's inverse must be a possible way to combine the information together. The joint estimation method may be valuable especially when the sample size is relatively small and the covariance

matrices (or say structures) are similar in different groups.

There are several ways to describe the common relationships of variables in two data groups. For example, variables x and y are marginally independent in both groups (this relationship can be shown by the common zeros in both covariance matrices); x and y are independent conditional on all the rest variables in both groups (it can be presented by common zeros in both precision matrices) or x and y are not correlated in a autoregressive model which can be presented by common zeros in both autoregressive matrices. In order to reveal the common relationships of variables between groups, we have to choose one of the above relations and try to jointly estimate the corresponding parameters. In this research, we investigate the third type of relationship by jointly estimating the autoregressive parameters, and try to improve the prediction accuracy.

Guo et al. (2009) considered the joint estimation technique, in which they jointly estimated the precision matrices using a reparameterize method. A hierarchical penalty function was added to the negative log likelihood function and a collection of homologous precision matrices were obtained. This approach is very efficient in discovering common zeros in precision matrices in different groups and it is also attractive in revealing the underlying graphical models of the data.

The hierarchical penalty term is equivalent to a nonconvex penalty function

which makes the computation procedure of the algorithm sensitive to the starting point, and their approach mainly focused on revealing the underlying graphical model. Alternatively, we use convex penalty terms and focus on estimating both the covariance matrix and precision matrix. The new methods were achieved by imposing penalties with group effect to the Cholesky factors of the data. The specific objectives of this research include:

1. Propose new methods which consider the similar structures of different groups.
2. Study both consistent properties of the resulting covariance matrices and precision matrices and also the sparsity property of the resulting autoregressive matrices.
3. Compare the performance of the new methods with the existing separated estimation method. Apply our new methods to real data.

This research may provide better choice for estimating the covariance matrix and precision matrix when multiple groups of data that have natural order is analyzed due to the following reasons.

1. The new methods provide more interpretable estimates of matrices for multiple groups of data. The estimates are guaranteed to be positive definite due to

the usage of Cholesky decomposition.

2. The estimation accuracy may be increased as the information in different groups were shared by our joint estimation approaches.

It has to be noted that the joint estimation methods are immature attempt in analyzing multiple groups of data. There may be some issues to be further considered. For instance, order of variables are assumed in this research and our methods work well when the structures of different groups are similar. How to find the order of variables when the natural order is not available and how to test the similarity of structures of different groups may need to be further investigated. However, they are not the central concern of this research.

Besides the work of Guo et al. (2009), recently, Lee et al. (2012), Lee and Liu (2012) also considered the joint estimation methods based on multicategory data. Nevertheless, they focused on multiple response regression problems. Another recent work of Danaher et al. (2012) investigated the joint precision matrix estimation problem using the log normal likelihood function with a sparse fusion LASSO or sparse group LASSO which is similar to our approach. However, the same as Guo et al. (2009), the main purpose of the research is to reveal the underling graphical model. The positive definite property can not be guaranteed.

The content of this thesis is organized as follows:

In Chapter 1, we have stated some necessary background knowledge related to our methods.

In Chapter 2, we have reviewed the literature of covariance matrix and precision matrix estimation.

In Chapter 3, we propose the new joint estimation methods. The computation issues as well as the theoretical properties are also given.

In Chapter 4, we compare our methods with the separated estimation method and other well known methods. A real data analysis is also conducted.

In Chapter 5, we do a summarization of our work. Potential future researches are also discussed.

Proofs of the theories are given in the Appendix part.

CHAPTER 3

Model Description

In this chapter, we assume the variables have a natural order. This assumption is reserved for the longitudinal data. On the other side, in the point of view of graphical modeling, if the relationship of the variables can be presented by a directed acyclic graph without cycles, we can always sort the variables into a specific order where the new sequence of variables satisfies that the variables afterwards only related to the previous variables. In this case, these variables satisfy the ordered assumption.

3.1 Penalized Joint Normal Likelihood Function

Assume we have J groups of observations. $Y_i^{(j)} = (y_{i1}^{(j)}, y_{i2}^{(j)}, \dots, y_{ip}^{(j)})^T$ is an observation from group j and it is a p dimensional vector. The number of observations from group j is n_j . Assume the observations in each group follow a multivariate normal distribution which is $Y_i^{(j)} \sim N(\mu_0^{(j)}, \Sigma_0^{(j)})$. For simplicity, we centralize the observations in each group, which means $\sum_{i=1}^{n_j} y_{ik}^{(j)} = 0$ for $k=1, 2, \dots, p$ and $j=1, 2, \dots, J$. The normal assumption is common, even the sample covariance matrix can be treated as the maximum likelihood estimator based on normal distribution. Assume matrix $A^{(j)} = (y_{ik}^{(j)})_{n_j \times p}$ is constructed by the observations in the j th group and $S^{(j)} = (A^{(j)})^T A^{(j)} / n_j$, then $S^{(j)}$ is the sample covariance matrix.

We write the modified Cholesky decomposition of the covariance matrix $\Sigma_0^{(j)}$ as $T_0^{(j)} \Sigma_0^{(j)} T_0^{(j)} = D_0^{(j)}$, where $T_0^{(j)}$ is the autoregressive matrix and $D_0^{(j)}$ is a diagonal matrix. The diagonal elements of $T_0^{(j)}$ are all 1 and lower triangular element $t_{0kl}^{(j)}$ is set to $-\phi_{0kl}^{(j)}$. Denote the diagonal matrix $D_0^{(j)}$ by $\text{diag}\{\sigma_{01}^{(j)}, \sigma_{02}^{(j)}, \dots, \sigma_{0p}^{(j)}\}$, under this modified decomposition, we can estimate the Cholesky factors $\phi_{0kl}^{(j)}$ and the parameters $\sigma_{0k}^{(j)}$ instead of the covariance matrix itself since once the estimators of $T_0^{(j)}$ and $D_0^{(j)}$ are available, the estimators of $\Sigma_0^{(j)}$ and $\Omega_0^{(j)}$ can be obtained.

In Chapter 1, we stated an explanation of modified Cholesky decomposition

of covariance matrix. Accordingly, we have the following explanation of modified Cholesky decomposition for sample covariance matrix. This property was generally stated in Pourahmadi (1999).

Proposition 3.1. *Assume $Y_i = (y_{i1}, y_{i2}, \dots, y_{ip})'$ ($i=1, 2, \dots, n$) are observed.*

Denote $S = \sum_{i=1}^n Y_i Y_i' / n$ and the modified Cholesky decomposition of S by $TST' = D$. Assume the least square regression equation of variable y_k regressed on variables $y_{k-1}, y_{k-2}, \dots, y_1$ is

$$\hat{y}_k = \sum_{l=1}^{k-1} \phi_{kl} y_l,$$

then the kl th element of matrix T is $-\phi_{kl}$ ($k > l$) and the kk th element of matrix D is $\sum_{i=1}^n (y_{ik} - \sum_{l=1}^{k-1} \phi_{kl} y_{il})^2 / n$.

This property can be proved as follows. Denote the regression residual for the i th observation of the k th regression by $\hat{\epsilon}_{ik}$, then one can write down the following equations

$$\begin{aligned} \hat{\epsilon}_{i1} &= y_{i1}, \\ \hat{\epsilon}_{i2} &= y_{i2} - \phi_{21} y_{i1}, \\ &\dots \\ \hat{\epsilon}_{ip} &= y_{ip} - \phi_{p(p-1)} y_{i(p-1)} - \phi_{p(p-2)} y_{i(p-2)} \dots - \phi_{p1} y_{i1}. \end{aligned} \tag{3.1}$$

Let $\hat{\epsilon}_i$ represent the vector $(\hat{\epsilon}_{i1}, \hat{\epsilon}_{i2}, \dots, \hat{\epsilon}_{ip})'$, and write T as a $p \times p$ matrix

with kl th element

$$\begin{cases} -\phi_{kl}, & k > l, \\ 1, & k = l, \\ 0, & k < l, \end{cases}$$

then the p equations in (3.1) can be written as

$$\hat{\epsilon}_i = TY_i.$$

This equation leads to

$$\hat{\epsilon}_i \hat{\epsilon}_i' = TY_i Y_i' T', \quad i = 1, 2, \dots, n. \quad (3.2)$$

Sum all these n equations up, we have

$$\sum_{i=1}^n \hat{\epsilon}_i \hat{\epsilon}_i' = T \left(\sum_{i=1}^n Y_i Y_i' \right) T'. \quad (3.3)$$

That is

$$\sum_{i=1}^n \hat{\epsilon}_i \hat{\epsilon}_i' / n = T S T'. \quad (3.4)$$

T is already a lower triangular matrix. Next we will show that $\sum_{i=1}^n \hat{\epsilon}_i \hat{\epsilon}_i' / n$ is a diagonal matrix. The kk th element of $\sum_{i=1}^n \hat{\epsilon}_i \hat{\epsilon}_i' / n$ is $\frac{\sum_{i=1}^n \hat{\epsilon}_{ik}^2}{n}$ which equals

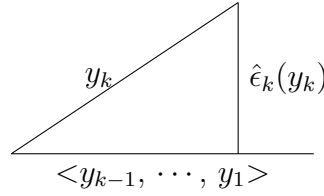
$$\sum_{i=1}^n (y_{ik} - \sum_{l=1}^{k-1} \phi_{kl} y_{il})^2 / n.$$

Consider the kl th element of $\sum_{i=1}^n \hat{\epsilon}_i \hat{\epsilon}_i' / n$ ($k > l$), which is $\sum_{i=1}^n \hat{\epsilon}_{ik} \hat{\epsilon}_{il} / n$. Denote $\hat{\epsilon}_k(y_k)$ as $(\hat{\epsilon}_{1k}, \hat{\epsilon}_{2k}, \dots, \hat{\epsilon}_{nk})'$ which is the residual vector of variable y_k regressed

on variables y_1, \dots, y_{k-1} , then $\hat{\epsilon}_k(y_k)$ is orthogonal to the space spanned by observations of variables y_1, \dots, y_{k-1} . Obviously, $\hat{\epsilon}_l(y_l)$ ($l < k$) is in this space, which implies

$$\langle \hat{\epsilon}_k(y_k), \hat{\epsilon}_l(y_l) \rangle = 0.$$

Thus $\sum_{i=1}^n \hat{\epsilon}_{il} \hat{\epsilon}_{ik} / n$ equals zero and the matrix $\sum_{i=1}^n \hat{\epsilon}_i \hat{\epsilon}_i' / n$ is a diagonal matrix.



Assume D is a diagonal matrix and the i th element is $\hat{\epsilon}_k(y_k)' \hat{\epsilon}_k(y_k) / n$ which equals $\frac{\sum_{i=1}^n (y_{ik} - \sum_{l=1}^{k-1} \phi_{kl} y_{il})^2}{n}$, then we have

$$D = TST'.$$

By the uniqueness of the modified Cholesky decomposition, we proved this property.

Go back to our problem. Consider the regression of $y_k^{(j)}$ on variables $y_{k-1}^{(j)}, y_{k-2}^{(j)}, \dots, y_1^{(j)}$, we assume the coefficients are $\phi_{k(k-1)}^{(j)}, \phi_{k(k-2)}^{(j)}, \dots, \phi_{k1}^{(j)}$. The regression residual term $\hat{\epsilon}_{ik}^{(j)}$ equals $y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl} y_{il}^{(j)}$. Following the above notations, we have

$$\hat{\epsilon}_i^{(j)} = T^{(j)} Y_i^{(j)}. \quad (3.5)$$

$\hat{\epsilon}_i^{(j)}$ can be treated as an observation of $\epsilon^{(j)}$ where $\epsilon^{(j)}$ follows a normal distribution of $N(0, D^{(j)})$. The likelihood of $\epsilon_i^{(j)}$ which is written as a function of the new parameters $T^{(j)}$ and $D^{(j)}$ is

$$\frac{1}{\sqrt{2\pi|D^{(j)}|}} \exp\left(-\frac{Y_i^{(j)'} T^{(j)'} D^{(j)-1} T^{(j)} Y_i^{(j)}}{2}\right). \quad (3.6)$$

If we take log of (3.5), multiply by -2 and omit the constant term, the likelihood function for the j th group is

$$l_j(T^{(j)}, D^{(j)}) = \sum_{i=1}^{n_j} (\log |D^{(j)}| + Y_i^{(j)'} T^{(j)'} D^{(j)-1} T^{(j)} Y_i^{(j)}). \quad (3.7)$$

This loss function (3.7) is based on the unconstrained parameters $\phi_{kl}^{(j)}$ and $\sigma_k^{(j)}$. Following (3.5), we can write $Y_i^{(j)'} T^{(j)'} D^{(j)-1} T^{(j)} Y_i^{(j)}$ as $\hat{\epsilon}_i^{(j)'} D^{(j)-1} \hat{\epsilon}_i^{(j)}$. So the term $Y_i^{(j)'} T^{(j)'} D^{(j)-1} T^{(j)} Y_i^{(j)}$ in (3.7) is

$$\sum_{k=1}^p \frac{(y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)} y_{il}^{(j)})^2}{\sigma_k^{(j)}}. \quad (3.8)$$

Recall that matrix $D^{(j)} = \text{diag}\{\sigma_1^{(j)}, \sigma_2^{(j)}, \dots, \sigma_p^{(j)}\}$. So the transformed log likelihood function for the j th group is

$$l_j(T^{(j)}, D^{(j)}) = n_j \sum_{k=1}^p \log(\sigma_k^{(j)}) + \sum_{i=1}^{n_j} \sum_{k=1}^p \frac{(y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)} y_{il}^{(j)})^2}{\sigma_k^{(j)}}. \quad (3.9)$$

In this study, for simplicity, we assume data from different groups are balanced which means that numbers of observations from different groups are all the same and the unbalanced case is almost the same as the balanced case. Assume the

number of observations in each group is n , then the transformed joint likelihood function is

$$l(T, D) = \sum_{j=1}^J \sum_{k=1}^p n \log(\sigma_k^{(j)}) + \sum_{j=1}^J \sum_{i=1}^n \sum_{k=1}^p \frac{(y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)} y_{il}^{(j)})^2}{\sigma_k^{(j)}}. \quad (3.10)$$

Without any extra information, the empirical estimator for the Cholesky factors and corresponding variances can be obtained by minimizing (3.10). The resulting covariance matrix estimators coincide with the standard MLE.

In this study, in order to introduce sparsity to the elements in the autoregressive matrix $T^{(j)}$ and impose group effect to the Cholesky factors in different groups, we add a penalty term to the above likelihood function which treats the elements in the same position as a group and penalize them together. The penalty term can be chosen as

$$p_\lambda(\phi) = \sum_{k=2}^p \sum_{l=1}^k g_\lambda(\phi_{kl}^{(1)}, \phi_{kl}^{(2)}, \dots, \phi_{kl}^{(J)}). \quad (3.11)$$

Consequently, the penalized method of jointly estimating the Cholesky factors and the variance matrices is to minimize the following function

$$\begin{aligned} & \sum_{j=1}^J \sum_{k=1}^p n \log(\sigma_k^{(j)}) + \sum_{j=1}^J \sum_{i=1}^n \sum_{k=1}^p \frac{(y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)} y_{il}^{(j)})^2}{\sigma_k^{(j)}} \\ & + \sum_{k=2}^p \sum_{l=1}^k g_\lambda(\phi_{kl}^{(1)}, \phi_{kl}^{(2)}, \dots, \phi_{kl}^{(J)}). \end{aligned} \quad (3.12)$$

An ideal choice of penalty function $g_\lambda(\phi)$ can provide sparse estimates of the

coefficients $\phi_{kl}^{(j)}$. More importantly, we would like to impose some particular structures to the Cholesky factors based on our prior information about the data. Specifically, if variables $y_k^{(j)}$ and $y_l^{(j)}$ are not correlated in any groups, the penalty term would shrink $\phi_{kl}^{(1)}, \phi_{kl}^{(2)}, \dots, \phi_{kl}^{(J)}$ all to zero. If the correlation of $y_k^{(j)}$ and $y_l^{(j)}$ are close in all groups, the penalty term may make $\phi_{kl}^{(1)}, \phi_{kl}^{(2)}, \dots, \phi_{kl}^{(J)}$ close.

3.2 IL-JMEC Method

The first choice of $g_\lambda(\phi)$ is a combination of l_1 penalty function and l_∞ penalty function. It can be roughly presented by $\lambda|\phi|_1 + \beta|\phi|_\infty$. This penalty has a nice property that it shrinks the J dimensional coefficient vector ϕ to some small subspace of R^J . In these subspaces, the coefficient vector ϕ will be encountered either of following:

1. Some coefficients will be exactly zero.
2. Some coefficients will be identical.

This shrinkage procedure is presented in figure (3.1). This figure shows how the minimizer of $\phi_1^2 - (a+b)\phi_1 + \phi_2^2 - a\phi_2 + \lambda|\phi|_1 + \beta||\phi||_\infty$ changes when different λ and β are employed where we assume $a > 0$ and $b > 0$.

