Large Dimensional Covariance Matrix Estimation with Decomposition-based Regularization

By

Hao Zheng

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Statistics)

at the
University of Wisconsin–Madison
2014

Date of final oral examination: 1/29/2014

The dissertation is approved by the following members of the Final Oral Committee:

Kam-Wah Tsui, Professor, Statistics
Paul J. Rathouz, Professor, Biostatistics and Medical Informatics
Sijian Wang, Assistant Professor, Biostatistics and Medical Informatics
Zhengjun Zhang, Associate Professor, Statistics
Jun Zhu, Professor, Statistics and Entomology

Abstract

Estimation of population covariance matrices from samples of multivariate data is of great importance. When the dimension of a covariance matrix is large but the sample size is limited, it is well known that the sample covariance matrix is dissatisfactory. However, the improvement of covariance matrix estimation is not straightforward, mainly because of the constraint of positive definiteness. This thesis work considers decomposition-based methods to circumvent this primary difficulty. Two ways of covariance matrix estimation with regularization on factor matrices from decompositions are included. One approach replies on the modified Cholesky decomposition from Pourahmadi, and the other technique, matrix exponential or matrix logarithm, is closely related to the spectral decomposition. We explore the usage of covariance matrix estimation by imposing $\mathcal{L}_1$ regularization on the entries of Cholesky factor matrices, and find the estimates from this approach are not sensitive to the orders of variables. A given order of variables is the prerequisite in the application of the modified Cholesky decomposition, while in practice, information on the order of variables is often unknown. We take advantage of this property to remove the requirement of order information, and propose an order-invariant covariance matrix estimate by refining estimates corresponding to different orders of variables. The refinement not only guarantees the positive definiteness of the estimated covariance matrix, but also is applicable in general situations without the order of variables being prespecified. The refined estimate can be approximated by only combining a moderate number of representative estimates. Numerical simulations are conducted to evaluate the performance of the proposed method in comparison with several other estimates. By applying the matrix exponential technique, the problem of estimating positive definite covariance matrices is transferred into a problem of estimating symmetric matrices. There are close connections between covariance matrices and their logarithm matrices, and thus, pursing a matrix logarithm with certain properties helps restoring the original covariance matrix. The covariance matrix estimate from applying $\mathcal{L}_1$ regularization to the entries of the matrix logarithm is compared to some other estimates in simulation studies and real data analysis.

Acknowledgements

I would like to express my sincere gratitude to my advisor Prof. Kam-Wah Tsui, for his continuous guidance and support of my Ph.D. study and research. It is due to his supervision that I can produce this work. His knowledge, patience, encouragement and inspiration accompany me throughout my time in Madison.

I would like to express my great appreciation to Prof. Xinwei Deng of Virginia Tech. Through beneficial discussion with him, I understand more about this area. His responses always come promptly to my questions and queries.

I am deeply grateful to Prof. Paul Rathouz for whom I worked as a research assistant in the past two years. Those computational techniques and writing skills I learn from him benefit me much in my thesis work, and would continue to benefit me in the long run.

I would also like to thank other members in my thesis committee: Prof. Sijian Wang, Prof. Zhengjun Zhang and Prof. Jun Zhu, for their kindness and insightful comments.

Contents

List of Figures

List of Tables

1

Introduction

Covariance matrices play an essential role in multivariate statistics, including principal components analysis, factor analysis, canonical correlation analysis, discriminant analysis and so on. Thus, estimation of covariance matrices from samples of multivariate data is of great importance. The sample covariance matrix, becomes less satisfactory in covariance matrix estimation when the number of variables increases. In many applications involving gene expression, spectroscopic imaging, functional magnetic resonance imaging, weather forecasting and others, the variable sizes largely override the sample sizes. In such a circumstance, the sample covariance matrix becomes degenerate with a distorted eigen-structure (Geman, 1980; Silverstein, 1985; Johnstone, 2001; Pourahmadi, 2011). Therefore, it is important to explore more appropriate covariance matrix estimation in large dimensions.

1.1 Empirical Covariance Matrix Estimation

In statistics, covariance is a measure of how two random variables change together.

The covariance between two random variables X and Y is defined as

$$\mathrm{cov}(X, Y) = \mathrm{E}\big(X - \mathrm{E}(X))(Y - \mathrm{E}(Y)\big),$$

where $\mathrm{E}(X)$ is the expected value or expectation of X. The sign of $\mathrm{cov}(X, Y)$ shows the tendency of linear relationship between X and Y. A positive $\mathrm{cov}(X, Y)$ implies similar behaviors of X and Y. More specifically, when one variable takes greater values, the other also takes greater values; when one takes smaller values, the other also takes smaller values. A negative $\mathrm{cov}(X, Y)$ implies opposite behaviors of X and Y. When one variable takes greater values, the other takes smaller ones. The magnitude of $\mathrm{cov}(X, Y)$ measures the extent of the linear relation with the consideration of individual variances of X and Y. If X and Y are identical, $\mathrm{cov}(X, Y)$ becomes the variance of X or Y.

In practice, $\mathrm{cov}(X, Y)$ is often unknown, and needs to be estimated. With the observations for X and Y available, say $(x_1, \ldots, x_n)^T$ and $(y_1, \ldots, y_n)^T$, the sample covariance, denoted as $\hat{\sigma}(X, Y)$ here, serves as an estimate for $\mathrm{cov}(X, Y)$. It is calculated as follows:

$$\hat{\sigma}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \tag{1.1}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ are sample means that are used to estimate corresponding expected values $\mathrm{E}(X)$ and $\mathrm{E}(Y)$. In a special case with both $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ known, $\hat{\sigma}(X, Y)$ is calculated using

$$\hat{\sigma}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mathrm{E}(X))(y_i - \mathrm{E}(Y)). \tag{1.2}$$

With the assumption of independence and identical distribution (*i.i.d.*) among observation points, $\hat{\sigma}(X, Y)$ is an unbiased estimate for $\mathrm{cov}(X, Y)$.

A covariance matrix $\Sigma = (\sigma_{st})$ is a square matrix that presents pairwise covariances for a group of random variables $\{X_1, \ldots, X_p\}$. This group of random variables compose a p-length random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, and the (s,t) position of Σ is the covariance between the s-th and t-th elements of $\boldsymbol{X}$. That is, $\sigma_{st} = \mathrm{cov}(X_s, X_t)$, or equivalently,

$$\Sigma = \mathrm{cov}(\boldsymbol{X}) = \mathrm{E}\big(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X}))(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})^T\big).$$

Σ is symmetric and positive semi-definite. The diagonal entries of Σ give the variances of individual random variables.