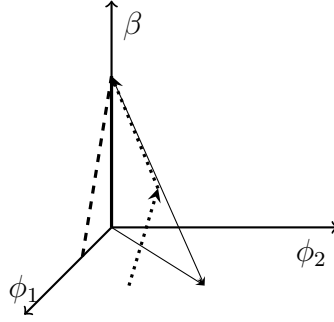


Figure 3.1 Minimizer of $\phi_1^2 - (a+b)\phi_1 + \phi_2^2 - a\phi_2 + \lambda|\phi|_1 + \beta\|\phi\|_\infty$.

1. The thick line in the figure shows that when the threshold parameter λ for the LASSO part is big enough ($\lambda > \phi_1 + \phi_2$), the coefficients are all estimated to be zero no matter what the threshold parameter β is.

2. The dashed line shows that when the threshold parameter λ is moderately large, the small coefficient ϕ_2 is always estimated to be zero while the larger coefficient ϕ_1 is shrunk by both l_1 and l_∞ term and finally goes to zero along with the increasing of threshold parameter β .

3. The dotted line is the solution of (ϕ_1, ϕ_2) when λ is smaller than a . When β increases, (ϕ_1, ϕ_2) will firstly hit the surface $\phi_1 = \phi_2$, and then is shrunk to zero.

In view of the effect of this combined penalty term, we can choose the following penalty function to impose group effect to parameters $\phi_{kl}^{(1)}, \dots, \phi_{kl}^{(J)}$,

$$g_{\lambda, \beta}(\phi) = \lambda \sum_{j=1}^J |\phi_{kl}^{(j)}| + \beta \max\{|\phi_{kl}^{(1)}|, \dots, |\phi_{kl}^{(J)}|\}. \quad (3.13)$$

Thus the resulting penalized likelihood function $l(T, D)$ for the reparameterized coefficients $\phi_{kl}^{(j)}$ and $\sigma_k^{(j)}$ is

$$\begin{aligned}
 l_{\lambda, \beta}(T, D) = & \sum_{j=1}^J \sum_{k=1}^p n \log(\sigma_k^{(j)}) + \sum_{j=1}^J \sum_{i=1}^n \sum_{k=1}^p \frac{(y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)} y_{il}^{(j)})^2}{\sigma_k^{(j)}} \\
 & + \lambda \sum_{k>l} \sum_{j=1}^J |\phi_{kl}^{(j)}| + \beta \sum_{k>l} \max_{j=1}^J \{|\phi_{kl}^{(1)}|, |\phi_{kl}^{(2)}|, \dots, |\phi_{kl}^{(J)}|\}.
 \end{aligned} \tag{3.14}$$

We call this method Infinity LASSO-Joint Matrix Estimation Approach via Cholesky Decomposition (IL-JMEC).

The objective function (3.14) is complicated, but a good news is that the minimization problem can be divided into p sub minimization problems and each of them can be solved iteratively.

3.3 GL-JMEC Method

Another choice of the penalty function is a combination of LASSO and group LASSO. As described in Yuan and Lin (2004), the group LASSO penalty can efficiently shrink a group of variables to zero. Nevertheless, it has to be noted that the group LASSO always performs an all in and all out strategy. Directly applying group LASSO in our joint estimation method can not reveal the unique relationship in some specific groups. Alternatively, we can combine the group LASSO penalty with the LASSO penalty together, and impose this joint penalty

to the autoregressive coefficients. The goal that reveals the common zeros and unique zeros may be achieved by the following joint penalty term

$$p_{\lambda,\beta}(\phi) = \lambda \sum_{j=1}^J \sum_{l < k} |\phi_{kl}^{(j)}| + \beta \sum_{l < k} \sqrt{\sum_{j=1}^J \phi_{kl}^{(j)2}}. \quad (3.15)$$

This joint penalty function was also investigated in Friedman et al. (2010).

Applying this penalty to our study, the penalized likelihood function becomes

$$\begin{aligned} l_{\lambda,\beta}(T, D) = & \sum_{j=1}^J \sum_{k=1}^p n \log(\sigma_k^{(j)}) + \sum_{j=1}^J \sum_{i=1}^n \sum_{k=1}^p \frac{(y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)} y_{il}^{(j)})^2}{\sigma_k^{(j)}} \\ & + \lambda \sum_{j=1}^J \sum_{k > l} |\phi_{kl}^{(j)}| + \beta \sum_{k > l} \sqrt{\sum_{j=1}^J \phi_{kl}^{(j)2}}. \end{aligned} \quad (3.16)$$

We call this sparse Group LASSO-Joint Matrix Estimation Approach Via Cholesky Decomposition(GL-JMEC).

The insight of sparse property of the sparse group LASSO can be examined by the following simple example. A different procedure which leads to the same result was investigated in Friedman et al. (2010). We consider the following penalized regression problem which minimizes the objective function

$$\frac{1}{2} \|Y - X\phi\|_2^2 + \lambda \|\phi\| + \beta \|\phi\|_2. \quad (3.17)$$

Let $C = X^T Y$ represents the correlation coefficients vector of variables x_1, x_2, \dots, x_p and y . The i th element of C is c_i . We write vector ϕ as $r\mu$ where $r = \|\phi\|_2$ and $\mu = \phi/r$. Let S be a p dimensional vector. The i th element s_i equals the sign of ϕ ,

then we have $|\phi| = \phi^T S = r\mu^T S$. The objective function (3.17) can be written as

$$\begin{aligned} & \frac{1}{2}\phi^T X^T X\phi - \phi^T X^T Y + \lambda\phi^T S + \beta r \\ &= \frac{1}{2}r^2\mu^T X^T X\mu - r\mu^T C + \lambda r\mu^T S + \beta r \\ &= r\left[\frac{1}{2}r\mu^T X^T X\mu - \mu^T(C - \lambda S) + \beta\right]. \end{aligned} \quad (3.18)$$

Here we omitted the constant term $\frac{1}{2}Y^T Y$. Denote

$$w_i = \begin{cases} c_i - \text{sign}(c_i)\lambda, & |c_i| > \lambda, \\ 0, & |c_i| \leq \lambda. \end{cases}$$

Let $W = (w_1, w_2, \dots, w_p)'$. It is easy to prove that

$$\mu^T(C - \lambda S) \leq \mu^T W, \quad (3.19)$$

for arbitrary vector μ where S is a function of μ . Consequently, we have the following property of the sparse group LASSO:

Proposition 3.2. $\phi=0$ is the minimizer of (3.17) if and only if $\beta \geq \|W\|_2$.

This property can be proved as follows. When $\beta \geq \|W\|_2$, by (3.19), we know

$$\beta - \mu^T(C - \lambda S) \geq \beta - \mu^T W.$$

Recall that $\|\mu\|_2 = 1$, we have $\mu^T W \leq \|W\|_2 \leq \beta$. Thus

$$r\left[\frac{1}{2}r\mu^T X^T X\mu - \mu^T(C - \lambda S) + \beta\right] \geq r\left[\frac{1}{2}r\mu^T X^T X\mu + \beta - \mu^T W\right] \geq 0.$$

This shows that the minimal value of (3.18) is greater or equal to 0. However, 0 can be obtained when $r = 0$ which is equivalent to $\phi = 0$. Thus $\phi = 0$ is the minimizer of (3.18). This gives us the sufficient condition.

In order to prove the necessary part, it suffices to show that when $\|W\|_2 > \beta$, the minimal value is not achieved when $\phi = 0$, which is equivalent to show that the minimal value of (3.18) is less than 0. Instead of directly finding the minimal value of (3.18), we will find a particular value of ϕ such that (3.18) is less than 0.

We choose $\hat{\mu} = W/\|W\|_2$. When $|c_i| > \lambda$, $s_i = \text{sign}(\hat{\mu}_i) = \text{sign}(c_i - \text{sign}(c_i)\lambda) = \text{sign}(c_i)$ which leads to $w_i(c_i - s_i\lambda) = w_i(c_i - \text{sign}(c_i)\lambda) = w_i^2$. When $|c_i| \leq \lambda$, we have $w_i = 0$ which also leads to $w_i(c_i - \text{sign}(c_i)\lambda) = 0 = w_i^2$. Thus $W^T(C - \lambda S) = W^T W$, which implies

$$\begin{aligned} & r\left[\frac{1}{2}r\hat{\mu}^T X^T X \hat{\mu} - \hat{\mu}^T(C - \lambda S) + \beta\right] \\ &= r\left[\frac{1}{2}r\hat{\mu}^T X^T X \hat{\mu} - W^T(C - \lambda S)/\|W\|_2 + \beta\right] \\ &= r\left[\frac{1}{2}r\hat{\mu}^T X^T X \hat{\mu} - W^T W/\|W\|_2 + \beta\right] \\ &= r\left[\frac{1}{2}r\hat{\mu}^T X^T X \hat{\mu} - \|W\|_2 + \beta\right]. \end{aligned}$$

Recall that $\|W\|_2 > \beta$, we can choose r small enough such that $\frac{1}{2}r\hat{\mu}^T X^T X \hat{\mu}$ is dominated by $\|W\|_2 - \beta$, then we have found a nonzero $\phi = r\hat{\mu}$ such that $\frac{1}{2}\phi^T X^T X \phi - \phi^T X^T Y + \lambda\phi^T S + \beta r < 0$. This proved the necessary part.

This rule will be used in the computation part. By using this property, computation procedure of GL-JMEC can be simplified.

We present the contour graphs of the sparse infinity LASSO and sparse group LASSO as follows.

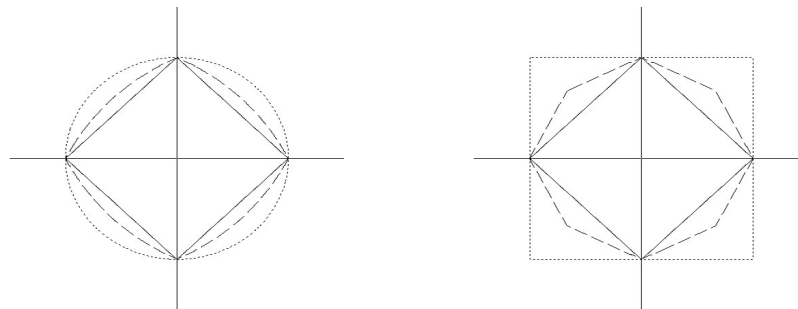


Figure 3.2 Contour graph for sparse group LASSO (left) and sparse l_∞ LASSO (right).

In this figure, solid line represents the contour curve of Lasso penalty, dashed line represents the contour curve of sparse group penalty and dotted line represents the contour curve of group LASSO penalty. In the right graph, solid line also represents the contour curve of LASSO penalty, dashed line represents the contour of sparse l_∞ penalty and dotted line represents the contour curve of l_∞ penalty.

3.4 Computation Issue

The likelihood function (3.14) can be split into J disjoint functions. Denote $T_{(k)}$ as the collection of coefficients in the k th row of $T^{(1)}, \dots, T^{(J)}$ and $\sigma_{(k)}$ as the collection of $\sigma_k^{(1)}, \dots, \sigma_k^{(J)}$. Denote

$$f_k(T_{(k)}, \sigma_{(k)}) = \sum_{j=1}^J n \log(\sigma_k^{(j)}) + \sum_{j=1}^J \sum_{i=1}^n \frac{(y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)} y_{il}^{(j)})^2}{\sigma_k^{(j)}} + \sum_{l=1}^{k-1} g_\lambda(\phi_{k1}^{(1)}, \dots, \phi_{kl}^{(J)}), \quad (3.20)$$

then we can write $l_{\lambda, \gamma}(T, D)$ as $\sum_{k=1}^p f_k(T_{(k)}, \sigma_{(k)})$. Minimizing $l_{\lambda, \gamma}(T, D)$ is equivalent to minimizing $f_k(T_{(k)}, \sigma_{(k)})$ ($k = 1, 2, \dots, p$) separately.

When $k = 1$, the function $f_k(T_{(k)}, \sigma_{(k)})$ degenerates to $\sum_{j=1}^J n \log(\sigma_1^{(j)}) + \sum_{j=1}^J \sum_{i=1}^n \frac{(y_{i1}^{(j)})^2}{\sigma_1^{(j)}}$. the minimal value is achieved when $\sigma_1^{(j)} = 1/n \sum_{i=1}^n (y_{i1}^{(j)})^2$.

When $k \geq 2$, minimizing (3.20) can be done through an iterative procedure. Assume the resulting estimates of the coefficients from the t th iteration are $\phi_{kl}^{(j)}(t)$ and $\sigma_k^{(j)}(t)$. In the $(t+1)$ th iteration, ϕ and σ can be updated through a two step procedure as follows:

Step 1. For fixed $\phi_{kl}^{(j)}$ which equals $\phi_{kl}^{(j)}(t)$, minimizing $\sum_{j=1}^J n \log(\sigma_k^{(j)}) + \sum_{j=1}^J \sum_{i=1}^n \frac{(y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)}(t) y_{il}^{(j)})^2}{\sigma_k^{(j)}}$ with respect to $\sigma_k^{(j)}$. The variance term $\sigma_k^{(j)}$ can be

directly updated by

$$\sigma_k^{(j)}(t+1) = \frac{\sum_{i=1}^n (y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)}(t) y_{il}^{(j)})^2}{n}.$$

Step 2. Fix $\sigma_k^{(j)} = \sigma_k^{(j)}(t+1)$, $\phi_{kl}^{(j)}$ can be updated by minimizing the objective function

$$\begin{aligned} & \sum_{j=1}^J 1/\sigma_k^{(j)}(t+1) \sum_{i=1}^n (y_{ik}^{(j)} - \sum_{l=1}^{k-1} \phi_{kl}^{(j)} y_{il}^{(j)})^2 \\ & + \sum_{l=1}^{k-1} g_\lambda(\phi_{kl}^{(1)}, \dots, \phi_{kl}^{(J)}). \end{aligned} \quad (3.21)$$

Minimizing (3.21) is our main concern. A group coordinate optimization procedure which was mentioned in Friedman et al. (2007) is used in our work. The idea of this method is to update a single or a group of variables while all the others are fixed. The initial value $\hat{\phi}_{kl}^{(j)}$ is set to the t th update of $\phi_{kl}^{(j)}$ in the minimization procedure of (3.20). That is $\hat{\phi}_{kl}^{(j)} = \phi_{kl}^{(j)}(t)$ ($l = 1, 2, \dots, k-1, j = 1, 2, \dots, J$). For fixed $1 \leq l_0 \leq k-1$, we would like to update a group of parameters $\phi_{kl_0}^{(j)}$ ($j = 1, 2, \dots, J$) while all other parameters are fixed by minimizing the following function:

$$\sum_{j=1}^J c_j \|R_{l_0}^{(j)}(r) - \phi_{kl_0}^{(j)} Y_{*l_0}^{(j)}\|_2^2 + g_\lambda(\phi_{kl_0}^{(1)}, \dots, \phi_{kl_0}^{(J)}). \quad (3.22)$$

Here $c_j = 1/\sigma_k^{(j)}(t+1)$ is always fixed in this sub iteration procedure. $Y_{*l_0}^{(j)}$ is the vector of observations for the l_0 th variable in j th group. That is $Y_{*l_0}^{(j)} = (y_{1l_0}^{(j)}, \dots, y_{nl_0}^{(j)})^T$. The vector $R_{l_0}^{(j)}(r)$ is an n dimensional vector in which the i th element equals $y_{ik}^{(j)} - \sum_{l \neq l_0} \hat{\phi}_{kl}^{(j)} y_{il}^{(j)}$.

So far, the minimization problem can be simplified to minimize (3.22). Optimization procedures for IL-JMEC and GL-JMEC are slightly different. So we will state these two procedures separately.