Since individual sample covariances can be used to estimate individual covariances, a matrix which consists of sample covariances could be used to estimate the covariance matrix. This provides an empirical covariance matrix estimation, and the estimate is the sample covariance matrix. Let $\hat{\sigma}_{st} = \hat{\sigma}(X_s, X_t)$ from (1.1) or (1.2), and the sample covariance matrix is $S = (\hat{\sigma}_{st})$. Without loss of generality, the assumption of $\mathrm{E}(\boldsymbol{X}) = 0$ holds throughout this thesis. Thus,

$$\Sigma = \mathrm{E}\ \boldsymbol{X}\boldsymbol{X}^T, \quad \text{and} \quad S = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i}\boldsymbol{x_i}^T,$$

where $\boldsymbol{x_i}$ represents the i-th observation for the random vector $\boldsymbol{X}$.

The properties of the sample covariance matrix S have been extensively studied. With *i.i.d.* observations, S is an unbiased estimate for Σ. If the sample is from a multivariate normal distribution with zero mean, the resulting S is the maximum likelihood estimate for Σ, and nS follows the (central) Wishart distribution (Wishart, 1928). Constantine (1963) rewrites the distribution of S in terms of a hypergeometric function of matrix argument.

As a special case of random matrices, the sample covariance matrix S is also widely studied from the perspective of spectral analysis, i.e., the distribution of its eigenvalues $\{\lambda_1, \ldots, \lambda_p\}$. The empirical distribution of eigenvalues is defined as follows:

$$F(t; n, p) = \frac{1}{p} \sum_{i=1}^{p} \mathbf{1}_{\{\lambda_i \le t\}}.$$

Marčenko and Pastur (1967) achieve a great advancement about the limiting distribution of $F(t; n, p)$ under a simple but important circumstance. Their result states if both the sample size n and the variable size p proportionally grow to infinity such that $\lim \frac{p}{n} = c$ for some positive $c > 0$ and all the variables X_k's are *i.i.d.* with mean zero and variance σ^2, then $F(t; n, p)$ converges to the so-called Marčenko-Pastur distribution with the density function

$$g(t) = \frac{1}{2\pi\sigma^2 ct}\sqrt{(b-t)(t-a)}, \quad a \le t \le b, \tag{1.3}$$

with $a = \sigma^2(1-\sqrt{c})^2$ and $b = \sigma^2(1+\sqrt{c})^2$. The distribution has a point mass $1 - \frac{1}{c}$ if $c > 1$. This fundamental result describes how the eigenvalues of the empirical sample covariance matrix deviate from the expected ones under in a common situation. Wachter (1978), Jonsson (1982) and Yin (1986) further explore the asymptotic properties of the empirical distribution of eigenvalues in similar conditions. Assuming a matrix with extra independent variables, Silverstein (1995) achieves a convergence in distribution about the empirical distribution of eigenvalues for the product matrix. Through an equation defining its Stieltjes transform, Silverstein and Choi (1995) discover certain analytical properties of the limiting distribution for $F(t;n,p)$ such as the limiting distribution has a continuous derivative away from zero.

Another branch of research is to study the largest eigenvalue $\lambda_{\max}$ and the smallest eigenvalue $\lambda_{\min}$ of the sample covariance matrix. Through developing an almost sure limit for the operator norm of a class of rectangular random matrices, Geman (1980) achieves the limiting distribution of the largest eigenvalue $\lambda_{\max}$ of the sample covariance matrix S with the *i.i.d* observation setting for all X_k's, $k = 1, \ldots, p$. Under the condition

$$E|X_1|^n \leq n^{\alpha n} \quad \text{for all } n \geq 3 \text{ and some } \alpha > 0,$$

the statement deduced from Geman's work gives

$$\lambda_{\max} \rightarrow \sigma^2(1+\sqrt{c})^2, \quad a.s., \tag{1.4}$$

where c is the asymptotic ratio of p and n, consistent with the previous definition. Jonsson (1982) obtains the same convergence under $E|X_1|^7 < \infty$, and Silverstein (1984) further weakens the condition to $E|X_1|^{6+\epsilon} < \infty$ for any small ϵ. Yin et al. (1988) prove that (1.4) still holds under $E|X_1|^4 < \infty$, and Bai et al. (1988) complete this procedure by showing that the fourth moment can not be further weakened.

Starting with normal distribution and $0 < c < 1$, Silverstein (1985) shows that if $X_k \sim \mathcal{N}(0, \sigma^2)$, $1 \leq k \leq p$, are independent, the smallest eigenvalue $\lambda_{\min}$ of the sample covariance matrix S has a similar asymptotic property:

$$\lambda_{\min} \rightarrow \sigma^2(1-\sqrt{c})^2, \quad a.s.. \tag{1.5}$$

Bai and Yin (1993) prove that (1.5) still holds without the normality assumption, as long as the fourth moment of variables are finite.

(1.4) and (1.5) state that when the true covariance matrix is an identity matrix and the original variables have finite fourth moment, $\lambda_{\max}$ and $\lambda_{\min}$ in the corresponding sample covariance matrix converges almost surely to the respective boundaries of the support of the Marčenko-Pastur distribution.

The further achievement about $\lambda_{\max}$ comes from Johnstone (2001) based on the development of Tracy-Widom distribution. By studying the limiting law of the largest eigenvalue of a square Gaussian symmetric matrix, Tracy and Widom (1994, 1996) develop the so-called Tracy-Widom distribution. The cumulative distribution has a form of

$$F(y) = \exp\Big(- \int_y^\infty (x-y)\, q^2(x)\, dx \Big),$$

where function $q(\cdot)$ solves the nonlinear Painlevé II differential equation (Painleve, 1902),

$$\frac{d^2}{dx^2}\, q(x) = x\, q(x) + 2\, q^3(x).$$

With assumption of independent $X_k \sim \mathcal{N}(0,1)$, $1 \leq k \leq p$, Johnstone presents the variability information for the convergence of $\lambda_{\max}$, and the asymptotic distribution of the re-scaled largest eigenvalue of the sample covariance matrix follows Tracy-Widom distribution, a more delicate result compared with (1.4).

1.2 Challenges from High Dimensional Data

Covariance matrices play an important role in statistical inference. In recent years, areas of statistical learning dealing with massive and high-dimensional data have been growing rapidly. In such areas, while the number of features or variables are dramatically high, the number of sample points are rather limited. One of the most influential areas among them deals with gene expression data. Gene expression is the most fundamental level where the genotype of an organism gives rise to the phenotype, and usually is achieved through genome-wide approaches such as microarrays and RNA sequencing. Not surprisingly, the genome-wide approaches accompany a large amount of features or variables. For instance, in a normal DNA microarray, every gene is supposed to have a expression value, which means over 30,000 variables get involved. Other types of gene expression data, such as data sets from exon array, tiling array, single nucleotide polymorphism (SNP) array, have even larger scale of variable dimensions. Functional magnetic resonance imaging (fMRI) is another area flooded with high-dimensional data. By measuring brain activity through detecting associated changes in blood flow, fMRI is a neuro-imaging technology that does not require medicine injection or radiation exposure, and thus, has dominated the medical diagnosis of brain imaging. When each brain image supplies huge amount of variable values, the number of images is quite small. In spectroscopic imaging, weather forecasting and so many other areas, there are more and more high-dimensional data emerging.