IL-JMEC : In this method, the penalty function $g_\lambda(\phi_{kl_0}^{(1)}, \dots, \phi_{kl_0}^{(J)}) = \lambda \sum_{j=1}^J |\phi_{kl_0}^{(j)}| + \beta \max\{\phi_{kl_0}^{(1)}, \dots, \phi_{kl_0}^{(J)}\}$. The explicit minimizer of (3.22) can be found (see Wu and Lange 2008). If we simplify the function (3.22), the minimization problem is equivalent to solve

$$\operatorname{argmin}_\theta \sum_{j=1}^J (a_j \theta_j^2 - b_j \theta_j) + \lambda \sum_{j=1}^J |\theta_j| + \beta |\theta|_\infty. \quad (3.23)$$

Here $a_j = c_j \|Y_{*l_0}^{(j)}\|_2^2$, $b_j = 2c_j \langle Y_{*l_0}^{(j)}, R_{l_0}^{(j)}(r) \rangle$ and $\theta_j = \phi_{kl_0}^{(j)}$. Since a_j is positive, the minimizer θ of (3.23) must satisfy that θ_j has the same sign as b_j . Thus (3.23) can be written as

$$\operatorname{argmin}_\theta \sum_{j=1}^J [a_j \theta_j^2 - (|b_j| - \lambda) |\theta_j|] + \beta |\theta|_\infty, \quad (\operatorname{sign}(\theta_j) = \operatorname{sign}(b_j)), \quad (3.24)$$

which is the same as minimizing $\sum_{j=1}^J [a_j \theta_j^2 - (|b_j| - \lambda)_+ |\theta_j|] + \beta |\theta|_\infty$. Assume $(|b_j| - \lambda)_+ / a_j$ are sorted in descending order and integer u ($1 \leq u \leq J$) is the first integer that satisfies $\frac{\sum_{j=1}^u (|b_j| - \lambda)_+ - \beta}{\sum_{j=1}^u a_j} > \frac{(|b_{u+1}| - \lambda)_+}{a_{u+1}}$ (denote $\frac{(|b_{J+1}| - \lambda)_+}{a_{J+1}} = 0$), then the minimizer of (3.23) is

$$\hat{\theta}_j = \begin{cases} \operatorname{sign}(b_j) \frac{\sum_{j=1}^u (|b_j| - \lambda)_+ - \beta}{2 \sum_{j=1}^u a_j}, & 1 \leq j \leq u, \\ \operatorname{sign}(b_j) \frac{(|b_j| - \lambda)_+}{2a_j}, & u < j \leq J, \end{cases}$$

when this u exists. Otherwise, the minimizer $\hat{\theta}$ is 0.

The estimates of $\hat{\phi}_{kl_0}^{(j)}$ ($j = 1, 2, \dots, J$) are updated by $\hat{\theta}$. Repeat this procedure for $l_0 = 1, 2, \dots, k-1, 1, \dots$ until convergence. Iteration will stop when the l_1 norm of the difference of two successive updates is less than 10^{-4} (We call two updates are successive if one is obtained by coordinately updating every element in the other one). It has to be noted that the function (3.21) is convex. So function (3.21) is always decreasing with each update. By iteratively updating T_k and σ_k , we can achieve the minimizer of (3.20) which is also the minimizer of (3.14).

GL-JEMC: In this case, the penalty function $g_\lambda(\phi_{kl_0}^{(1)}, \dots, \phi_{kl_0}^{(J)}) = \lambda \sum_{j=1}^J |\phi_{kl_0}^{(j)}| + \beta \sqrt{\sum_{j=1}^J \phi_{kl_0}^{(j)}}$. The same as the previous situation, function (3.22) can be also simplified as

$$\operatorname{argmin}_\theta \sum_{j=1}^J (a_j \theta_j^2 - b_j \theta_j) + \lambda \sum_{j=1}^J |\theta_j| + \beta \|\theta\|_2, \quad (3.25)$$

where $a_j = c_j \|Y_{*l_0}^{(j)}\|_2^2$, $b_j = 2c_j \langle Y_{*l_0}^{(j)}, R_{l_0}^{(j)}(r) \rangle$ and $\theta_j = \phi_{kl_0}^{(j)}$. Unlike the IL-GEMC method, (3.25) has no explicit solution. However, the minimizer $\hat{\theta}$ of (3.25) can be quickly found by the following iteration method:

Assume $w_j = b_j - \operatorname{sign}(b_j)\lambda$ if $\lambda < |b_j|$ and $w_j = 0$ if $\lambda \geq |b_j|$ for $j = 1, \dots, J$.

We can check if $\sqrt{\sum w_j^2} > \beta$.

If $\sqrt{\sum w_j^2} > \beta$ is true, $\hat{\theta}_j$ are all zero. This is based on proposition 3.2.

If $\sqrt{\sum w_j^2} > \beta$ fails, an iterative approach is employed to minimize (3.25). The initial value $\hat{\theta}_j(0)$ was set to $\phi_{kl_0}^{(j)}(t)$. In the $(r + 1)$ th iteration, according to Zou and Li (2008), the term $\sqrt{\sum \theta_j^2}$ can be approximated by local linear approximation

$$\frac{\sum \theta_j^2}{2\sqrt{\sum \hat{\theta}_j^2(r)}} + \frac{1}{2}\sqrt{\sum \hat{\theta}_j^2(r)}. \quad (3.26)$$

Substitute the term $\sqrt{\sum \theta_j^2}$ in (3.25) with (3.26), we can update $\hat{\theta}_j(r + 1)$ by

$$\frac{w_j}{2a_j + \beta/\sqrt{\sum \hat{\theta}_j(r)^2}}.$$

Repeat this procedure until convergence. It has to be noted that this iteration procedure is very fast. Assume the minimizer of (3.25) is $\hat{\theta}$, then $\phi_{kl_0}^{(j)}$ ($j = 1, \dots, J$) are updated by $\hat{\theta}$. Repeat this for $l_0 = 1, 2, \dots, k - 1, 1, 2, \dots$ until convergence. Similar to IL-JEMC, iteration will stop when the l_1 norm difference of two successive updates is smaller than 10^{-4} .

So far we have updated $\phi_{kl}^{(j)}$ while $\sigma_k^{(j)}$ are fixed. The objective function $f_k(T_{(k)}, \sigma_{(k)})$ is always decreasing when we update $\phi_{kl}^{(j)}$ and $\sigma_k^{(j)}$ iteratively following Step 1 and Step 2. The minimizing procedure will be terminated within finite steps.

The threshold parameters were obtained by cross validation method. In this study, we use the 5-fold cross validation approach. The observations are divided into 5 subsets. For a given combination of λ and β , let $T_{-u}^{(j)}(\lambda, \beta)$ and $D_{-u}^{(j)}(\lambda, \beta)$

denote the estimates from our joint estimation methods based on all the observations besides the u th subset. We apply these estimates on the u th subset and evaluate the performance based on the following loss function

$$l(\lambda, \beta) = \sum_{j=1}^J \sum_u n \log |D_{-u}^{(j)}(\lambda, \beta)| + \text{tr}(T_{-u}^{(j)'}(\lambda, \beta) D_{-u}^{(j)-1}(\lambda, \beta) T_{-u}^{(j)}(\lambda, \beta) S_u^{(j)}).$$

The threshold parameters λ and β are chosen so as to minimize $l(\lambda, \beta)$.

3.5 Main Results

We state some notations first. Denote the Frobenius norm of matrix A by $\|A\|_F = \sqrt{\text{tr}AA'}$. The singular values of matrix A are denoted as the square root of the eigenvalues of matrix AA' . Assume matrix AA' is a $p \times p$ matrix, we denote the square root of eigenvalues of matrix AA' by $s_p(A), s_{p-1}(A), \dots, s_1(A)$ and assume they have an increasing order. Then the operator norm of matrix A is defined as $s_1(A)$ which is denoted by $\|A\|$. The true covariance matrices and precision matrices are denoted as $\Sigma_0^{(j)}$ and $\Omega_0^{(j)}$ while the estimates are denoted as $\Sigma^{(j)}$ and $\Omega^{(j)}$. In this part, we will assume the matrices are growing along with the sample size n and also the dimension p .

Let $Z_j = \{(k, l) : k > l, \phi_{0kl}^{(j)} \neq 0\}$ which is the collection of nonzero points in the lower triangular part of matrix $L_0^{(j)}$. We write $Z = \cup_{j=1}^J Z_j$ and let the

cardinality of Z_j equals s_j and the cardinality of $\cup_{j=1}^J Z_j$ equals s .

In order to achieve the consistent property of minimizer of (3.14) and (3.16), we have to make a basic assumption that there exists a constant $c > 1$ such that the singular values of the covariance matrices are bounded. That is

$$1/c < s_p(\Sigma_0^{(j)}) \leq s_1(\Sigma_0^{(j)}) < c \quad (j = 1, 2, \dots, J). \quad (3.27)$$

This assumption is common. In Rothman (2008), Lam and Fan (2009), Guo et al. (2011). They all made the same assumption. This assumption makes inverting the covariance matrices meaningful when the dimension and sample size are growing. The following theorem gives the convergence rate of our estimates based on the Frobenius norm.

Theorem 3.1. *Assume the threshold parameters λ and β satisfy $\lambda + \beta = O(\log(p)/n)$. $T^{(j)}$ and $D^{(j)}$ ($j = 1, \dots, J$) are minimizers of the penalized likelihood function $L_{\lambda, \beta}(T, D)$ in (3.16) and (3.14). Assume s and p satisfy $(s + p) \log(p)/n = o(1)$, then we have following consistent properties*

$$\sum_{j=1}^J \|T_0^{(j)} - T^{(j)}\|_F = O_p(\sqrt{s \log(p)/n}), \quad (3.28)$$

$$\sum_{j=1}^J \|D_0^{(j)} - D^{(j)}\|_F = O_p(\sqrt{p \log(p)/n}). \quad (3.29)$$

This theorem shows the consistent properties of the estimates of the autoregressive matrices and also the corresponding variance matrices. It is close to the

Theorem 9 in Lam and Fan (2009). Based on the consistent properties of the Cholesky factors and variance matrices, the consistent properties of the resulting covariance matrices and precision matrices can be obtained by the following theorem.

Theorem 3.2. *Assume all the conditions in Theorem 3.1 hold. $T^{(j)}$ and $D^{(j)}$ ($j = 1, 2, \dots, J$) are the minimizer of (3.14) or (3.16). Let $\Omega^{(j)} = T^{(j)'} D^{(j)-1} T^{(j)}$ and $\Sigma^{(j)} = T^{(j)-1} D^{(j)} T^{(j)-1'}$, then we have the following properties*

$$\sum_{j=1}^J \|\Omega^{(j)} - \Omega_0^{(j)}\|_F = O_p(\sqrt{(s+p) \log(p)/n}), \quad (3.30)$$

$$\sum_{j=1}^J \|\Sigma^{(j)} - \Sigma_0^{(j)}\|_F = O_p(\sqrt{(s+p) \log(p)/n}). \quad (3.31)$$

The convergence rates for covariance matrices and precision matrices are the same. It has to be noted that, the direct thresholding or shrinking methods usually provide either consistent properties of the covariances matrices or precision matrices. However, they hardly provide both of them because one can not even guarantee the other exists. In this research, with the advantage of the Cholesky decomposition, both consistent rates of the resulting covariance matrices and precision matrices are given.

The last theorem illustrates sparsity of the autoregressive matrices. Due to the singular property of LASSO type penalties, we can prove the following result.

Theorem 3.3. *Suppose all conditions in Theorem 3.1 hold. Moreover, we assume $\sum_{j=1}^J \|T^{(j)} - T_0^{(j)}\| = O_p(\zeta_n)$ and $\sum_{j=1}^J \|D^{(j)} - D_0^{(j)}\| = O_p(\eta_n)$ which satisfy $\sqrt{\log p/n} + \sqrt{\zeta_n} + \sqrt{\eta_n} = O_p(\lambda)$, then with probability tending to 1, the minimizers of (3.14) or (3.16) satisfy $\phi_{kl}^{(j)} = 0 \quad ((k, l) \in Z_j^c, 1 \leq j \leq J)$.*

From the above theorem, we know that the consistent properties of the estimates from both joint estimation methods are identical. According to the proof, this attributes to the convex property of the penalty functions. It has to be noted that the consistent properties of this penalized approach are related to the rate of the sum of the two threshold parameters $\lambda + \beta$. On the other hand, only the threshold parameter λ controls the sparsity of the resulting estimates.

CHAPTER 4

Simulation Results

4.1 Simulation Settings

In this chapter, we compare our approaches with the method mentioned in Huang et al. (2006). In that paper, they proposed a penalized likelihood method based on the Cholesky decomposition. A LASSO penalty term was imposed to the Cholesky factors in their research. Cholesky decomposition is employed in our methods as well as the method in Huang et al. (2006), which makes these approaches comparable. Meanwhile, the estimates proposed by Ledoit and Wolf (2004) are also simulated as a benchmark work. Besides these approaches, the

results of the empirical estimators are also presented.

Since the Cholesky factors and corresponding variance parameters are obtained in this research, both the covariance matrices and precision matrices can be obtained directly. By Theorem 3.2, we proved that the precision matrices and covariance matrices resulted from our methods have the same convergence rates under Frobenius norm. So we only focus on the prediction accuracy of precision matrices. Another reason is that we mainly compare our joint estimation approaches with the method mentioned in Huang et al. (2006). In both approaches, It is more straightforward to obtain the precision matrix estimates.

Two loss functions are used to measure the prediction accuracy of the estimators. The first one is Frobenius loss which is defined as

$$FL = \sum_{j=1}^J \|\Omega^{(j)} - \Omega_0^{(j)}\|_F^2 / p.$$

This loss will provide us with the information of the elementwise closeness of estimators and the true matrices.

The other loss is called Operator loss which can be calculated as

$$OP = \sum_{j=1}^J \|\Omega_0^{(j)} - \Omega^{(j)}\|.$$

This loss roughly gives us some information about the closeness of the whole structure of the estimates and the true matrices.

4.2 Simulation with Respect to Different Data Sets

Example 1: In this study, 4 data sets are generated with sample size 70, 100, 200, 400 respectively. The structures of these four data sets are the same while the sample sizes vary. There are two groups in each data set. For the first group, data are generated from an $AR(1)$ model. Each variable is only related to its previous neighbor. The structure of data in the second group follows an $AR(2)$ model. Each variable is affected by its previous two neighbors.

The autoregressive matrices L_1 and L_2 are accordingly set as follows

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -0.5 & 1 & 0 & 0 & \dots & 0 \\ 0 & -0.5 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \\ 0 & 0 & \dots & 0 & -0.5 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -0.5 & 1 & 0 & 0 & \dots & 0 \\ -0.5 & -0.5 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \\ 0 & 0 & \dots & -0.5 & -0.5 & 1 \end{pmatrix}$$

The diagonal entries of variance matrix D_1 and D_2 are independently drawn from distribution $U(0, 1) + 1$. The covariance matrices for these two groups are set to $L_1^{-1}D_1L_1^{-1'}$ and $L_2^{-1}D_2L_2^{-1'}$ correspondingly. The dimensions are all set to 50. Simulation results are shown in Table 4.1. The ratios of Frobenius loss in different

methods as well as the ratios of Operator loss are presented in Figure 4.1. Trend of the ratios when the sample size increases can be found in this graph.

Example 2: In this example, the purpose is to simulate the performance of the new methods when the group size is growing. In order to keep consistency, the structures of all groups are set to the same and every variable in each group is independently generated, which means the autoregressive matrices are all identity matrix I_p .

Four data sets are simulated in this example. They have group size 2, 3, 5 and 10 accordingly. In the first data set, the diagonal entries in the variance matrices are all generated from distribution $U(0, 1) + 0.1$ independently. In the second data set, diagonal entries are generated from distribution $U(0, 1) + 0.3$. In the third and fourth data set, the diagonal entries in the variance matrices are generated from $U(0, 1) + 0.4$ and $U(0, 1) + 0.5$ accordingly. Dimensions are all set to 50 while numbers of observations are all set to 100. Simulation of each data set is conducted 100 times. The average Frobenius loss and Operator loss are presented in Table 4.2.

Example 3: In this simulation study, randomly generated sparse autoregressive matrices are employed. There are four data sets in this simulation study. Similar to the previous example, the group size is growing for these four data sets. They have

group size 2, 3, 5, 10 accordingly. In each data set, the autoregressive matrices in different groups are set to the same. They are generated as follows: We randomly select p positions in the lower part of an identity matrix. The values at these positions are drawn from an uniform distribution $-U(0, 0.5)$ independently. The rest positions are set to 0 except the diagonal entries. Under this setting, the number of nonzero positions equals to the dimension p in each group.

The diagonal entries of matrices D_1 , D_2 and D_3 are independently generated from distribution $U(0, 1) + 1$. Dimensions are all set to 50 while the sample sizes are all set to 100. The simulation results are listed in Table 4.3. The ratios of the Frobenius loss between joint estimation methods and separated estimation method as well as the ratios of Operator loss are presented in Figure 4.2.

Example 4: In the fourth simulation study, we exam the performance of the new methods when the data have different degrees of similarity. Three data sets are simulated. In each data set, there are three groups. The data set is generated as follows: We denote by set \mathcal{P} the collection of the positions in the lower triangular part of a $p \times p$ matrix. Four disjoint subsets with size k are randomly selected from \mathcal{P} . They are denoted by \mathcal{P}_0 , \mathcal{P}_1 , \mathcal{P}_2 and \mathcal{P}_3 . Denote by l_1 , l_2 and l_3 three identity matrices.