Estimation of the covariance matrix in such areas encounters a new challenge that the samples sizes are much smaller compared with the number of variables where the number of parameters grows quadratically in terms of the number of variables. As summarized by Pourahmadi (2011), the sample covariance matrix, based on a sample of size n from a mean zero normal population with the covariance matrix Σ, is not a satisfactory estimator when the ratio $\dfrac{p}{n}$ goes large. Applying (1.4), one can see $\lambda_{\max}$ goes in the same order of the ratio $\frac{p}{n}$ even when the true $\lambda_{\max}$ should be constant. In the studies involving high-dimensional data, the sample covariance matrix systematically distorts the eigen-structure of Σ, making the the largest sample eigenvalues upward biased and the smallest sample eigenvalues downward biased. This phenomenon explains the motivation of efforts for developing more appropriate covariance matrix estimation in high-dimensional statistical studies.

1.3 Development of Covariance Matrix Estimation

While the distorted eigen-structure of the sample covariance matrix S become a common issue, especially when the sample size is less than the number of variables, the development of covariance matrix estimation never ceases to advance. Here we roughly introduce four groups of methods: (1) Stein-type shrinkage estimation; (2) estimation from regularizing the sample covariance matrix; (3) regularized covariance matrix estimation; (4) decomposition-based estimation with regularization.

1.3.1 Stein-type Shrinkage Estimation

An early achievement of covariance matrix estimation should be attributed, at least partially, to the practice of ridge regression (Hoerl and Kennard, 1970). Somehow, the resultant covariance matrix estimate is Stein-type, though there is no shrinkage. One may argue that ridge regression is developed in a totally different domain, and the motivation is to improve ordinary least squares in fitting linear regressions. However, in its implementation of imposing $\mathcal{L}_2$ penalty on the regression coefficients, the treatment is equivalent to adding small positive quantities to the diagonal entries of

$\sum_{i=1}^{n} \boldsymbol{x_i}\boldsymbol{x_i}^T$. If the data are centered first to remove the intercept in the regression, we have $\sum_{i=1}^{n} \boldsymbol{x_i}\boldsymbol{x_i}^T = nS$, so this treatment supplies a covariance matrix estimate in the form of

$$\hat{\Sigma} = S + \eta\, I_p\,, \tag{1.6}$$

where η corresponds to the added quantities, and I_p is the p-dimensional identity matrix. This strategy is straightforward, and practically useful especially when the original S is singular. Although the development of ridge regression does not have a focus on improving covariance matrix estimation, it is a pioneer in providing standard procedures to improve S. Through adding small quantities to the diagonal entries of S, the underestimated small eigenvalues of S are increased. On the other hand, because adding quantities to the diagonal entries further enlarge the eigenvalues, the distortion of largest eigenvalues of S gets worse.

An important event in the development of covariance matrix estimation is Stein's Rietz lecture (Stein, 1975), where he brings up the issues of improving the sample covariance matrix and provides many useful recommendations. Since then, many people explore the problem within the framework of Stein's estimation by shrinking S. Among them, one of the most ambitious estimate is from Haff (1980), which has the form of

$$\hat{\Sigma} = a\, S + u \cdot t(u)\, \mathbb{C}. \tag{1.7}$$

a is a constant number; $\mathbb{C}$ could be any arbitrary positive definite matrix; u is determined by S and $\mathbb{C}$, and $t(\cdot)$ is any non-increasing function. While the motivation aims high, the setting involving the arbitrary $\mathbb{C}$ and $t(\cdot)$ increases its complexity and decreases its feasibility. Important results are achieved with $t(\cdot) \equiv 0$, which reduces the form into $a\,S$. a is determined by the choice of loss functions. Denote $\hat{\Sigma}$ for an estimate of Σ. When the data are *i.i.d* from normal distribution with mean 0 and covariance matrix Σ, Haff shows that if the loss function is entropy loss which is defined as

$$\text{entropy loss} = \text{tr}(\Sigma^{-1}\hat{\Sigma}) - \log|\Sigma^{-1}\hat{\Sigma}| - p, \tag{1.8}$$

the resultant a is equal to 1. A general choice of the loss function does not correspond to a close-form a. Through numerical investigations, Lin and Perlman (1985) and (Haff, 1991) suggest that Haff's estimate adequately shrinks the eigen-values of S, even though the theoretical result about its accuracy is quite limited. They also notice that the eigenvalues of such an estimate could be over-shrank. Meanwhile, because the distortion of eigenvalues of S is not in the same direction, shrinking the eigenvalues through $\hat{\Sigma} = a\,S$ with $a < 1$ further underestimates the small eigenvalues.

Based on (1.6) and (1.7), it seems natural to investigate potential estimate for the covariance matrix in the form of

$$a\,S + \eta\,I_p\,, \tag{1.9}$$

especially when this form carries the possibility of simultaneously decreasing the large eigenvalues and increasing the small eigenvalues of S. Such a form is used by Efron and Morris (1976) to estimate an inverse covariance matrix. A delicate covariance matrix estimate in this form is used in discriminant analysis by Friedman (1989), but there is no optimality analysis for the choice of a and η. It is Ledoit and Wolf (2004) that first systematically investigate the covariance matrix estimate in the form of

(1.9). The very important strategy of Ledoit and Wolf is that they adopt the usage of Frobenius norm, which could be used to define an inner product. Projection in the defined inner product space brings their estimate. Equivalently, they rewrite the form (1.9) using

$$\hat{\Sigma}_{LW} = (1-\rho)\, S + \rho\, \nu\, I_p, \tag{1.10}$$

and estimate the parameters through

$$\min_{\rho,\nu} \mathrm{E}\left\{||\hat{\Sigma}_{LW} - \Sigma||_F^2\right\}.$$

$\|\cdot\|_F$ represents Frobenius norm (F norm). The F norm of a matrix $A = (a_{st})$ is denoted by $||A||_F$, which is defined as follows:

$$||A||_F = \sqrt{\sum_s \sum_t a_{st}^2}\,. \tag{1.11}$$

The most convenient part is that Ledoit and Wolf present closed-form estimates for ρ and ν in simple formula, which greatly speeds up the acceptance of this approach. The eigenvalues of $\hat{\Sigma}_{LW}$ are weighted averages of the ones from the sample covariance matrix and the identity matrix. In practice, the largest eigenvalues of $\hat{\Sigma}_{LW}$ are better controlled, while the smallest eigenvalues of the estimate are often exaggerated.

1.3.2 Estimation from Regularizing Sample Covariance Matrix

Different from shrinking the sample covariance matrix towards some pre-specified positive definite matrix, a group of methods achieve covariance matrix estimates by regularizing the entries of the sample covariance matrix. Bickel and Levina (2008b) consider thresholding small entries of the sample covariance matrix to zeros. Dealing with covariance matrices with banded structures, Bickel and Levina (2008a) consider banding the sample covariance matrix through only keeping entries in the diagonal and certain sub-diagonals non-zeros. Where these methods are easy to implement and have nice asymptotic properties, one drawback is that they can not guarantee the positive definiteness of the estimated covariance matrix. In statistical inference, positive definiteness is a desirable property for a covariance matrix estimate. Many applications including evaluating the likelihood of multivariate normal data and measuring the variance proportion in applying principal components analysis (PCA) require positive definite covariance matrices.

1.3.3 Regularized Covariance Matrix Estimation

To pursue improved covariance matrix estimate with guaranteed positive definiteness, one perspective is to apply regularization on the covariance entries while treating them as parameters. Such a strategy usually requires sophisticated optimization techniques in order to meet the positive definiteness. Bien and Tibshirani (2011) propose an estimate through optimizing the $\mathcal{L}_1$ penalized log-likelihood using a majorization-minimization (Hunter and Lange, 2000) technique. An alternative direction algorithm is used by Liu et al. (2013) when they add an eigenvalue constraint to the employment of thresholding methods. Such sophisticated optimization often involves intensive computation and convergence issue.

1.3.4 Decomposition-based Estimation with Regularization

Another perspective of improving covariance matrix estimates with guaranteed positive definiteness is not to directly regularizing the covariance entries. Rather, through appropriate matrix decomposition, the regularization could be placed on the entries of the factor matrices instead of on the original covariance entries. Therefore, the constraint of positive definiteness is circumvented. We follow the direction in this thesis. More specifically, two ways of reparameterization for the covariance matrix are considered. One is associated with the modified Cholesky decomposition, and the other one, using matrix exponential, or equivalently, using matrix logarithm, is associated with the spectral decomposition.

The modified Cholesky decomposition from Pourahmadi (1999) is a widely-used tool in dealing with covariance matrices. The sequential regressions in accordance with the modified Cholesky decomposition provide an unconstrained reparameterization of the covariance matrix, and regularization can be easily applied to the Cholesky factor matrix for it is equivalent to regularizing the coefficients of the linear regressions. Incorporating the advantages of Bickel and Levina's banding idea, Rothman et al. (2010) proposed banding the Cholesky factor matrix of the covariance matrix so that the estimated covariance matrix is always positive definite. The covariance matrix estimation through regularizing the Cholesky factor matrix is not necessarily limited to the scenarios in which the covariance matrices are banded. We employ the $\mathcal{L}_1$ regularization on the Cholesky factor matrix to estimate the covariance matrix in a more general situation where particular assumption of the matrix structure is not necessary. Besides that, unlike the approach of banding the Cholesky factor matrix using ordinary least squares, the covariance matrix estimation through employing $\mathcal{L}_1$ regularization on the Cholesky factor matrix does not suffer the constraint of insufficient sample points. More importantly, we find that the estimate from $\mathcal{L}_1$ regularization is not sensitive towards the order of variables. One prerequisite condition of using the modified Cholesky decomposition is the order information of variables. Often, such information is not available, or can not be reasonably assumed. Weakening or even getting rid of this requirement can greatly broaden the usage of this technique. In summary, Cholesky-based estimate from regularizing the Cholesky factor matrix has its own features and advantages.

The idea of applying spectral decomposition to improve the covariance matrix estimation has a straightforward starting point. Since the sample eigenvalues of S tend to spread out compared with the ones of the true covariance matrix, a straightforward way to deal with it is directly working on the eigenvalues of S. Suppose the sample covariance matrix has a spectral decomposition as follows:

$$S = R\,\Lambda\,R^T, \tag{1.12}$$

where R is the matrix of normalized eigenvectors ($RR^T = R^TR = I_p$), Λ is the diagonal matrix of corresponding eigenvalues with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. In Stein's Rietz lecture (Stein, 1975), he proposes estimates in the form of

$$S = R\,\boldsymbol{\varphi}(\Lambda)\,R^T, \tag{1.13}$$

where $\boldsymbol{\varphi}(\Lambda) = \text{diag}\big\{\varphi_1(\lambda_1, \ldots, \lambda_p), \ldots, \varphi_p(\lambda_1, \ldots, \lambda_p)\big\}$. Stein provides a few suggestions for $\boldsymbol{\varphi}(\cdot)$, and there is further discussion along this line (Dey and Srinivasan, 1985). Although the guideline of selecting $\varphi(\cdot)$ is obvious that the large eigenvalues should be shrank while the small ones should be expanded, there is no clear optimal choice for $\varphi(\cdot)$, partly because the domain of the function is so large and behavior of eigenvalues varies so greatly along with the change of covariance matrix structure and the increase of matrix dimension.

Other than proposing an explicit function to control the eigenvalues, similar effect can be achieved differently. A technique called matrix exponential, or its inverse form, matrix logarithm, helps to implement this purpose in an implicit way. Because the sum of logarithms of eigenvalues of original covariance matrix is equal to the squared F norm of its matrix logarithm, Deng and Tsui (2013) propose optimizing a penalized log-likelihood, where the penalty controls the largest and smallest eigenvalues simultaneously. So the unknown part is changed from a function difficult to determine to be a tuning parameter which is achievable with data-driven methods. Besides the eigenvalue connection, there are more direct structure connection. If a covariance matrix has a banded structure, its matrix logarithm has a banded structure; if a covariance matrix has a block diagonal structure, its matrix logarithm also has a block diagonal structure. While the constraint of positive definiteness does not exist for the matrix logarithm, the approach that restores its structure and transfers it back to the covariance matrix, is often easier than directly restoring the structure of the covariance matrix, especially when the restore is conducted through the application of certain regularization.

With the belief that applying regularization on the factor matrices from matrix decompositions rather than directly on the covariance matrix has its own advantages, we work in such a direction. In this thesis, the modified Cholesky decomposition and matrix exponential technique based on spectral decomposition are mainly considered. That's where the title "Large Dimensional Covariance Matrix Estimation with Decomposition-based Regularization" comes from.