Group 1: Setting the values at positions in $\mathcal{P}_0 \cup \mathcal{P}_1$ of l_1 to -0.5.

Group 2: Setting the values at positions in $\mathcal{P}_0 \cup \mathcal{P}_2$ of l_2 to -0.5.

Group 3: Setting the values at positions in $\mathcal{P}_0 \cup \mathcal{P}_3$ of l_3 to -0.5.

Thus, each of these three matrices has $2k$ nonzero elements except the diagonal entries. For each pair of them, there are $2k$ positions at which they have different values.

Denote by D_1 a diagonal matrix. The diagonal entries of D_1 are independently generated from $U(0, 0.5) + 0.5$. Diagonal matrices D_2 and D_3 are generated the same as D_1 . So the covariance matrices for these three groups are obtained by

$$\Sigma_1 = T_1^{-1}D_1T_1^{-1'}, \quad \Sigma_2 = T_2^{-1}D_2T_2^{-1'}, \quad \Sigma_3 = T_3^{-1}D_3T_3^{-1'}.$$

For the first data set, the integer k is chosen as 30. For the second data set, the integer k is chosen as 50. For the third data set, the integer k is chosen as 100. Dimensions for all groups are set to 50 and the sample sizes are all set to 100. The simulation results are shown in Table 4.4.

Example 5: The purpose of the last simulation is to compare the performance when the autoregressive matrices have many small elements. Two data sets are simulated and each of them has two groups. In each data set, the autoregressive matrices are chosen to be the same. The ij th element in the lower part of the autoregressive matrices in both groups are chosen to be ρ^{i-j} . In the first data set, the parameter ρ equals 0.2. In the second group, ρ equals 0.5 which leads to a less

sparse autoregressive matrix. The i th elements of the variance matrices in both groups follow distribution $U(0, 1) + 0.5$. Dimensions of variables are all set to 50 and the sample sizes are all 100. Both Operator loss and Frobenius loss of the estimators are given in Table 4.5.

Through all these 5 data sets, we know the performance of our new methods is at least the same as the separated estimation method in Huang et al. (2006). All the simulation results outperform the empirical method and Ledoit and Wolf's method (Ledoit and Wolf 2004).

In Table 4.1, it is clear that the Ledoit-Wolf's method and the empirical method are not comparable to the penalized approach. This mainly attributes to the underlining sparse structure of the data. The joint estimation methods with sparse group LASSO and sparse l_∞ LASSO penalty both outperform the separated estimation method. The ratios of Frobenius loss and Operator loss between joint estimation methods and separated estimation method were shown in Figure 4.1. As shown in this graph, the ratios roughly have increasing trend and approach 1 when the sample size is growing. This reminds us that the joint estimation methods do perform well when the sample size is relatively small. Information from other groups is borrowed and the estimation accuracy is improved by our joint estimation methods. When the sample size is large, performance between the joint estimation methods and separated estimation method becomes less significant.

This result is also expected. All the penalized methods performs almost the same in this circumstance.

In Table 4.2, Frobenius loss and Operator loss are very similar between different penalizing methods. There is no much difference between the numbers of common zeros neither (in the worst scenario that $J = 10$, 98% of common zeros have been identified). This may result from the fact that the autoregressive matrices are complete sparse. The LASSO penalty is already capable to recover the structure of the data. Meanwhile, our joint estimation methods are also capable to identify the structure with very little improvement. Check the detailed computation results, we find that the threshold parameter corresponding to the joint penalty part is close to 0, which means β is close to 0. This reminds us that, in this completely sparse case, the separated estimation method is good enough and performs similar to our joint estimation methods.

In example 3, the performance of the new methods when the group size is growing is tested. The numbers of common zeros that are estimated from the joint estimation methods are relatively stable compared to the number of common zeros from separated estimation method when the sample size is growing. As shown in Figure 4.2, the ratios of Frobenius loss and Operator loss between joint estimation methods and separated estimation method are always bellow 1. Meanwhile, there exists a clear trend that along with the growing of group size, the ratios

are decreasing. This shows that the joint estimation methods can actually borrow information from other groups. When the number of groups is big, we do obtain more information from the other groups and increase the prediction accuracy by applying the joint estimation methods.

In example 4, autoregressive matrices are randomly generated. The degree of similarity between groups varies for these three data sets. Through Table 4.4, it is clear that, the joint methods always outperform the separated method for different degrees of similarity. According to the simulation result of example 5, our joint estimation methods also outperform the separated estimation method when the autoregressive matrices have many small elements.

		GL-JMEC	IL-JMEC	SEP	LW	SAM
	fe	0.539(0.117)	0.476(0.107)	0.635(0.136)	1.639(0.236)	17.336(5.655)
n=70	op	3.083(0.882)	3.053(0.830)	3.562(1.031)	4.377(0.221)	28.629(6.167)
	cz	976(26.9)	978(20.1)	930(14.7)		
	fe	0.346(0.048)	0.315(0.051)	0.393(0.067)	1.334(0.079)	6.415(0.930)
n=100	op	2.433(0.4849)	2.446(0.5029)	2.799(0.646)	3.658(0.154)	14.296(1.693)
	cz	965(34.8)	944(21)	888(34.9)		
	fe	0.163(0.018)	0.134(0.018)	0.175(0.020)	0.865(0.037)	1.512(0.110)
n=200	op	1.519(0.214)	1.387(0.223)	1.680(0.280)	3.347(0.123)	5.721(0.411)
	cz	918(21.7)	947(29.0)	783(53.3)		
	fe	0.078(0.007)	0.064(0.006)	0.076(0.006)	0.431(0.021)	0.642(0.052)
n=400	op	1.019(0.096)	0.929(0.103)	1.009(0.109)	2.275(0.123)	3.300(0.301)
	cz	890(23.5)	954(23.6)	721(57.1)		

Table 4.1 Simulation result when sample size is growing.

Here fe means Frebenius loss, op means Operator loss and cz means common zeros in different groups. GL-JMEC means joint estimation method with sparse group LASSO penalty and IL-JMEC means joint estimation method with sparse l_∞ penalty. SEP represents the separated estimation method with l_1 penalty term. LW represents Ledoit and Wolf's method while SAM is the empirical method. The rest tables follow the same notations.

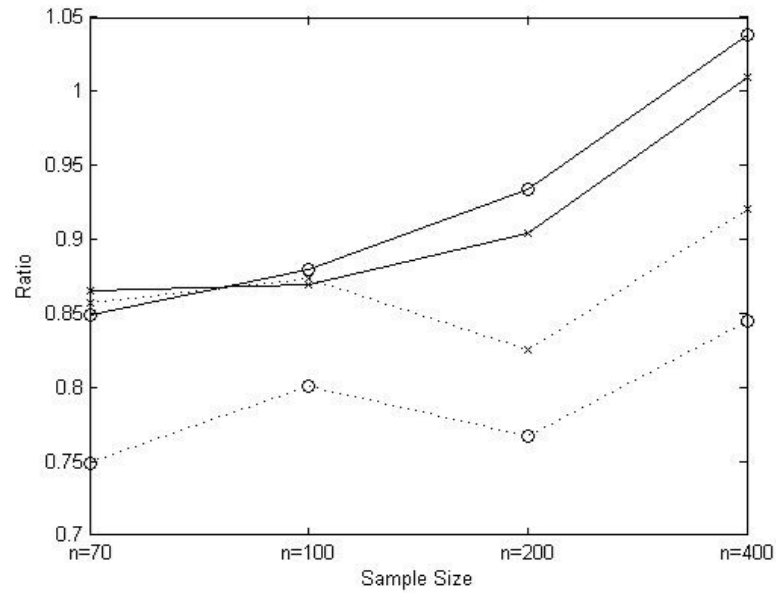


Figure 4.1 Ratio of Frobenius loss and Operator loss in example 1.

Solid line represents the ratios of the losses between GL-JMEC method and SEP method while dotted line represents the ratios of the losses between IL-JMEC method and SEP method. Circle represents the ratios of Frobenius loss between joint estimation methods and separated estimation method while star represents the ratios of Operator loss between joint estimation methods and separated estimation method. The rest figures follow the same notations.

		GL-JMEC	IL-JMEC	SEP	LW	SAM
	fe	0.416(0.169)	0.417(0.169)	0.434(0.182)	6.267(0.160)	14.335(2.480)
J=2	op	3.676(1.174)	3.680(1.176)	3.796(1.224)	12.259(0.130)	22.168(3.119)
	cz	1222(3.2)	1221(6.4)	1220(8.2)		
	fe	0.187(0.042)	0.188(0.043)	0.187(0.042)	0.999(0.021)	7.567(0.758)
J=3	op	2.526(0.465)	2.536(0.470)	2.521(0.474)	4.639(0.081)	19.141(1.766)
	cz	1219(11.8)	1214(17.9)	1217(13.3)		
	fe	0.218(0.033)	0.219(0.034)	0.222(0.033)	1.099(0.018)	8.608(0.637)
J=5	op	3.345(0.468)	3.352(0.472)	3.374(0.468)	5.7362(0.077)	25.939(1.713)
	cz	1219(8.5)	1212(16.0)	1212(13.2)		
	fe	0.316(0.033)	0.317(0.034)	0.324(0.034)	1.067(0.010)	14.4547(0.889)
J=10	op	5.602(0.565)	5.614(0.571)	5.709(0.582)	8.959(0.084)	47.5414(2.599)
	cz	1215(16.3)	1208(29.7)	1200(16.6)		

Table 4.2 Simulation result when number of groups is growing while the autoregressive matrices are identity matrix.

		GL-JMEC	IL-JMEC	SEP	LW	SAM
	fe	0.254(0.041)	0.237(0.040)	0.369(0.064)	0.642(0.022)	3.648(0.429)
J=2	op	2.321(0.459)	2.273(0.479)	2.872(0.551)	3.345(0.074)	11.030(1.440)
	cz	1075(34.8)	1037(18.3)	948(22.8)		
	fe	0.288(0.031)	0.267(0.032)	0.446(0.049)	0.986(0.023)	5.116(0.603)
J=3	op	2.813(0.340)	2.718(0.334)	3.525(0.450)	5.140(0.079)	15.757(1.597)
	cz	1051(32.4)	985(25.8)	879(15.6)		
	fe	0.484(0.048)	0.447(0.046)	0.908(0.089)	1.620(0.030)	8.355(0.732)
J=5	op	4.748(0.507)	4.648(0.510)	6.776(0.701)	8.921(0.102)	25.719(2.148)
	cz	918(18.9)	903(28.9)	670(32.4)		
	fe	0.855(0.052)	0.762(0.054)	1.742(0.085)	3.042(0.037)	16.901(1.210)
J=10	op	8.417(0.579)	8.103(0.580)	13.133(0.682)	16.116(0.117)	52.539(3.518)
	cz	802(34.2)	757(25.7)	447(21.1)		

Table 4.3 Simulation result when number of groups is growing while autoregressive matrices are randomly generated.

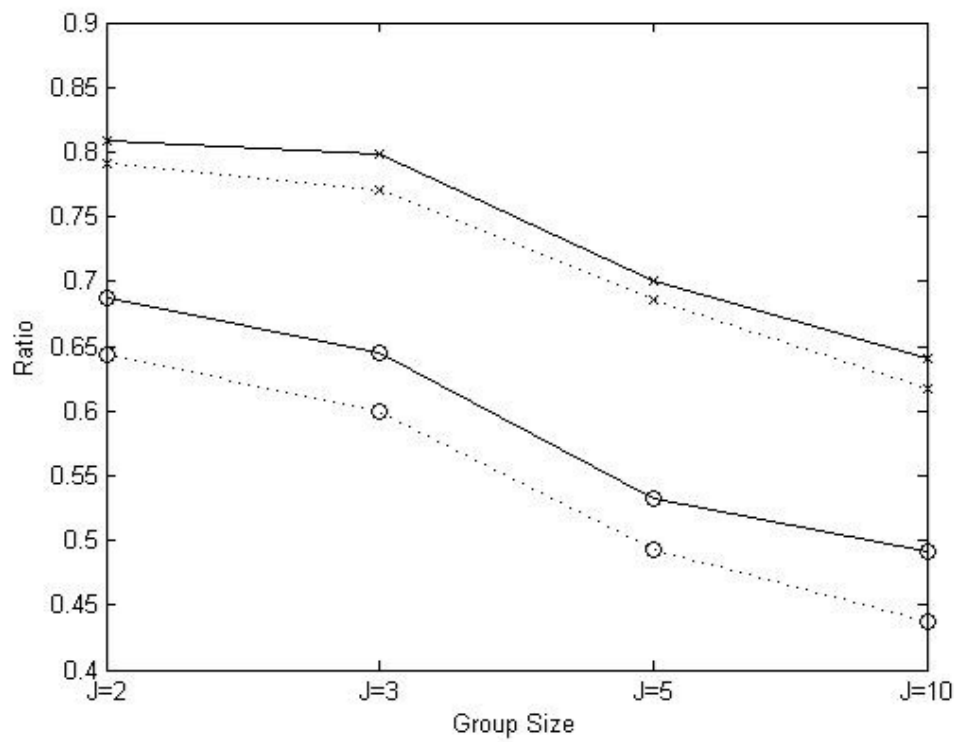


Figure 4.2 Ratio of Frobenius loss and Operator loss in example 3.

		GL-JMEC	IL-JMEC	SEP	LW	SAM
	fe	0.678(0.104)	0.675(0.099)	0.745(0.107)	1.846(0.038)	9.010(0.995)
k=30	op	4.396(0.732)	4.411(0.680)	4.676(0.791)	8.003(0.176)	21.078(2.144)
	cz	958(43.3)	958(66.3)	891(20.0)		
	fe	1.162(0.167)	1.132(0.157)	1.225(0.139)	3.046(0.166)	10.678(1.289)
k=50	op	5.677(1.036)	5.620(0.994)	5.859(0.782)	10.064(0.572)	22.873(2.492)
	cz	809(18.8)	811(20.6)	830(16.2)		
	fe	2.851(0.293)	2.661(0.301)	3.306(0.438)	5.693(0.216)	17.795(2.599)
k=100	op	9.748(1.175)	9.323(1.286)	10.775(1.639)	11.946(0.582)	30.541(3.874)
	cz	653(32.1)	613(14.1)	492(71.4)		

Table 4.4 Simulation result when data have different degrees of similarity.

		GL-JMEC	IL-JMEC	SEP	LW	SAM
	fe	0.201(0.030)	0.194(0.031)	0.240(0.035)	0.395(0.027)	3.231(0.091)
$\rho=0.2$	op	1.808(0.416)	1.814(0.422)	1.818(0.444)	2.145(0.071)	10.219(1.364)
	cz	1060(42.5)	1049(23.7)	1096(80.7)		
	fe	0.385(0.061)	0.381(0.062)	0.664(0.098)	1.426(0.253)	4.506(0.742)
$\rho=0.5$	op	2.738(0.469)	2.765(0.468)	3.714(0.624)	4.292(0.664)	11.971(1.816)
	cz	895(14.1)	878(16.3)	763(25.6)		

Table 4.5 Simulation result when when autoregressive matrices have many non zero elements.

4.3 A Real Data Set Analysis

Kenward (1987) reported an experiment about the weights of cattle in a farm. The cattle will probably be infected by roundworm which were developed from eggs on the pasture. Once a cattle was infected by roundworm, its resistance to diseases was lowered and this would affect its growth. In order to study the effect of two treatments on the weight of cattle, 60 cattle were randomly assigned to two groups with treatments A and B. In each group, there were 30 cattle. Their weights were recorded in order to study the effects of the treatments. The weights are observed 11 times in 133 days. The first 10 records were measured every 14 days and the last record was measured 7 days later. So we have 60×11 records with no missing data.

This data set was investigated extensively (see Kenward 1987; Pourahmadi 1999; Pourahmadi, 2000; Pan and Mackenzie 2003; Wu and Pourahmadi 2003). The covariance matrices are separately estimated using various methods. If one looks at the graph of weights which is presented in Figure 4.3 (solid lines represent the cattle from group A while dashed lines represent group B). It is clear that there exist two time periods. In the first period, these two groups are hardly identifiable. In the second time period. The trends of group A and group B vary. Group A still has an upward trend while group B may have a downward trend. This reminds

us that using the joint estimation approach, one may increase estimation accuracy because of the similarity of both groups in the first period. We calculate the precision matrices for both groups using different methods. In order to evaluate the performance of estimators, a discrimination study is conducted for this data set.