2

Two Reparametrizations of Covariance Matrix

While approaches of regularizing covariance matrix entries are expected to produce improved covariance matrix estimation, the obstacle of positive-definiteness constraint from the covariance matrix stands in the way. To circumvent the encumbrance, this thesis work applies regularization on the entries of factor matrices from decomposing the covariance matrix so that the parameters are released from the positive-definiteness restriction. This chapter covers the basics of the two decompositions used later as well as two reparametrizations for the covariance matrix to be estimated.

2.1 Reparametrization using Modified Cholesky Decomposition

In the study of joint mean-covariance modeling, Pourahmadi (1999) provides unconstrained parameterization for a covariance matrix. In the further decomposition of the Cholesky factor matrix from the traditional Cholesky decomposition into a unit lower triangle and a unique diagonal matrix, Pourahmadi shows the entries of the unit lower triangular are unconstrained and have meanings as regression coefficients when regressing a measurement on its predecessors, while the non-zero entries of the diagonal matrix correspond to the prediction standard deviations.

Pourahmadi's work of the regression-based Cholesky decomposition builds a bridge connecting covariance matrix estimation and regression analysis. Not only the decomposition provides an unconstrained and statistically interpretable reparameterization and guarantees the positive-definiteness of the estimated covariance matrix, but also many developed techniques in regression analysis can be applied. The regression-based Cholesky decomposition is referred as the modified Cholesky decomposition throughout this thesis, and is explained in details next.

2.1.1 Modified Cholesky Decomposition

Discovered by French military officer André-Louis Cholesky and named after him, the Cholesky decomposition is a decomposition of a positive definite matrix into the product of a lower triangular matrix and its (conjugate) transpose as follows:

$$\Sigma = CC^T,$$

where $C = (c_{ij})$ is a lower triangular matrix with positive diagonal entries. When Σ is only positive semi-definite, there are many forms of $\Sigma = CC^T$ for C may have different forms. When Σ is positive definite, the Cholesky decomposition is unique, and the lower matrix C is often referred as the Cholesky factor matrix of Σ.

The Cholesky decomposition is extensively applicable in a wide filed. One of the most popular ways of simulating multivariate normal distribution data with a known covariance matrix is to multiplying its Cholesky factor matrix with the independently distributed normal data, which are directly available in majority statistical softwares. The primary usage of the Cholesky decomposition is to solve a system of linear equations $A\mathbf{x} = \mathbf{b}$, where A is the coefficient matrix and $\mathbf{x}$ is the variable vector. With a lower triangular matrix L or a upper triangular matrix U, a linear system in the form $L\mathbf{x} = \mathbf{b}$ or $U\mathbf{x} = \mathbf{b}$ is straightforward to solve by an iterative process, called forward substitution for the lower triangular matrix case and back substitution for the upper triangular matrix case. Therefore, to solve a general linear system $A\mathbf{x} = \mathbf{b}$, it is usually to decompose the coefficient matrix A first into a product of a lower triangular matrix and a upper triangular matrix, $A = LU$, so called LU decomposition. As a special situation, linear systems with positive definite A arise often in applications, such as in solving linear regressions using ordinary least squares. When A is a positive definite matrix, it is possible that U is the (conjugate) transpose of L, which is the Cholesky decomposition. While computing the Cholesky decomposition is numerically more stable and more efficient than computing some other LU decompositions, and the decomposition always exists and is unique, applying Cholesky decomposition becomes the routine in solving linear systems when the coefficient matrix is positive definite. Because of that, the Cholesky decomposition is widely used in many optimization algorithms. For instance, in Newton-Raphson method, each iteration needs to solve a linear system with an approximate Hessian matrix as the coefficient matrix. While the approximate Hessian matrix is often numerically ill-conditioned (the condition number is high, or the output value from multiplying this matrix changes much for a small change in the input variable vector), application of the Cholesky decomposition of the approximate Hessian matrix is preferred than approximating the inverse of the Hessian matrix.

Pourahmadi discovers another relationship between the Cholesky decomposition and the inverse matrix through changing the original Cholesky factor matrix into a unit lower triangular matrix. Define the diagonal matrix

$$D = \mathrm{diag}(c_{11}, c_{22}, \ldots, c_{pp}),$$

the decomposition of Σ can be rewritten as

$$\Sigma = CC^T = CD^{-1}DDD^{-1}C^T = LD^2L^T, \tag{2.1}$$

where $L = CD^{-1}$ is a unit lower triangular matrix. L can be extracted directly from C, and from now on, we also call L as the Cholesky factor matrix of Σ.

While the modification is trivial and has been used ever since the beginning of the Cholesky decomposition, Pourahmadi points out the statistical interpretation of the modified decomposition.

2.1.2 Representation of Sequential Regressions

The importance of obtaining a unit lower triangular matrix is to connect Σ with linear regression techniques. Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a random vector with mean zero and the positive-definite covariance matrix Σ. The auto-regression predictor of X_j based on previous $\{X_1, \ldots, X_{j-1}\}$ is denoted by $\hat{X}_j$. The residual is $E_j = X_j - \hat{X}_j$ with variance σ_j^2 for $j = 1, \ldots, p$, and the residual vector is $\boldsymbol{E} = (E_1, \ldots, E_p)^T$. For

$j = 1$, we set $\hat{X}_1 = E(X_1) = 0$ and thus $E_1 = X_1$; for $1 < j \leq p$, there are unique coefficients ϕ_{kj} satisfying

$$X_j = \sum_{k<j} \phi_{jk} X_k + E_j, \quad j = 2, \ldots, p. \tag{2.2}$$

Let Φ be the lower matrix with entries $\phi_{jk}, 1 \leq k < j \leq p$ and other entries zeros, and I_p be the $p \times p$ identity matrix. (2.2) can be rewritten as

$$(I_p - \Phi)\boldsymbol{X} = \boldsymbol{E}. \tag{2.3}$$

Based on regression techniques, it can be shown E_j is in the linear space spanned by $\{X_1, \ldots, X_j\}$ but perpendicular to the sub-space spanned by $\{X_1, \ldots, X_{j-1}\}$, and hence, the residuals are uncorrelated. Computing the variance of the both sides of (2.3) gives

$$(I_p - \Phi)\, \Sigma\, (I_p - \Phi)^T = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2), \tag{2.4}$$

and (2.4) gives the expression of the covariance matrix as follows:

$$\Sigma = (I_p - \Phi)^{-1}\, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)\, \{(I_p - \Phi)^{-1}\}^T. \tag{2.5}$$

Note Φ is a lower triangular matrix with diagonal entries zeros, and $(I_p - \Phi)$ is a unit lower triangular matrix. Because the inverse matrix of a unit lower triangular matrix is still a unit lower triangular matrix, the form of (2.5) matches the modified Cholesky decomposition formula (2.1). Correspondingly,

$$L = (I_p - \Phi)^{-1} \quad \text{and} \quad D = \mathrm{diag}(\sigma_1, \ldots, \sigma_p). \tag{2.6}$$

For the parameters of ϕ_{jk}'s and σ_j's are all achievable through linear regressions, the modified Cholesky decomposition of a covariance matrix provides a reparameterization of the covariance matrix by associating the covariance covariates with sequential regression coefficients. Unlike the original covariance entries of Σ, which should function together to make the whole matrix positive definite, ϕ_{jk}'s and σ_j's don't have such constraints. As long as σ_j's are non-zeros, the parameters are free to vary. The release of positive definiteness accompanies great potential in further applications using the representation from sequential regressions.

The representation has been connected to the computation of likelihood for multivariate normal data. Assume the sample size is n. Then, each observation is a p-length vector, denoted by $\boldsymbol{x}_i$, $1 \leq i \leq n$, and the data matrix is $\mathscr{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, a $n \times p$ matrix. When the data are from a multivariate normal population with mean zero and covariance matrix Σ, we have the likelihood function

$$\text{likelihood} = \prod_{i=1}^{n} f(\boldsymbol{x}_i) = \Big\{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\Big\}^n \exp\{-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{x}_i\},$$

and the log-likelihood function

$$\text{log-likelihood} = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{x}_i.$$

Recall $\Sigma = (I_p - \Phi)^{-1}\ D^2\ \{(I_p - \Phi)^{-1}\}^T$ and $\Sigma^{-1} = (I_p - \Phi)^T D^{-2} (I_p - \Phi)$. For Φ is a lower triangular matrix with diagonal entries zeros, $(I_p - \Phi)$ is a unit lower triangular matrix, and so is $(I_p - \Phi)^{-1}$. Hence, $|(I_p - \Phi)^{-1}| = 1$ and $|\Sigma| = |D^2|$.

Thus, the log-likelihood has a form as follows:

$$\text{log-likelihood} = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|D^2| - \frac{1}{2}\sum_{i=1}^{n}\Big((I_p - \Phi)\boldsymbol{x}_i\Big)^T D^{-2}\Big((I_p - \Phi)\boldsymbol{x}_i\Big).$$

For $D = \text{diag}(\sigma_1, \ldots, \sigma_p)$, we have

$$-2 \times \text{log-likelihood} = np\log(2\pi) + n\sum_{j=1}^{p} \log \sigma_j^2 + \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(x_{ij} - \sum_{k<j}\phi_{jk}x_{ik})^2}{\sigma_j^2}, \quad (2.7)$$

where x_{ij} represents the j-th component of the $\boldsymbol{x}_i$.

While the appearance of ϕ_{jk}'s and σ_j's would shape Σ^{-1} and impact Σ, the reformulation (2.7) supplies a platform for potential treatments for ϕ_{jk}'s and σ_j's so that Σ^{-1} as well as Σ can be indirectly adjusted. In contrast, directly manipulating the entries of Σ^{-1} might damage its structure. In the work of Huang et al. (2006), $\mathcal{L}_1$ regularization of LASSO type penalty (Tibshirani, 1996) on ϕ_{jk}'s has been applied so that Σ^{-1} has parsimonious properties. With focus on Σ^{-1} that has a banded structure, Levina et al. (2008) propose a modified version of $\mathcal{L}_1$ regularization, so called Nested LASSO, to select the bandwidth of Σ^{-1} adaptively by setting $\phi_{jk} = 0$ for $k < K(j)$, where the cutoff $K(j)$ depends on j.

Besides the reparameterization for Σ^{-1}, the reparameterization for Σ is also covered through sequential regressions. Combining (2.3) and (2.6), we have

$$\boldsymbol{X} = L\,\boldsymbol{E}.$$

Using the entries of L, i.e. l_{jk}'s, the individual regressions can be written as follows:

$$X_j = \sum_{k<j} l_{jk}E_k + E_j, \quad j = 2, \ldots, p. \quad (2.8)$$

The set of regressions (2.8) can be viewed as one form of Schmidt decomposition (see Horn and Johnson, 2012). Thus, a procedure as the Gram-Schmidt process of orthogonalizing the data into p orthogonal vectors of Euclidean space R^n becomes applicable. Unlike (2.2), where individual regressions do not depend on each other, the implementation of such a procedure needs to be performed in a sequential manner. These sequential regressions of (2.8) gives the representation of the covariance matrix.

We would like to see how the procedure works under a usual situation. If the sample covariance matrix $S = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i {\boldsymbol{x}_i}^T$ is non-singular, it is relatively straightforward to show the least squares estimates $\hat{l}_{jk}$ from sequentially regressing X_j on E_k, $1 \le k < j$, together with the corresponding estimates of $\hat{\sigma}_j$'s from sample variances of E_j, $1 \le j \le p$, would restore the sample covariance matrix. Denote ε_{ij} for the realized E_j from i-th observation. Below we show the estimated covariance matrix from sequential regressions is consistent as the sample covariance matrix.

$$X_1 = E_1 \quad \Rightarrow \quad \varepsilon_{i1} = x_{i1}, 1 \le i \le n, \quad \hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{i1}^2$$

$$X_2 = l_{21}E_1 + E_2 \quad \Rightarrow \quad \begin{cases} \hat{l}_{21} = \dfrac{\sum_{i=1}^{n} x_{i2}\varepsilon_{i1}}{\sum_{i=1}^{n} \varepsilon_{i1}^2}, \ \varepsilon_{i2} = x_{i2} - \hat{l}_{21}\varepsilon_{i1}, 1 \le i \le n \\ \hat{\sigma}_2^2 = \dfrac{1}{n}\sum_{i=1}^{n} \varepsilon_{i2}^2, \ \sum_{i=1}^{n} \varepsilon_{i2}\varepsilon_{i1} = 0 \end{cases}$$

$$X_3 = l_{31}E_1 + l_{32}E_2 + E_3 \quad \Rightarrow \quad \begin{cases} \hat{l}_{31} = \dfrac{\sum_{i=1}^{n} x_{i3}\varepsilon_{i1}}{\sum_{i=1}^{n} \varepsilon_{i1}^2}, \ \hat{l}_{32} = \dfrac{\sum_{i=1}^{n} x_{i3}\varepsilon_{i2}}{\sum_{i=1}^{n} \varepsilon_{i2}^2} \\ \varepsilon_{i3} = x_{i3} - \hat{l}_{31}\varepsilon_{i1} - \hat{l}_{32}\varepsilon_{i2}, 1 \le i \le n \\ \hat{\sigma}_3^2 = \dfrac{1}{n}\sum_{i=1}^{n} \varepsilon_{i3}^2, \ \sum_{i=1}^{n} \varepsilon_{i3}\varepsilon_{i1} = 0, \ \sum_{i=1}^{n} \varepsilon_{i3}\varepsilon_{i2} = 0 \end{cases}$$