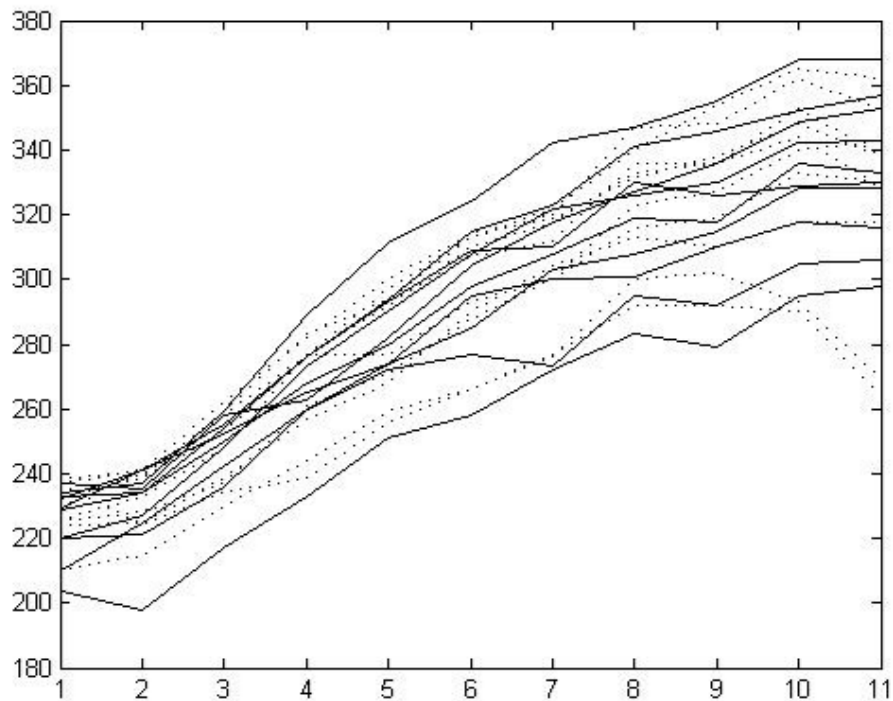


Figure 4.3 Trend of weights for the two groups of cattle.

Following T. D. Anderson (2003), we choose two discrimination functions to classify these two balanced groups. The first approach is likelihood procedure.

	GL-JMEC	IL-JMEC	SEP	LW	SAM
LDM(sd)	8.7(0.67)	7.7(0.45)	7.2(1.34)	7.5(1.65)	6.6(1.95)
QDM(sd)	8.7(0.45)	7.6(0.40)	7.4(1.61)	7.3(1.37)	6.9(1.78)

Table 4.6 Performance of discrimination study for the cattle weight data.

The second is the quadratic method. The corresponding score functions are

$$\text{LDM} : \delta_j = -\frac{n+1}{2} \log(1 + (x - \hat{\mu}_j)' \hat{\Sigma}_j^{(-1)} (x - \hat{\mu}_j)) + \frac{1}{2} \log |\hat{\Sigma}_j^{-1}|,$$

$$\text{QDM} : \delta_j = -(x - \hat{\mu}_j)' \hat{\Sigma}_j^{-1} (x - \hat{\mu}_j) + \log |\hat{\Sigma}_j^{-1}|.$$

Here $\hat{\mu}_1$ and $\hat{\mu}_2$ are the estimated mean vectors for the first and second group. In each group, 25 observations are randomly selected as training data and the rest 5 are treated as testing data. An observation x is classified into the i th group if $i = \arg \max_j \delta_j$. This procedure is repeated 10 times and the number of true predictions are recorded and the average numbers of true predictions are presented in Table 4.6.

Following our analysis setting, the testing data set has size 10. Through the likelihood ratio method, the GL-JMEC approach can predict 8.7 positive true observations and the IL-JMEC approach can predict 7.7 positive true observations. meanwhile, Ledoit and Wolf's estimators (see Ledoit and Wolf 2004), separated l_1

regularized estimators and sample covariance matrices can predict 7.5, 7.2, 6.6 true observations accordingly. The sample covariance matrices are inappropriate when one classifies these observations since the average true prediction number is close to 5 even when we randomly classify the 10 testing observations. However, according to our simulation, the number of true classified observations by using sample covariance matrices is not far from the mean value 5. Nevertheless, the joint estimation method with group LASSO penalty term (GL-JMEC) can identify 8.7 positive true observations. This result is significantly better than others and quite attractive in this discrimination study. Moreover, the variances of numbers of true predictions for our two joint estimation approaches are much smaller than the rest three methods. This reminds us that the joint estimation approach is much more stable than the separated estimation methods.

The autoregressive matrices that estimated from the IL-JMEC approach is presented as follows.

1.000	0	0	0	0	0	0	0	0	0	0
-0.806	1.000	0	0	0	0	0	0	0	0	0
-0.049	$\begin{pmatrix} -0.901 \\ -0.961 \end{pmatrix}$	1.000	0	0	0	0	0	0	0	0
0	0	-0.969	1.000	0	0	0	0	0	0	0
0	0	0	-0.985	1.000	0	0	0	0	0	0
0	0	0	-0.267	-0.782	1.000	0	0	0	0	0
0	0	0	0	0	-0.934	1.000	0	0	0	0
0	0	0	0	$\begin{pmatrix} -0.025 \\ 0.025 \end{pmatrix}$	-0.497	-0.460	1.000	0	0	0
0	0	0	0	$\begin{pmatrix} 0.012 \\ -0.012 \end{pmatrix}$	0	-0.226	-0.815	1.000	0	0
0	0	0	0	0	0	0	$\begin{pmatrix} -0.041 \\ 0 \end{pmatrix}$	-0.971	1.000	0
0	0	0	0	0	$\begin{pmatrix} 0.029 \\ -0.029 \end{pmatrix}$	0	0	-0.015	-0.961	1.000

The value at top is the estimated parameters for the first group and the value at bottom is the parameters for the second group. This also applies to the following matrices which are the autoregressive matrices that are estimated from the GL-JMEC approach.

1.000	0	0	0	0	0	0	0	0	0	0
$\begin{pmatrix} -0.857 \\ -0.800 \end{pmatrix}$	1.000	0	0	0	0	0	0	0	0	0
$\begin{pmatrix} -0.032 \\ -0.051 \end{pmatrix}$	$\begin{pmatrix} -0.864 \\ -1.011 \end{pmatrix}$	1.000	0	0	0	0	0	0	0	0
0	$\begin{pmatrix} -0.011 \\ -0.013 \end{pmatrix}$	$\begin{pmatrix} -0.925 \\ -1.015 \end{pmatrix}$	1.000	0	0	0	0	0	0	0
0	0	$\begin{pmatrix} -0.003 \\ -0.002 \end{pmatrix}$	$\begin{pmatrix} -1.033 \\ -0.956 \end{pmatrix}$	1.000	0	0	0	0	0	0
0	0	0	$\begin{pmatrix} -0.238 \\ -0.272 \end{pmatrix}$	$\begin{pmatrix} -0.788 \\ -0.787 \end{pmatrix}$	1.000	0	0	0	0	0
0	0	0	0	0	$\begin{pmatrix} -0.925 \\ -0.942 \end{pmatrix}$	1.000	0	0	0	0
0	0	0	$\begin{pmatrix} -0.021 \\ -0.078 \end{pmatrix}$	0	$\begin{pmatrix} -0.536 \\ -0.394 \end{pmatrix}$	$\begin{pmatrix} -0.461 \\ -0.459 \end{pmatrix}$	1.000	0	0	0
0	0	0	0	$\begin{pmatrix} 0.064 \\ -0.146 \end{pmatrix}$	$\begin{pmatrix} -0.003 \\ -0.131 \end{pmatrix}$	$\begin{pmatrix} -0.239 \\ -0.098 \end{pmatrix}$	$\begin{pmatrix} -0.852 \\ -0.685 \end{pmatrix}$	1.000	0	0
0	0	0	0	0	0	0	$\begin{pmatrix} -0.057 \\ -0.029 \end{pmatrix}$	$\begin{pmatrix} -0.982 \\ -0.847 \end{pmatrix}$	1.000	0
0	0	0	0	0	$\begin{pmatrix} 0.025 \\ -0.007 \end{pmatrix}$	0	0	$\begin{pmatrix} -0.086 \\ -0.018 \end{pmatrix}$	$\begin{pmatrix} -0.891 \\ -1.038 \end{pmatrix}$	1.000

The autoregressive matrices which are obtained from the separated estimation method in Huang et al.(2006) with l_1 penalty are as follows

1.000	0	0	0	0	0	0	0	0	0	0
$\begin{pmatrix} -0.905 \\ -0.8001 \end{pmatrix}$	1.000	0	0	0	0	0	0	0	0	0
$\begin{pmatrix} -0.010 \\ -0.072 \end{pmatrix}$	$\begin{pmatrix} -0.892 \\ -0.998 \end{pmatrix}$	1.000	0	0	0	0	0	0	0	0
$\begin{pmatrix} 0 \\ 0.1232 \end{pmatrix}$	$\begin{pmatrix} -0.004 \\ -0.1387 \end{pmatrix}$	$\begin{pmatrix} -0.943 \\ -1.011 \end{pmatrix}$	1.000	0	0	0	0	0	0	0
$\begin{pmatrix} 0 \\ -0.015 \end{pmatrix}$	0	$\begin{pmatrix} -0.039 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1.016 \\ -0.9504 \end{pmatrix}$	1.000	0	0	0	0	0	0
0	0	0	$\begin{pmatrix} -0.236 \\ -0.2855 \end{pmatrix}$	$\begin{pmatrix} -0.801 \\ -0.776 \end{pmatrix}$	1.000	0	0	0	0	0
0	$\begin{pmatrix} 0.010 \\ 0 \end{pmatrix}$	0	0	0	-0.943	1.000	0	0	0	0
0	0	0	$\begin{pmatrix} 0 \\ -0.216 \end{pmatrix}$	$\begin{pmatrix} -0.045 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -0.553 \\ -0.223 \end{pmatrix}$	$\begin{pmatrix} -0.434 \\ -0.513 \end{pmatrix}$	1.000	0	0	0
$\begin{pmatrix} -0.036 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.010 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0.010 \end{pmatrix}$	$\begin{pmatrix} 0.165 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -0.0726 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -0.388 \end{pmatrix}$	$\begin{pmatrix} -0.298 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -0.858 \\ -0.601 \end{pmatrix}$	1.000	0	0
0	0	0	0	0	$\begin{pmatrix} 0 \\ -0.192 \end{pmatrix}$	0	$\begin{pmatrix} -0.148 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -0.908 \\ -0.699 \end{pmatrix}$	1.000	0
$\begin{pmatrix} -0.020 \\ 0 \end{pmatrix}$	0	0	$\begin{pmatrix} 0 \\ 0.129 \end{pmatrix}$	0	$\begin{pmatrix} 0.130 \\ 0 \end{pmatrix}$	0	0	$\begin{pmatrix} -0.230 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -0.833 \\ -1.064 \end{pmatrix}$	1.000

The estimates from IL-JMEC and GL-JMEC remind us that the underlying structures for both groups are close to an AR(2) model. When the observation time of two records are far away, these autoregressive matrix estimates show that the corresponding autoregressive parameters are 0, which coincides with our experience. At the other side, the autoregressive matrices estimated from separated estimation approach are relatively noisy. It fails to reveal the underlying structure.

CHAPTER 5

Conclusion

This study investigated the joint covariance matrix estimation methods based on multiple groups of ordered data. The reparameterizing technique of modified Cholesky decomposition was used in this research. Penalty functions with group effect were imposed to the Cholesky factors based on log normal likelihood function. The penalty functions were chosen to be sparse group LASSO, and sparse l_∞ LASSO which may perform group selection effect. Consistent properties of the resulting estimates from both approaches were explored. The simulation results showed that the joint estimation methods outperform the separated estimation methods and achieved smaller prediction error.

The simulation results in Chapter 4 showed that both joint estimation methods and also the separated penalized estimation method outperform the empirical estimation method when the true covariance matrices are sparse. This may attribute to the denoise function of the penalizing approaches. When the autoregressive matrices are totally sparse which means they are all identity matrices, the simulation results showed that our new joint estimation methods perform almost the same as the separated estimation method. All the penalizing methods including the separated estimation method and our joint estimation methods achieved the true structure of the autoregressive matrices. So the differences between them are not significant.

The joint approaches perform better when the autoregressive matrices of different groups have homogeneous structures. Both Operator loss and Frobenius loss for the joint approaches are smaller than the losses from separated approach. The idea of borrowing information from other groups is achieved through the joint estimation approaches. This is very attractive when one analyzes multiple groups data with homogeneous structures. As we expected, more true common zeros in the autoregressive matrices were generated in the joint approaches according to our simulation results. Common links in different groups are revealed by our new methods.

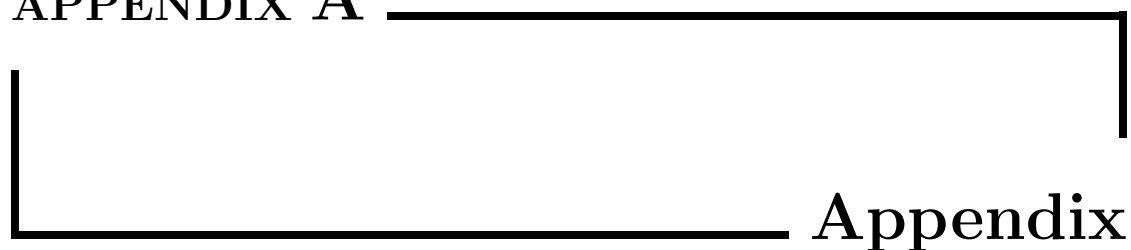
An application of our new methods on a real data set was conducted. The results of classification study of the cattle data set which was mentioned in Kenward (1987) showed that the sparse group LASSO approach has the best classification accuracy which is significantly better than the empirical approach and the separated estimation approach. This study illustrates the power of our new methods when we handle multiple groups of data. Our approaches can efficiently shrink a group of parameters towards zeros if the true parameters are all zero. This effect can make the homogeneous parts close which in turns emphasizes the inhomogeneous parts.

It has to be noted that the whole procedure is based on multiple groups data which have a natural order. This assumption is needed when we apply the Cholesky decomposition which guarantees the positive definiteness of the resulting covariance matrices and precision matrices. If the order assumption fails, the Cholesky decomposition is statistically meaningless and the performance of the joint estimation methods can not be guaranteed. Moreover, this study did not consider the procedure that tests the homogeneity of the autoregressive parameters in different groups. The homogeneity assumption is only based on our prior information about the data. Nevertheless, most of the time, prior information is sufficient enough for us to identify whether homogeneity assumption is valid.

Joint estimation of covariance matrices and precision matrices need to be further investigated since multiple groups data are common. In this study, the sparse l_∞ LASSO and sparse group LASSO are used. In order to further reduce bias, one can try the adapted version of these penalties.

In this study, the convergence rate is obtained by constraining the number of nonzero elements. It is possible that the result can be improved by constraining the structure of the underlying autoregressive matrices like Bickel and Levina (2008a) or Cai and Liu (2011) did. The convergence property of the estimates can be further investigated.

APPENDIX A



Appendix

A.1 Three Lemmas

In this chapter, we denote $s_i(A)$ the i th singular value of matrix A . $s_1(A)$, $s_2(A)$, . . . $s_p(A)$ are in a descending order. Before we prove the theorems, we state some lemmas first.

Lemma A.1. *Let A and B be two matrices with order $p \times n$ and $n \times m$. For any $i, j \geq 0$, we have*

$$s_{i+j+1}(AB) \leq s_{i+1}(A)s_{j+1}(B),$$

and

$$s_p(A)\|B\|_F \leq \|AB\|_F \leq s_1(A)\|B\|_F \leq \|A\|_F\|B\|_F.$$

Especially if B is an identity matrix, we have

$$s_p(A) \leq \|A\|_F/p \leq s_1(A) \leq \|A\|_F.$$

The proof can be found in Bai and Silverstein (2010).

Lemma A.2. Assume a sequence of positive definite matrices Σ_n have corresponding Cholesky decompositions

$$T_n \Sigma_n T_n' = D_n,$$

and the singular values of matrices Σ_n are bounded which means there exist c_1 and c_2 such that $0 < c_1 < s_p(\Sigma_n) < s_1(\Sigma) < c_2 < \infty$, then there exist constants d_1, d_2 such that

$$d_1 < s_p(T_n) \leq s_1(T_n) < d_2,$$

and

$$d_1 < s_p(D_n) \leq s_1(D_n) < d_2.$$

Proof: For an arbitrary matrix in the sequence Σ_n , say Σ , assume it has decomposition

$$\Sigma = RR', \tag{A.1}$$

where R is a lower triangular matrix with ij th element r_{ij} ($i \geq j$) or 0 ($i > j$) where we assume $r_{ii} > 0$. Using simple algebra, we have $\sigma_{ii} = \sum_{i \leq j} r_{ij}^2$ where σ_{ii}

is the i th diagonal element of matrix Σ . This implies

$$r_{ii}^2 \leq \sigma_{ii}.$$

Denote e_i a p dimensional vector with the i th element equals 1 and the others equal 0, then we have

$$r_{ii}^2 \leq \sigma_{ii} = e_i^T \Sigma e_i \leq s_1(\Sigma) \leq c_2, \quad (\text{A.2})$$

for $i=1, 2, \dots, p$. Set $D=\text{diag}\{r_{11}^2, r_{22}^2, \dots, r_{pp}^2\}$, then it can be induced from (A.2) that $s_1(D) \leq c_2$.