$$\cdots$$

$$X_j = \sum_{k<j} l_{jk}E_k + E_j \ \Rightarrow \ \begin{cases} \hat{l}_{j1} = \dfrac{\sum_{i=1}^{n} x_{ij}\varepsilon_{i1}}{\sum_{i=1}^{n} \varepsilon_{i1}^2}, \ldots, \hat{l}_{jk} = \dfrac{\sum_{i=1}^{n} x_{ij}\varepsilon_{ik}}{\sum_{i=1}^{n} \varepsilon_{ik}^2}, \ldots, \hat{l}_{j,j-1} = \dfrac{\sum_{i=1}^{n} x_{ij}\varepsilon_{i,j-1}}{\sum_{i=1}^{n} \varepsilon_{i,j-1}^2} \\ \varepsilon_{ij} = x_{ij} - \sum_{k<j} \hat{l}_{jk}\varepsilon_{ik}, 1 \le i \le n \\ \hat{\sigma}_j^2 = \dfrac{1}{n}\sum_{i=1}^{n} \varepsilon_{ij}^2, \ \sum_{i=1}^{n} \varepsilon_{ij}\varepsilon_{i1} = 0, \ \ldots, \ \sum_{i=1}^{n} \varepsilon_{ij}\varepsilon_{i,j-1} = 0 \end{cases}$$

Therefore, the (s, t) entry of the covariance matrix estimate from the sequential regression process is

$$(\hat{\Sigma})_{st} = (\hat{L}\hat{D}^2\hat{L}^T)_{st} = \sum_{u=1}^{\min(s,t)} \hat{l}_{su}\hat{l}_{tu}\hat{\sigma}_u^2 \quad (\hat{l}_{uu} = 1).$$

Meanwhile,

$$\begin{aligned} x_{is} &= \sum_{u=1}^{s} \hat{l}_{su}\varepsilon_{iu} \quad (\hat{l}_{uu} = 1), \quad 1 \le i \le n, \\ x_{it} &= \sum_{v=1}^{t} \hat{l}_{tv}\varepsilon_{iv} \quad (\hat{l}_{vv} = 1), \quad 1 \le i \le n, \end{aligned}$$

and the (s, t) entry of the sample covariance matrix is

$$\begin{aligned} (S)_{st} &= \frac{1}{n}\sum_{i=1}^{n} x_{is}x_{it} = \frac{1}{n}\sum_{i=1}^{n}\big(\sum_{u=1}^{s}\hat{l}_{su}\varepsilon_{iu}\big)\big(\sum_{v=1}^{t}\hat{l}_{tv}\varepsilon_{iv}\big) \\ &= \frac{1}{n}\sum_{u=1}^{s}\sum_{v=1}^{t}\hat{l}_{su}\hat{l}_{tv}\big(\sum_{i=1}^{n}\varepsilon_{iu}\varepsilon_{iv}\big) \\ &= \sum_{u=1}^{\min(s,t)} \hat{l}_{su}\hat{l}_{tu}\hat{\sigma}_u^2 \quad (\hat{l}_{uu} = 1). \end{aligned}$$

The last equality holds because of

$$\sum_{i=1}^{n}\varepsilon_{iu}\varepsilon_{iv} = \begin{cases} n\sigma_u^2 & u = v; \\ 0 & u \ne v. \end{cases}$$

Thus, we've obtained the consistency

$$S = \hat{L}\ \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)\ \hat{L}^T.$$

The importance of adopting the representation for the covariance matrix using sequential regressions is to free the parameters from the positive definiteness constraint, so that potential regularization leading to improved estimates can be applied, just as in the estimation of inverse covariance matrices (Huang et al., 2006; Levina et al., 2008).

2.2 Reparametrization using Matrix Exponential

The essence of decomposition-based covariance matrix estimation is through the application of possible matrix decomposition, the parameters can be set free from the constraint of positive definiteness and potential regularization aiming to improve the estimation can be adopted. While the Cholesky decomposition is one of the most popular matrix decompositions, the spectral decomposition is also widely used. Somehow the regularization on matrix factor matrices from the spectral decomposition is not evident, and the matrix exponential technique closely related to the spectral decomposition helps filling the gap.

2.2.1 Matrix Exponential and Spectral Decomposition

The matrix exponential is a matrix function correspondent to the ordinary exponential function. Analogous to the Taylor expansion for the ordinary exponential function at the zero point, the exponential of a real or complex $p \times p$ matrix A, denoted by e^A or $\exp(A)$, is the $p \times p$ matrix given by the power series

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k, \tag{2.9}$$

where A^0 equals the $p \times p$ identity matrix, and A^k denotes ordinary matrix multiplication of A for k times. The power series always converges, so the matrix exponential of A is well-defined. Denote $A = \log(e^A)$ if A satisfies the definition of (2.9), and it is called the matrix logarithm of e^A.

For the definition involves matrix multiplication, the entries of e^A usually are not equal to the exponential of entries of A. Taking a special case for instance, when A is a diagonal matrix and the diagonal entries are $a_{11}, \dots, a_{pp}$, we write the exponential of A as

$$e^A = \begin{pmatrix} e^{a_{11}} & & & \\ & e^{a_{22}} & & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & & & \ddots \\ & & & & e^{a_{pp}} \end{pmatrix}, \quad \text{rather than} \quad \begin{pmatrix} e^{a_{11}} & & & \\ & e^{a_{22}} & & \mathbf{1} \\ & & \ddots & \\ \mathbf{1} & & & \ddots \\ & & & & e^{a_{pp}} \end{pmatrix}.$$

The matrix on the right hand side has entries equal to the exponential of the entries of A.

One of the motivations of developing matrix exponential is that it helps solving systems of linear differential equations. In Chapter 5, we use one matrix exponential equality (5.7) intensively, and the equality is also deduced in solving linear differential equations.

From the definition (2.9), A and e^A always commute, i.e. $A\,e^A = e^A A$. If they are diagonalizable, they are simultaneously diagonalizable (see Horn and Johnson, 2012). A common approach of diagonalization is through the spectral decomposition. The spectral decomposition is the factorization of a diagonalizable matrix B using its eigenvalues and eigenvectors, such as

$$B = \Gamma\,\Lambda\,\Gamma^{-1},$$

where Γ is the square matrix whose columns are the normalized eigenvectors of B and

Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues. As a special case, if B is a real symmetric matrix, the eigenvectors can be chosen to be orthonormal, and thus,

$$\Gamma^{-1} = \Gamma^T \quad \text{and} \quad B = \Gamma \, \Lambda \, \Gamma^T.$$

Therefore, with a real symmetric A, its exponential matrix e^A is also symmetric, and they can be diagonalized under the same set of eigenvectors.