By (A.1) we have the following decomposition

$$D^{\frac{1}{2}} R^{-1} \Sigma R'^{-1} D^{\frac{1}{2}} = D.$$

Denote $T = D^{\frac{1}{2}} R^{-1}$, then matrix T is a lower triangular matrix with 1 on its diagonal. Rewriting the above equation, it becomes $T \Sigma T' = D$. This is exactly the modified Cholesky decomposition. Take determinant on both sides, we have $|T \Sigma T'| = |D|$ which induces

$$s_1(D) s_2(D) \dots s_p(D) = s_1(\Sigma) s_2(\Sigma) \dots s_p(\Sigma).$$

We already proved that $s_1(D) \leq c_2$. Therefore

$$s_p^p(\Sigma) \leq \prod_{i=1}^p s_i(\Sigma) = \prod_{i=1}^p s_i(D) \leq c_2^{p-1} s_p(D),$$

which implies $s_p(D) \geq \frac{c_1^p}{c_2^{p-1}}$. Denote $d_1 = \frac{c_1^p}{c_2^{p-1}}$ and $d_2 = c_2$, we have

$$0 < d_1 \leq s_p(D) \leq s_1(D) \leq d_2 < \infty. \quad (\text{A.3})$$

By Lemma 1 we have

$$s_p(D) = s_p(T\Sigma T') \leq s_p(T)s_1(\Sigma)s_p(T').$$

Thus, $s_p(T) \geq \sqrt{s_p(D)/s_1(\Sigma)} \geq \sqrt{d_1/c_2}$. Recall that $T\Sigma T' = D$, this induces $\Sigma = T^{-1}DT^{-1'}$. Apply Lemma 1 again, we have

$$s_p(\Sigma) = s_p(T^{-1}DT^{-1'}) \leq s_p^2(T^{-1})s_1(D) = 1/s_1^2(T)s_1(D).$$

Consequently, we can bound $s_1(T)$ from above by $\sqrt{s_1(D)/s_p(\Sigma)} = \sqrt{d_2/c_2}$.

Choosing d_1 small enough and d_2 big enough, the following inequality hold,

$$d_1 \leq s_p(T) \leq s_1(T) \leq d_2. \quad (\text{A.4})$$

Combine (A.3) and (A.4), we proved Lemma 2. This Lemma shows that the singular values of the corresponding autoregressive matrix and the variance matrix are all bounded once the singular values of the covariance matrix are bounded.

Next we restate a lemma which was given in Lam and Fan (2009).

Lemma A.3. *Let S be the sample covariance matrix of observations which are drawn from distribution $N(0, \Sigma_0)$. Let A and B be two $p \times p$ matrices which satisfy*

$\|A\| = O_p(1)$ and $\|B\| = O_p(1)$. Assume matrix S , Σ_0 , A and B are growing with dimension p while $p/n \rightarrow c \in [0, 1)$, then we have

$$\max_{i,j} |(A(S - \Sigma_0)B)_{i,j}| = O_p(\sqrt{\log(p)/n}).$$

A.2 Proof of Theorems

Proof of Theorem 3.1: The proof is enlighten by the proof in Rothman et al. (2008). Let D represent set $\{D^{(1)}, D^{(2)}, \dots, D^{(J)}\}$ and T represent $\{T^{(1)}, T^{(2)}, \dots, T^{(J)}\}$ where $D^{(j)}$ are diagonal matrices and $T^{(j)}$ are lower triangular matrices with one on their diagonal. Denote

$$Q(D, T) = \sum_{j=1}^J (\log |D^{(j)}| + \text{tr}(T'^{(j)} D^{-1} T^{(j)} S^j)) + \sum_{r < s} g_{\lambda, \beta}(\phi_{rs}^{(1)}, \dots, \phi_{rs}^{(J)}).$$

Let $G(\Delta_D, \Delta_T) = Q(D_0 + \Delta_D, T_0 + \Delta_T) - Q(D_0, T_0)$. Denote $\mathcal{A}_{U_1} = \{\Delta_T : \sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2 \leq U_1^2 s \log(p)/n\}$ and $\mathcal{B}_{U_2} = \{\Delta_D : \sum_{j=1}^J \|\Delta_D^{(j)}\|_F^2 \leq U_2^2 p \log(p)/n\}$. We will prove that for every $\Delta_T \in \partial \mathcal{A}_{U_1}$ and $\Delta_D \in \partial \mathcal{B}_{U_2}$, probability $P(G(\Delta_T, \Delta_D) > 0)$ is tending to 1 for sufficiently large U_1 and U_2 . Here $\partial \mathcal{A}_{U_1}$ and $\partial \mathcal{B}_{U_2}$ represent the boundary of \mathcal{A}_{U_1} and \mathcal{B}_{U_2} . Because $G(\Delta_T, \Delta_D)$ will achieve 0 when $\Delta_T = 0$ and $\Delta_D = 0$, we know the minimal point of $G(\Delta_D, \Delta_T)$ is achieved when $\Delta_T \in \mathcal{A}_{U_1}$ and $\Delta_D \in \mathcal{B}_{U_2}$. That is $\sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2 = O_p(s \log(p)/n)$ and $\sum_{j=1}^J \|\Delta_D^{(j)}\|_F^2 = O_p(p \log(p)/n)$.

Assume $\sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2 = U_1^2 s \log(p)/n$, $\sum_{j=1}^J \|\Delta_D^{(j)}\|_F^2 = U_2^2 p \log(p)/n$. According to the assumption, we know there exists a constant d such that $0 < 1/d \leq s_p(\Sigma_0^{(j)}) \leq s_1(\Sigma_0^{(j)}) \leq d < \infty$. By Lemma 2, without losing any generality, we can also assume $0 < 1/d \leq s_p(T_0^{(j)}) \leq s_1(T_0^{(j)}) \leq d < \infty$ and $0 < d \leq s_p(D_0^{(j)}) \leq s_1(D_0^{(j)}) \leq d < \infty$.

Denote $E^{(j)} = D^{(j)-1}$, $E_0^{(j)} = D_0^{(j)-1}$ and $\Delta_E^{(j)} = E^{(j)} - E_0^{(j)}$. Recall that the singular values of diagonal matrices $D^{(j)}$ and $D_0^{(j)}$ are bounded, therefore the diagonal matrices $E_0^{(j)}$ also satisfies the inequality $1/d \leq s_p(E_0^{(j)}) \leq s_1(E_0^{(j)}) \leq d$. Due to the constraint $\sum_{j=1}^J \|\Delta_D^{(j)}\|_F^2 = U_2^2 p \log(p)/n$, it is straightforward to conclude that $1/d^2 U_2^2 p \log(p)/n \leq \sum_{j=1}^J \|\Delta_E^{(j)}\|_F^2 \leq d^2 U_2^2 p \log(p)/n$. Expand $\log |D_0^{(j)}| - \log |D^{(j)}| = \log |E_0^{(j)} + \Delta_E^{(j)}| - \log |E_0^{(j)}|$ into integration form using Taylor's expansion, we have

$$\begin{aligned} & \log |D_0^{(j)}| - \log |D^{(j)}| \\ &= \log |E_0^{(j)} + \Delta_E^{(j)}| - \log |E_0^{(j)}| \\ &= \text{tr} \Delta_E^{(j)} D_0^{(j)} - (\text{Vec} \Delta_E^{(j)})^T \int_0^1 (1-v) E_v^{(j)-1} \otimes E_v^{(j)-1} dv (\text{Vec} \Delta_E^{(j)}), \end{aligned}$$

where $E_v^{(j)} = E_0^{(j)} + v \Delta_E^{(j)}$. Using this decomposition, we can divide $G(\Delta_T, \Delta_D)$ into several parts.

$$\begin{aligned} & G(\Delta_T, \Delta_D) \\ &= - \sum_{j=1}^J \text{tr} \Delta_E^{(j)} D_0^{(j)} + \sum_{j=1}^J (\text{Vec} \Delta_E^{(j)})^T \int_0^1 (1-v) E_v^{-1} \otimes E_v^{-1} dv (\text{Vec} \Delta_E^{(j)}) \end{aligned}$$

$$\begin{aligned}
& + \sum_{j=1}^J (\text{tr} T^{(j)'} D^{(j)-1} T^{(j)} S - \text{tr} T_0^{(j)'} D_0^{(j)-1} T_0^{(j)} S) \\
& + \sum_{r>s} (g_{\lambda,\beta}(\phi_{rs}^{(1)}, \dots, \phi_{rs}^{(J)}) - g_{\lambda,\beta}(\phi_{0rs}^{(1)}, \dots, \phi_{0rs}^{(J)})) \\
= & \sum_{j=1}^J (\text{Vec} \Delta_E^{(j)})^T \int_0^1 (1-v) E_v^{-1} \otimes E_v^{-1} dv (\text{Vec} \Delta_E^{(j)}) \\
& - \sum_{j=1}^J \text{tr} \Delta_E^{(j)} D_0^{(j)} + \sum_{j=1}^J \text{tr} (T^{(j)'} D^{(j)-1} T^{(j)} S - \text{tr} T^{(j)'} D_0^{(j)-1} T^{(j)} S) \\
& + \sum_{j=1}^J (\text{tr} T^{(j)'} D_0^{(j)-1} T^{(j)} S - \text{tr} T_0^{(j)'} D_0^{(j)-1} T_0^{(j)} S) \\
& + \sum_{r>s} (g_{\lambda,\beta}(\phi_{rs}^{(1)}, \dots, \phi_{rs}^{(J)}) - g_{\lambda,\beta}(\phi_{0rs}^{(1)}, \dots, \phi_{0rs}^{(J)})) \\
= & \sum_{j=1}^J (\text{Vec} \Delta_E^{(j)})^T \int_0^1 (1-v) E_v^{-1} \otimes E_v^{-1} dv \text{Vec} \Delta_E^{(j)} \\
& + \sum_{j=1}^J \text{tr} (D^{(j)-1} - D_0^{(j)-1}) [T^{(j)} (S^{(j)} - \Sigma_0^{(j)}) T^{(j)'}] \\
& + \sum_{j=1}^J \text{tr} D_0^{(j)-1} (T^{(j)} S^{(j)} T^{(j)'} - T_0^{(j)} S^{(j)} T_0^{(j)'}) \\
& + \sum_{j=1}^J \text{tr} (D^{(j)-1} - D_0^{(j)-1}) (T^{(j)} \Sigma_0^{(j)} T^{(j)'} - D_0^{(j)}) \\
& + \sum_{rs \in Z^c} (g_{\lambda,\beta}(\phi_{rs}^{(1)}, \dots, \phi_{rs}^{(J)}) - g_{\lambda,\beta}(\phi_{0rs}^{(1)}, \dots, \phi_{0rs}^{(J)})) \\
& + \sum_{rs \in Z} (g_{\lambda,\beta}(\phi_{rs}^{(1)}, \dots, \phi_{rs}^{(J)}) - g_{\lambda,\beta}(\phi_{0rs}^{(1)}, \dots, \phi_{0rs}^{(J)})).
\end{aligned}$$

Thus according to our two estimation methods, the function $Q(\Delta_T, \Delta_D)$ can be

divided into $M_1 + M_2 + M_3 + I_1 + I_2$ or $M_1 + M_2 + M_3 + G_1 + G_2$. Here

$$M_1 = \sum_{j=1}^J (\text{Vec} \Delta_E^{(j)})^T \int_0^1 (1-v) E_v^{(j)-1} \otimes E_v^{(j)-1} dv \text{Vec} \Delta_E^{(j)},$$

$$\begin{aligned}
M_2 &= \sum_{j=1}^J \text{tr}(D^{(j)-1} - D_0^{(j)-1})[T^{(j)}(S^{(j)} - \Sigma_0^{(j)})T^{(j)'}], \\
M_3 &= \sum_{j=1}^J \text{tr}D_0^{(j)-1}(T^{(j)}S^{(j)}T^{(j)'} - T_0^{(j)}S^{(j)}T_0^{(j)'}) \\
&\quad + \sum_{j=1}^J \text{tr}(D^{(j)-1} - D_0^{(j)-1})(T^{(j)}\Sigma_0^{(j)}T^{(j)'} - D_0^{(j)}), \\
I_1 &= \beta \sum_{rs \in Z^c} \text{Max}_{j=1}^J |t_{rs}^{(j)}| + \lambda \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}|, \\
I_2 &= \beta \sum_{rs \in Z} (\text{Max}_{j=1}^J |t_{rs}^{(j)}| - \text{Max}_{j=1}^J |t_{0rs}^{(j)}|) + \lambda \sum_{rs \in Z} \sum_{j=1}^J (|t_{rs}^{(j)}| - |t_{0rs}^{(j)}|), \\
G_1 &= \beta \sum_{rs \in Z^c} \sqrt{\sum_{j=1}^J t_{rs}^{(j)2}} + \lambda \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}|, \\
G_2 &= \beta \sum_{rs \in Z} \left(\sqrt{\sum_{j=1}^J t_{rs}^{(j)2}} - \sqrt{\sum_{j=1}^J t_{0rs}^{(j)2}} \right) + \lambda \sum_{rs \in Z} \sum_{j=1}^J (|t_{rs}^{(j)}| - |t_{0rs}^{(j)}|).
\end{aligned}$$

We consider these terms separately. For the first term,

$$M_1 = \sum_{j=1}^J (\text{Vec} \Delta_E^{(j)})^T \int_0^1 (1-v) E_v^{(j)-1} \otimes E_v^{(j)-1} dv (\text{Vec} \Delta_E^{(j)}).$$

Recall that $E_v^{(j)} = E_0^{(j)} + v \Delta_E^{(j)}$, using the inequality of the operator norm, we have $\|E_v^{(j)}\| \leq \|E_0^{(j)}\| + v \|\Delta_E^{(j)}\|$. By Lemma 1, we have $\|\Delta_E^{(j)}\| \leq \|\Delta_E^{(j)}\|_F = O(\sqrt{p \log(p)/n}) = o(1)$, thus $\|E_v^{(j)}\| \leq \|E_0^{(j)}\| + o(1) \leq 2\|E_0^{(j)}\| \leq 2d$. Therefore,

$$\begin{aligned}
M_1 &= \sum_{j=1}^J \int_0^1 (1-v) (\text{Vec} \Delta_E^{(j)})^T E_v^{-1} \otimes E_v^{-1} (\text{Vec} \Delta_E^{(j)}) dv \\
&\geq \sum_{j=1}^J \|\text{Vec} \Delta_E^{(j)}\|_2^2 \int_0^1 (1-v) s_{\min}(E_v^{-1} \otimes E_v^{-1}) dv \\
&\geq \sum_{j=1}^J \|\Delta_E^{(j)}\|_F^2 \int_0^1 (1-v) s_{\min}^2(E_v^{-1}) dv
\end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^J \|\Delta_E^{(j)}\|_F^2 \int_0^1 (1-v) 1/s_{\max}^2(E_v) dv \\
&\geq 1/8d^2 \sum_{j=1}^J \|\Delta_E^{(j)}\|_F^2.
\end{aligned}$$

As we have mentioned at the beginning of this proof, $\|\Delta_E^{(j)}\|_F^2 \geq 1/d^2 \|\Delta_D^{(j)}\|_F^2$, therefore $M_1 \geq 1/8d^4 \|\Delta_D^{(j)}\|_F^2$.

Bounding the term M_2 is relatively easy. Recall that $\|T_0^{(j)}\| = O(1)$ and $\|\Delta_T^{(j)}\| = o(1)$, by Lemma 3 we know, for arbitrary ϵ there exist V_2 which only related to ϵ and n such that $p(\max |(T_0^{(j)}(S^{(j)} - \Sigma_0^{(j)})T_0^{(j)'})_{rs}| > V_2\sqrt{\log(p)/n}) < \epsilon/4$. Use Lemma 3 again, we know there exists C such that $p(\max |(\frac{\Delta_T^{(j)}}{\|\Delta_T^{(j)}\|}(S^{(j)} - \Sigma_0^{(j)})T_0^{(j)'})_{rs}| > C\sqrt{\log(p)/n}) < \epsilon/4$ which implies

$$p(\max |(\Delta_T^{(j)}(S^{(j)} - \Sigma_0^{(j)})T_0^{(j)'})_{rs}| > C\|\Delta_T^{(j)}\|\sqrt{\log(p)/n}) < \epsilon/4.$$

When n sufficiently large such that $C\|\Delta_T^{(j)}\| < V_2$, we have

$$p(\max |(\Delta_T^{(j)}(S^{(j)} - \Sigma_0^{(j)})T_0^{(j)'})_{rs}| > V_2\sqrt{\log(p)/n}) < \epsilon/4.$$

The same

$$p(\max |(T_0^{(j)}\Delta_T^{(j)}(S^{(j)} - \Sigma_0^{(j)})\Delta_0^{(j)'})_{rs}| > V_2\sqrt{\log(p)/n}) < \epsilon/4,$$

and

$$p(\max |(\Delta_0^{(j)}\Delta_T^{(j)}(S^{(j)} - \Sigma_0^{(j)})\Delta_0^{(j)'})_{rs}| > V_2\sqrt{\log(p)/n}) < \epsilon/4.$$

Combine all these four together, we know $p(\max |(T^{(j)}(S^{(j)} - \Sigma_0^{(j)})T^{(j)'})_{rs}| > V_2 \sqrt{\log(p)/n}) < \epsilon$.

So with probability greater than $1 - \epsilon$, we have

$$\begin{aligned}
|M_2| &= \left| \sum_{j=1}^J \text{tr}(D^{(j)-1} - D_0^{(j)-1})[T^{(j)}(S^{(j)} - \Sigma_0^{(j)})T^{(j)'}] \right| \\
&\leq \sum_{j=1}^J \max |(T^{(j)}(S^{(j)} - \Sigma_0^{(j)})T^{(j)'})_{rs}| \|D^{(j)-1} - D_0^{(j)-1}\|_1 \\
&\leq V_2 \sqrt{\log(p)/n} \sum_{j=1}^J \|D^{(j)} - D_0^{(j)}\|_1 \\
&\leq V_2 \log(p)/n \sqrt{pJ \sum_{j=1}^J \|\Delta_D\|_F^2}
\end{aligned}$$

The term M_3 is relatively complicated. We can rewrite M_3 as

$$\begin{aligned}
&\sum_{j=1}^J \text{tr} D_0^{(j)-1} (T^{(j)} S^{(j)} T^{(j)'} - T_0^{(j)} S^{(j)} T_0^{(j)'}) \\
&+ \sum_{j=1}^J \text{tr} (D^{(j)-1} - D_0^{(j)-1}) (T^{(j)} \Sigma_0^{(j)} T^{(j)'} - D_0^{(j)}) \\
&= \sum_{j=1}^J \text{tr} D_0^{(j)-1} [T^{(j)} (S^{(j)} - \Sigma_0^{(j)}) T^{(j)'} - T_0^{(j)} (S^{(j)} - \Sigma_0^{(j)}) T_0^{(j)'}] \\
&+ \sum_{j=1}^J \text{tr} D_0^{(j)-1} (T^{(j)} \Sigma_0^{(j)} T^{(j)'} - T_0^{(j)} \Sigma_0^{(j)} T_0^{(j)'}) \\
&+ \sum_{j=1}^J \text{tr} (D^{(j)-1} - D_0^{(j)-1}) (T^{(j)} \Sigma_0^{(j)} T^{(j)'} - D_0^{(j)}) \\
&= \sum_{j=1}^J \text{tr} D_0^{(j)-1} [T^{(j)} (S^{(j)} - \Sigma_0^{(j)}) T^{(j)'} - T_0^{(j)} (S^{(j)} - \Sigma_0^{(j)}) T_0^{(j)'}] \\
&+ \sum_{j=1}^J \text{tr} D_0^{(j)-1} (T^{(j)} \Sigma_0^{(j)} T^{(j)'} - T_0^{(j)} \Sigma_0^{(j)} T_0^{(j)'}) \\
&= \sum_{j=1}^J \text{tr} D_0^{(j)-1} [\Delta_T^{(j)} (S^{(j)} - \Sigma_0^{(j)}) T_0^{(j)'} + T_0^{(j)} (S^{(j)} - \Sigma_0^{(j)}) \Delta_T^{(j)'} + \Delta_T^{(j)} (S^{(j)} - \Sigma_0^{(j)}) \Delta_T^{(j)'}]
\end{aligned}$$

$$+ \sum_{j=1}^J \text{tr} D^{(j)-1} (\Delta_T^{(j)} \Sigma_0^{(j)} T_0^{(j)'} + T_0^{(j)} \Sigma_0^{(j)} \Delta_T^{(j)'} + \Delta_T^{(j)} \Sigma_0^{(j)} \Delta_T^{(j)}).$$

Thus M_3 can be decomposed into $L_1 + L_2$, where L_1 equals

$$\sum_{j=1}^J \text{tr} D_0^{(j)-1} [\Delta_T^{(j)} (S^{(j)} - \Sigma_0^{(j)}) T_0^{(j)'} + T_0^{(j)} (S^{(j)} - \Sigma_0^{(j)}) \Delta_T^{(j)'} + \Delta_T^{(j)} (S^{(j)} - \Sigma_0^{(j)}) \Delta_T^{(j)'}]$$

and L_2 equals

$$\sum_{j=1}^J \text{tr} D^{(j)-1} (\Delta_T^{(j)} \Sigma_0^{(j)} T_0^{(j)'} + T_0^{(j)} \Sigma_0^{(j)} \Delta_T^{(j)'} + \Delta_T^{(j)} \Sigma_0^{(j)} \Delta_T^{(j)}).$$

Since $T_0^{(j)} \Sigma_0^{(j)} T_0^{(j)'} = D_0^{(j)}$, we know $\Sigma_0^{(j)} T_0^{(j)'} = T_0^{(j)-1} D_0^{(j)}$ is a lower triangle matrix, therefore $\Sigma_0^{(j)} T_0^{(j)'} D^{(j)-1}$ is also a lower triangle matrix. As we already know that $\Delta_T^{(j)}$ is also a lower triangle matrix in which the diagonal entries are all zero, thus $\text{tr} D^{(j)-1} \Delta_T^{(j)} \Sigma_0^{(j)} T_0^{(0)'} = 0$. With the same argument we have $\text{tr} \Delta_D^{(j)-1} T_0^{(0)} \Sigma_0^{(j)} \Delta_T^{(j)'} = 0$. So

$$L_2 = \sum_{j=1}^J \text{tr} D^{(j)-1} \Delta_T^{(j)} \Sigma_0^{(j)} \Delta_T^{(j)'} = \sum_{j=1}^J \text{Vec}(\Delta_T^{(j)})^T \Sigma_0^{(j)} \otimes D^{(j)-1} \text{Vec}(\Delta_T^{(j)}).$$

Since $\|\Delta_D^{(j)}\|_F = \|D^{(j)} - D_0^{(j)}\|_F = O(\sqrt{p \log(p)/n}) = o(1)$, $\|\Delta_D^{(j)}\| \leq \|\Delta_D^{(j)}\|_F = o(1)$, we know $\|D^{(j)}\| \leq \|D_0^{(j)}\| + \|\Delta_D^{(j)}\| = \|D_0^{(j)}\| + o(1) \leq 2d$, therefore $L_2 = \sum_{j=1}^J \text{Vec}(\Delta_T^{(j)})^T \Sigma_0^{(j)} \otimes D^{(j)-1} \text{Vec}(\Delta_T^{(j)}) \geq \sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2 s_{\min}(\Sigma_0^{(j)}) s_{\min}(D^{(j)-1}) \geq 1/2d^2 \sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2$.

Let us go back to L_1 . Using the Theorem (5.11) in Bai and Silverstein (2010), we know $\|S^{(j)} - \Sigma_0^{(j)}\| = o_p(1)$, therefore $\sum_{j=1}^J \text{tr} D_0^{(j)-1} \Delta_T^{(j)} (S^{(j)} - \Sigma_0^{(j)}) \Delta_T^{(j)'} =$

$\sum_{j=1}^J \text{Vec}(\Delta_T^{(j)})^T D_0^{(j)-1} \otimes (S^{(j)} - \Sigma_0^{(j)}) \text{Vec}(\Delta_T^{(j)}) \leq o_p(1) \sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2$. This part will be dominated by the positive term L_2 . Since $\|D_0^{(j)}\| = O(1)$ and $\|T_0^{(j)}\| = O(1)$, applying Lemma A.3 again, for ϵ , we can find V_1 such that $p(|((S^{(j)} - \Sigma_0^{(j)})T_0^{(j)}D_0^{(j)-1})_{rs}| > V_1 \log(p)/n) < \epsilon$. This implies that with probability greater than $1 - \epsilon$

$$\begin{aligned}
& \left| \sum_{j=1}^J \text{tr} D_0^{(j)-1} (\Delta_T^{(j)} (S^{(j)} - \Sigma_0^{(j)}) T_0^{(j)} + T_0^{(j)} (S^{(j)} - \Sigma_0^{(j)}) \Delta_T^{(j)'}) \right| \\
& \leq \sum_{j=1}^J |\Delta_T^{(j)}|_1 (\max |((S^{(j)} - \Sigma_0^{(j)}) T_0^{(j)'} D_0^{(j)-1})_{rs}| + \max |(D_0^{(j)-1} T_0^{(j)} (S^{(j)} - \Sigma_0^{(j)}))_{rs}|) \\
& \leq 2V_1 \sqrt{\log(p)/n} \sum_{j=1}^J |\Delta_T^{(j)}|_1 \\
& = 2V_1 \sqrt{\log(p)/n} \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| + 2V_1 \sqrt{\log(p)/n} \sum_{rs \in Z} \sum_{j=1}^J |t_{rs}^{(j)} - t_{0rs}^{(j)}| \\
& \leq 2V_1 \sqrt{\log(p)/n} \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| + 2V_1 \sqrt{\log(p)/n} \sqrt{s} \sqrt{\sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2}.
\end{aligned}$$

Recall that $L_1 > 1/2d^2 \sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2$, we can induce from the above inequality that

$$\begin{aligned}
L_1 - |L_2| & > 1/2d^2 \sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2 - 2V_1 \sqrt{\log(p)/n} \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| \\
& \quad - 2V_1 \sqrt{\log(p)/n} \sqrt{s} \sqrt{\sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2}.
\end{aligned}$$

Next, consider $I_1 + I_2$ and $G_1 + G_2$. It has to be noted that the term I_1 is positive,

$$I_1 = \beta \sum_{rs \in Z^c} \max_{j=1}^J |t_{rs}^{(j)}| + \lambda \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| \geq (\frac{\beta}{J} + \lambda) \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}|.$$

The term G_1 is also positive. $G_1 = \beta \sum_{rs \in Z^c} \sqrt{\sum_{j=1}^J t_{rs}^{(j)2}} + \lambda \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| \geq \beta \sum_{rs \in Z^c} \sqrt{(\sum_{j=1}^J |t_{rs}^{(j)}|)^2 / J} + \lambda \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| \geq (\frac{\beta}{J} + \lambda) \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}|$.

At the other side, the term

$$\begin{aligned}
|I_2| &= |\beta \sum_{rs \in Z} (\max_{j=1}^J |t_{rs}^{(j)}| - \max_{j=1}^J |t_{0rs}^{(j)}|) + \lambda \sum_{rs \in Z} \sum_{j=1}^J (|t_{rs}^{(j)}| - |t_{0rs}^{(j)}|)| \\
&\leq \beta \sum_{rs \in Z} |\max_{j=1}^J |t_{rs}^{(j)}| - \max_{j=1}^J |t_{0rs}^{(j)}|| + \lambda \sum_{rs \in Z} \sum_{j=1}^J ||t_{rs}^{(j)}| - |t_{0rs}^{(j)}|| \\
&\leq \beta \sum_{rs \in Z} \max_{j=1}^J |t_{rs}^{(j)} - t_{0rs}^{(j)}| + \lambda \sum_{rs \in Z} \sum_{j=1}^J |t_{rs}^{(j)} - t_{0rs}^{(j)}| \\
&\leq (\lambda + \beta) \sum_{rs \in Z} \sum_{j=1}^J |t_{rs}^{(j)} - t_{0rs}^{(j)}| \\
&\leq (\lambda + \beta) \sqrt{s} \sqrt{\sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2}.
\end{aligned}$$

The upper bound for term G_2 can be similarly obtained.

$$\begin{aligned}
|G_2| &= |\beta \sum_{rs \in Z} (\sqrt{\sum_{j=1}^J t_{rs}^{(j)2}} - \sqrt{\sum_{j=1}^J t_{0rs}^{(j)2}}) + \lambda \sum_{rs \in Z} \sum_{j=1}^J (|t_{rs}^{(j)}| - |t_{0rs}^{(j)}|)| \\
&\leq \beta \sum_{rs \in Z} \sum_{j=1}^J |t_{rs}^{(j)} - t_{0rs}^{(j)}| \frac{|t_{rs}^{(j)}| + |t_{0rs}^{(j)}|}{\sqrt{\sum_{j=1}^J t_{rs}^{(j)2}} + \sqrt{\sum_{j=1}^J t_{0rs}^{(j)2}}} + \lambda \sum_{rs \in Z} \sum_{j=1}^J ||t_{rs}^{(j)}| - |t_{0rs}^{(j)}|| \\
&\leq \beta \sum_{rs \in Z} \sum_{j=1}^J |t_{rs}^{(j)} - t_{0rs}^{(j)}| + \lambda \sum_{rs \in Z} \sum_{j=1}^J |t_{rs}^{(j)} - t_{0rs}^{(j)}| \\
&\leq (\lambda + \beta) \sum_{rs \in Z} \sum_{j=1}^J |t_{rs}^{(j)} - t_{0rs}^{(j)}| \\
&\leq (\lambda + \beta) \sqrt{s} \sqrt{\sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2}.
\end{aligned}$$

Recall that $M_1 \geq 0$, $L_2 \geq 0$, $I_1 \geq 0$ and $G_1 \geq 0$. Combine all the above 5 terms together, with probability greater than $1 - 2\epsilon$, we have

$$\begin{aligned}
& |G(\Delta_T, \Delta_D)| \\
& \geq M_1 + I_1(G_1) + L_2 - |L_1| - |M_2| - |I_2|(|G_2|) \\
& \geq 1/8d^4 \sum_{j=1}^J \|\Delta_D^{(j)}\|_F^2 + \left(\frac{\beta}{J} + \lambda\right) \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| + 1/2d^2 \sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2 \\
& \quad - V_2 \sqrt{\log(p)/n} \sqrt{pJ \sum_{j=1}^J \|\Delta_D\|_F^2} - (\lambda + \beta) \sqrt{s} \sqrt{\sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2} \\
& \quad - V_1 \sqrt{\log(p)/n} \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| - V_1 \sqrt{\log(p)/n} \sqrt{s} \sqrt{\sum_{j=1}^J \|\Delta_T^{(j)}\|_F^2} \\
& = \frac{U_2^2}{8d^4} p \log(p)/n + \left(\frac{\beta}{J} + \lambda\right) \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| + \frac{U_1^2}{2d^2} s \log(p)/n \\
& \quad - V_2 U_2 \sqrt{J} p \log(p)/n - (\lambda + \beta) s \sqrt{\log(p)/n} \\
& \quad - V_1 \sqrt{\log(p)/n} \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| - V_1 U_1 s \log(p)/n \\
& \geq U_2 p \log(p) \left(\frac{U_2}{8d^4} - V_2 \sqrt{J}\right) + \sum_{rs \in Z^c} \sum_{j=1}^J |t_{rs}^{(j)}| (\beta/J + \lambda - V_1 \sqrt{\log(p)/n}) \\
& \quad + U_1 s \log(p)/n \left(\frac{U_1}{2d^2} - \frac{\lambda + \beta}{\sqrt{\log(p)/n}} - V_1\right).
\end{aligned}$$

Here V_1 and V_2 are only related to n and ϵ . Assume $\lambda + \beta = K(\log(p)/n)$ where $K > JV_1$ and choose $U_2 > 8d^4 V_1 \sqrt{J}$, $U_1 > 2d^2(K + V_1)$, then we have $G(\Delta_T, \Delta_D) > 0$.

So far, we have proved that $G(\Delta_T, \Delta_D) > 0$ with probability $1 - 2\epsilon$ when U_1

and U_2 big enough. This establishes the theorem.

Proof of Theorem 3.2: Assume $\Omega = T'D^{-1}T$ and $\Omega_0 = T'_0D_0^{-1}T_0$ with $\|\Delta_T\|_F^2 = \|T - T_0\|_F^2 = o_p(1)$, $\|\Delta_D\|_F^2 = \|D - D_0\|_F^2 = o_p(1)$. Further assume that $s_p(\Omega_0)$ and $s_1(\Omega_0)$ are bounded. Using Lemma A.2, we have $\|T_0\| = O(1)$ and $\|D_0\| = O(1)$. In this proof, we bound $\|\Omega - \Omega_0\|_F^2$ by a combination of $\|\Delta_T\|_F^2$ and $\|\Delta_D\|_F^2$ as follows,

$$\begin{aligned} \|\Omega - \Omega_0\|_F^2 &= \|T'D^{-1}T - T'_0D_0^{-1}T_0\|_F^2 \\ &= \|(\Delta'_T + T'_0)D^{-1}(\Delta_T + T_0) - T'_0D_0^{-1}T_0\|_F^2 \\ &\leq 4[\|\Delta'_T D^{-1}T_0\|_F^2 + \|T'_0 D^{-1}\Delta_T\|_F^2 \\ &\quad + \|\Delta'_T D^{-1}\Delta_T\|_F^2 + \|T'_0(D^{-1} - D_0^{-1})T_0\|_F^2]. \end{aligned}$$

We bound these four terms separately. By Lemma A.1 we have

$$\|\Delta'_T D^{(j)-1}T_0\|_F^2 \leq \|T_0\|^2 \|\Delta'_T D^{-1}\|_F^2 \leq \|T_0\|^2 \|D^{-1}\|^2 \|\Delta'_T\|_F^2.$$

Because $\|D - D_0\|_F^2 = o_p(1)$ and $\|D_0\| = O(1)$, we have $\|D\| = \|D_0 + D - D_0\| \leq \|D_0\| + \|D - D_0\| \leq \|D_0\| + \|D - D_0\|_F = O_p(1)$. Along with $\|T_0\| = O(1)$, we have $\|T_0\|^2 \|D^{-1}\|^2 \|\Delta'_T\|_F^2 = O_p(\|\Delta'_T\|_F^2)$. Using the same argument, we have $\|T'_0 D^{-1}\Delta_T\|_F^2 = O_p(\|\Delta'_T\|_F^2)$.

For the second term,

$$\begin{aligned}
\|\Delta'_T D^{-1} \Delta_T\|_F^2 &\leq \|\Delta_T\|^2 \|D^{-1}\|^2 \|\Delta_T\|_F^2 \\
&\leq \|\Delta_T\|_F^2 \|D^{-1}\|^2 \|\Delta_T\|_F^2 \\
&= o_p(\|\Delta_T\|_F^2).
\end{aligned}$$

As to the third term,

$$\begin{aligned}
\|T'_0(D^{-1} - D_0^{-1})T_0\|_F^2 &\leq \|T'_0\|^2 \|(D^{-1} - D_0^{-1})T_0\|_F^2 \\
&\leq \|T'_0\|^2 \|T_0\|^2 \|D^{-1} - D_0^{-1}\|_F^2 \\
&= O_p(\|D - D_0\|_F^2).
\end{aligned}$$

By the assumptions of Theorem 3.1, we know the singular values of $\Sigma_0^{(j)}$ are bounded. This induces the property that the corresponding autoregressive matrix $T_0^{(j)}$ and variance matrix $D_0^{(j)}$ satisfy $\|T_0^{(j)}\| = O(1)$ and $\|D_0^{(j)}\| = O(1)$. Recall that $\sum_{j=1}^J \|T^{(j)} - T_0^{(j)}\|_F^2 = O_p(s \log(p)/n)$ and $\sum_{j=1}^J \|D^{(j)} - D_0^{(j)}\|_F^2 = O_p(p \log(p)/n)$, following the above argument, we have

$$\|\Omega^{(j)} - \Omega_0^{(j)}\|_F^2 = O_p(\|\Delta_T^{(j)}\|_F^2) + O_p(\|\Delta_D^{(j)}\|_F^2) = O_p((s+p) \log(p)/n).$$

Consequently, we have

$$\sum_{j=1}^J \|\Omega^{(j)} - \Omega_0^{(j)}\|_F = O_p(\sqrt{(s+p) \log(p)/n}).$$

The same argument also applies to the covariance matrices. Thus, the following

property also holds

$$\sum_{j=1}^J \|\Sigma^{(j)} - \Sigma_0^{(j)}\|_F = O_p(\sqrt{(s+p) \log(p)/n}).$$

This gives Theorem 3.2.

Prove of Theorem 3.3: For parameters $\phi_{kl}^{(j)}$, where $(k, l) \in Z_j^c$ and $K > l$, we want to prove $\phi_{kl}^{(j)} = 0$. There are two cases, in the first case $(k, l) \notin \cap_{j=1}^J Z_j^c$, which means not all the parameters $\phi_{0kl}^{(1)}, \phi_{0kl}^{(2)}, \dots, \phi_{0kl}^{(J)}$ are zero. Assume $0 = |\phi_{0kl}^{(1)}| = |\phi_{0kl}^{(2)}| = \dots \leq |\phi_{0kl}^{(J)}|$ and $|\phi_{0kl}^{(j+1)}|$ is the first element that not equals 0. We consider a small space that contains $(\phi_{0kl}^{(1)}, \phi_{0kl}^{(2)} \dots \phi_{0kl}^{(J)})$ and suppose $(\phi_{kl}^{(1)}, \phi_{kl}^{(2)}, \dots \phi_{kl}^{(J)})$ is in this space which satisfies $|\phi_{kl}^{(1)}| \leq |\phi_{kl}^{(2)}| \leq \dots \leq |\phi_{kl}^{(J)}|$. Taking the derivative of the objective function with respect to $\phi_{kl}^{(j)}$ at 0, we have

$$\frac{\partial Q}{\partial \phi_{kl}^{(j)}} = \sum_{j=1}^J 2(S^{(j)} T^{(j)'} D^{(j)-1})_{lk} + \lambda \text{sign}(\phi_{kl}^{(j)}).$$

The term $(S^{(j)} T^{(j)'} D^{(j)-1})_{lk}$ can be divided into 4 terms K_1, K_2, K_3, K_4 , where

$$\begin{aligned} K_1 &= \sum_{j=1}^J ((S^{(j)} - \Sigma_0^{(j)}) T^{(j)'} D^{(j)-1})_{lk}, \\ K_2 &= \sum_{j=1}^J (\Sigma_0^{(j)} (T^{(j)'} - T_0^{(j)'}) D^{(j)-1})_{lk}, \\ K_3 &= \sum_{j=1}^J (\Sigma_0^{(j)} T_0^{(j)'} (D^{(j)} - D_0^{(j)-1}))_{lk}, \\ K_4 &= \sum_{j=1}^J (\Sigma_0^{(j)} T_0^{(j)'} D_0^{(j)-1})_{lk}. \end{aligned}$$

For the term K_4 , $K_4 = (T_0^{(j)-1})_{lk}$, we know that $T_0^{(j)-1}$ is a lower triangle matrix. Therefore its lk th element equals 0. So we only need to consider the rest 3 terms.

As we have already proved, $|T^{(j)'}| = O_p(1)$ and $|D^{(j)-1}| = O_p(1)$. By Lemma A.3, we know term K_1 have order $\max_{rs} |((S^{(j)} - \Sigma_0^{(j)})T^{(j)'})_{rs}| = O_p(\sqrt{\log(p)/n})$.

It can be concluded from lemma A.1 that $|K_2| = |\sum_{j=1}^J (\Sigma_0^{(j)} (T^{(j)'}) - T_0^{(j)'}) D^{(j)-1}|_{lk} \leq \sum_{j=1}^J \|\Sigma_0^{(j)} T^{(j)' - T_0^{(j)'}} D^{(j)-1}\| \leq \sum_{j=1}^J \|\Sigma_0^{(j)}\| \|T^{(j)' - T_0^{(j)'}}\| \|D^{(j)-1}\|$. Since $\|\Sigma_0^{(j)}\| = O_p(1)$ and $\|D^{(j)-1}\| = O_p(1)$, we have $|K_2| \leq O_p(\sum_{j=1}^J \|T^{(j)' - T_0^{(j)'}}\|)$.

Following the same procedure, we can prove that the term $K_3 \leq O_p(\sum_{j=1}^J \|D^{(j)} - D_0^{(j)-1}\|)$. According to our assumption that $\sum_{j=1}^J \|T^{(j)' - T_0^{(j)'}}\| = O_p(\zeta_n)$ and $\sum_{j=1}^J \|D^{(j)' - D_0^{(j)'}}\| = O_p(\eta_n)$, we know the term $\sum_{j=1}^J 2(S^{(j)} T^{(j)'})_{lk}$ has rate $\log(p)/n + \zeta_n + \eta_n$. According to our assumption, $\log(p)/n + \zeta_n + \eta_n = O_p(\lambda)$, so $\sum_{j=1}^J 2(S^{(j)} T^{(j)'})_{lk}$ is dominated by λ . Thus the sign of the derivative is the same as the sign of parameter $\phi_{kl}^{(j)}$, which will lead to the conclusion that $\phi_{kl}^{(j)} = 0$. Due to the assumption $|\phi_{kl}^{(1)}| \leq |\phi_{kl}^{(2)}| \leq \dots \leq |\phi_{kl}^{(J)}|$, we have $\phi_{kl}^{(1)} = \phi_{kl}^{(2)} = \dots = \phi_{kl}^{(j)} = 0$ and $|\phi_{kl}^{(j+1)}| = |\phi_{kl}^{(j+2)}| = \dots = |\phi_{kl}^{(J)}| > 0$ with probability tending to 1.

The second case is $(k, l) \in \cap_{j=1}^J Z_j^c$ where $0 = \phi_{0kl}^{(1)} = \phi_{0kl}^{(2)} = \dots = \phi_{0kl}^{(J)}$. Similarly, assume $\{\phi_{kl}^{(1)}, \phi_{kl}^{(2)}, \dots, \phi_{0kl}^{(J)}\}$ falls in a small space containing the original point of R^J space. Without losing any generality, we assume these J

parameters have an ascending order. Taking the derivative of Q with respect to $\phi_{kl}^{(j)}$, we have

$$\frac{\partial Q}{\partial \phi_{kl}^{(j)}} = \sum_{j=1}^J 2(S^{(j)}T^{(j')}D^{(j-1)})_{lk} + (\beta + \lambda)\text{sign}(\phi_{kl}^{(j)}).$$

As we have already proved, $\sum_{j=1}^J 2(S^{(j)}T^{(j')}D^{(j-1)})_{lk}$ will be dominated by β , certainly it will be dominated by $\beta + \lambda$, which tells us that Q will achieve its minimum when $\phi_{kl}^{(j)}$ equals 0. So, with probability tending to 1, $\phi_{kl}^{(1)} = \phi_{kl}^{(2)} = \dots = \phi_{kl}^{(j)} = 0$.

Bibliography

- [1] ANDERSON, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. Wiley Series in Probability and Statistics. Wiley, New York.
- [2] ANTONIADIS, A. (1997). Wavelets in statistics: a review (with discussion). *Italian Jour. Stat.* **6**, 97-144.
- [3] BAI, Z.D. and SILVERSTEIN, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer Series in Statistics. Springer, New York.
- [4] BAI, Z.D. and YIN, Y,Q. (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab.* **21**, 1275C1294.
- [5] BICKEL, P.J. and LEVINA, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Stat.* **36**, 199-227.
- [6] BICKEL, P.J. and LEVINA, E. (2008b). Covariance regularization by thresholding. *Ann. Stat.* **48**, 2577-2604.

-
- [7] BONDELL, H.D. and REICH, B.J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR *Biometrics*. **64**, 115-123.
- [8] BONDELL, H.D. and REICH, B.J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*. **65**, 169-177.
- [9] CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Stat. Assoc.* **106**, 672-684.
- [10] CAI, T., LIU, W. and LUO, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.* **106** 594-607.
- [11] DEY, D.K. and SRINIVASAN, C. (1986). Estimation of a covariance matrix under Stein's loss. *Ann. Stat.* **13**, 1581-1591.
- [12] DANAHER, P., WANG P. and WITTEN D. (2012) The joint graphical lasso for inverse covariance estimation across multiple classes. Available at <http://arxiv.org/pdf/1111.0324.pdf>
- [13] D'ASPREMONT, A., BANERJEE, O. and EL GHAOU, L. (2008). First-Order methods for sparse covariance selection. *Siam. J. Matrix Anal. & Appl.* **30**, 56-66.
- [14] EFRON, B., HASTIE, T., JOHNSTONE, T. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Stat.* **32**, 407-499.
- [15] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **48**, 2717-2756.
- [16] FAN, J., LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348-1360.
- [17] FRANK, I.E., FRIEDMAN, J.H. (1993). A statistical view of some chemometrics regression tool. *Technometrics*. **35**, 109-148
- [18] FRIEDMAN, J., HASTIE, T. and HOFLING, H. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1**, 302-332.

-
- [19] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics* **9**, 432-441.
- [20] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). A note on the group lasso and a sparse group LASSO. Available at <http://www-stat.stanford.edu/tibs/ftp/sparse-grlasso.pdf>
- [21] GUO, J., LEVINA, E., MICHAELIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1-15.
- [22] HUANG, J., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalized normal likelihood. *Biometrika* **93**, 85-98.
- [23] JOHNSTONE, I.M. and LU, A.Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* **104**, 682-693.
- [24] KASS, R. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57**, 1173-1184.
- [25] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.* **37**, 4254-4278.
- [26] LEDOIT, O. and WOLF, M. (2003a). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance.* **10**, 603-621
- [27] LEDOIT, O. and WOLF, M. (2003b). Honey, I shrunk the sample covariance matrix. *J. Portfolio Management.* **30**, 110-119
- [28] LEDOIT, O. and WOLF, M. (2004a). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multiv. Anal.* **88**, 365-411.
- [29] LEE, W., DU., Y., SUN, W., HAYES, D. N. and LIU, Y. (2012a) Multiple response regression for Gaussian mixture models with known labels. *Stat. Anal. Data Mining* to appear.

-
- [30] LEE, W. and LIU, Y. (2012b). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Multiv. Anal.* **111**, 241-255.
- [31] LEVINA, E., ROTHMAN, A. and ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Stat.* **2**, 245-263.
- [32] LIU, H., PALATUCCI, M. and ZHANG, J. (2009) Blockwise coordinate descent procedures for the multi-task Lasso, with applications to neural semantic basis discovery. *International Conference on Machine Learning, June, 2009*.
- [33] LENG, C., WANG, W. and PAN, J. (2010) Semiparametric mean-covariance regression analysis for longitudinal data. *J. Am. Stat. Assoc.* **105**, 181-193.
- [34] MEINSHAUSEN, N. and BUHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* **34**, 1436-1462.
- [35] PAN, J. and MACKENZIE, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika.* **90**, 239-244.
- [36] PEARL, J. (2000). Causality: models, reasoning and inference. Cambridge University Press, Cambridge.
- [37] PETRY, S., FLEXEDER, C. and TUTZ, G. (2011). Pairwise fused LASSO. Available at http://epub.ub.uni-muenchen.de/12164/1/petry_etal_TR102_2011.pdf
- [38] POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameters. *Biometrika.* **86**, 677-690.
- [39] POURAHMADI, M. (2000). Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix. *Biometrika.* **87**, 625-635.
- [40] ROTHMAN, A.J., BICKEL, P.J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2**, 494-515.
- [41] ROTHMAN, A.J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Am. Stat. Assoc.* **104**, 177-186.

-
- [42] ROTHMAN, A.J., LEVINA, E. and ZHU, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*. **97**, 539-550.
- [43] SHOJAIE, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high dimensional directed acyclic graphs. *Biometrika* **97**, 519-538.
- [44] SMITH, M. and KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Am. Stat. Assoc.* **97**, 1141-1153.
- [45] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Stat. Soc. B.* **58**, 267-288.
- [46] TIBSHIRANI, R., ROSSET, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. B.* **67**, 91-108.
- [47] WU, W. and POURAHMADI, M. (2003) Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*. **90**, 831-844.
- [48] WU, T.T. and LANGE, K. (2008) Coordinate descent algorithms for Lasso penalized regression. *Ann. Appl. Stat.* **2**, 224-244.
- [49] YIN, Y.Q. ,BAI, Z.D. and KRISHNAIAH, P.R. (1988) On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probab. Theory Related Fields.* **78**, 509-521.
- [50] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*. **94**, 19-35.
- [51] ZHANG, H., LIU, Y., WU, Y. and ZHU, J. (2008) Variable selection for the multcategory SVM via adaptive sup-norm regularization. *Electron. J. Stat.* **2**, 149-167.
- [52] ZHAO, P., Rocha, G. and Yu, B. (2009)The composite absolute penalties family for grouped and hierarchical variable selection . *Ann. Stat.* **37**, 3468-3497.
- [53] ZOU, H. (2006) The adaptive lasso and its oracle properties . *J. Am. Stat. Assoc.* **101**, 1418-1429.

-
- [54] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B.* **67**, 301-320.
- [55] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.* **36**, 1108-1126.
- [56] ZHOU, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *Stat. Interface.* **3**, 557-574.