2.2.2 Covariance Matrix in Matrix Exponential

In this presentation, we put our focus on positive definite covariance matrices and use $\Sigma = e^A$ to build the connection between a covariance matrix and its matrix logarithm. In such a circumstance, e^A is a positive definite matrix and the diagonalization is shown below,

$$e^A = \Gamma \, \Lambda \, \Gamma^T,$$

where Γ takes the eigenvectors of Σ as columns and Λ takes the eigenvalues $\{\lambda_1, \ldots, \lambda_p\}$ of Σ as its diagonal elements. All λ_j's are positive, $1 \le j \le p$, because of the positive definiteness of Σ. With the definition of (2.9) and $\Gamma^T = \Gamma^{-1}$, it is straightforward to show

$$\log(\Lambda) = \begin{pmatrix} \log(\lambda_1) & & \\ & \ddots & \\ & & \log(\lambda_p) \end{pmatrix}, \quad \text{and} \quad A = \log(\Sigma) = \Gamma \, \log(\Lambda) \, \Gamma^T. \qquad (2.10)$$

Result (2.10) is trivial, but the converse is not. A lemma from Chiu et al. (1996) states that for any symmetric matrix A, its exponential e^A is a positive definite matrix. Recall the motivation of the decomposition-based covariance matrix estimation, and we can tell the statement from Chiu et al. (1996) provides an alternative way of modeling the positive definite covariance matrix Σ through a potential symmetric matrix $\log(\Sigma)$ while the lower entries of $\log(\Sigma)$ are free of constraints. Potential regularization on $\log(\Sigma)$ indirectly shapes the covariance matrix estimate itself while maintaining its positive definiteness, and connections between $\log(\Sigma)$ and Σ supports this thinking. For instance, if Σ has a diagonal block structure, so does $\log(\Sigma)$ since they share the eigen-space. However, retrieving a diagonal block covariance matrix indirectly through a symmetric block matrix whose lower entries are free is much safer than retrieving the covariance matrix from direct approaches like trimming the sample covariance matrix.

There are a few other properties of matrix exponential that would be useful in covariance matrix modeling, and we list some of them here in advance.

- $\log|\Sigma| = \operatorname{tr}\big(\log(\Sigma)\big)$ where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix.
- $\Sigma^{-1} = e^{-A} \quad \text{if} \quad e^{A} = \Sigma$.
- For any $p \times p$ orthnormal matrix Q, $\log(Q\,\Sigma\,Q^T) = Q\,\log(\Sigma)\,Q^T$.

3

Regularized Estimation using Modified Cholesky Decomposition

Pourahmadi (1999) discovers that regression-based Cholesky decomposition provides an unconstrained and statistically interpretable reparameterization for a covariance matrix, and guarantees positive definiteness of its estimate based on the decomposition. Along this line of thought, Wu and Pourahmadi (2003) use the modified Cholesky decomposition of the covariance matrix in longitudinal data analysis and Huang et al. (2006) propose to impose $\mathcal{L}_1$ or $\mathcal{L}_2$ penalty on the entries of the Cholesky factor matrix for the inverse covariance matrix. From another point of view, Bickel and Levina (2008a) consider the banding structure of covariance matrices, and develop an approach of banding sample covariance matrices. In spite of the convenient usage and nice theoretical properties, there is one drawback of this approach that the banded matrix estimate is not necessarily positive definite. In order to address this issue, Rothman et al. (2010) combine the advantages from both the modified Cholesky decompsoition and the banding idea, and consider banding the Cholesky factor matrix of the covariance matrix rather than directly banding the sample covariance matrix. Application of $\mathcal{L}_1$ penalty as well as the Nested LASSO penalty (Levina et al., 2008) on the Cholesky factor matrix has also been mentioned by Rothman et al. (2010), but they do not explore the possible usage of such regularization in general situations other than the special scenarios in which covariance matrices have banded structures.

The banding assumption is rather strong and limits the usage of covariance matrix estimation through regularizing the Cholesky factor matrix. There are more general situations in which covariance structures are not banded. Meanwhile, because ordinary least squares technique is used by Rothman et al. (2010) to determine the width of the band in the Cholesky factor matrix, the band could be overly narrow if there are not enough sample points. In contrast, $\mathcal{L}_1$ regularization does not suffer as much as using the ordinary least squares from insufficient sample observations. In addition to the general feasibility of $\mathcal{L}_1$ regularization for estimating covariance matrices without assuming particular structures, we choose to impose $\mathcal{L}_1$ regularization on the Cholesky factor matrices in order to pursue parsimonious covariance matrix estimates in general situations.

For completeness, we revisit the work of Rothman et al. (2010) before we investigate the performance of applying $\mathcal{L}_1$ regularization in general cases.

3.1 Covariance Matrix Estimation via Banding Regularization

Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be a random vector with mean zero and a positive definite covariance matrix Σ. $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are *i.i.d* observations for $\boldsymbol{X}$ with $\boldsymbol{x}_i =$

$(x_{i1}, \dots, x_{ip})^T, 1 \leq i \leq n$. Denote the data matrix $\mathscr{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T$.

Recall the modified Cholesky decomposition in (2.1),

$$\Sigma = L\, D^2\, L^T,$$

where L is the Cholesky factor matrix for the covariance matrix Σ. The corresponding sequential regressions are

$$X_j = \sum_{k<j} l_{jk} E_k + E_j, \quad j = 2, \dots, p, \tag{3.1}$$

and the meanings of notations are consistent with the ones in Chapter 2.

One important property of the modified Cholesky decomposition for covariance matrix estimation is, regularization applied to the Cholesky factor matrix L would directly impact the estimated covariance matrix. For instance, a banded Cholesky factor matrix L leads to a banded covariance matrix estimate, where a banded L is simply achievable by letting only the first b subdiagonals of L non zeros and setting the rest zeros. Such an approach for estimation of inverse covariance matrices is used by Wu and Pourahmadi (2003) and Bickel and Levina (2008a), and the one for covariance matrices is used by Rothman et al. (2010).

Denote ε_{ij} for the realized E_j from i-th observation. Estimating the covariance matrix by banding its Cholesky factor matrix supplies a method of introducing parsimony while positive definiteness is guaranteed. The estimate is obtained by banding L in the decomposition of $\Sigma = LD^2L^T$ as follows:

$$\min_{\boldsymbol{l}_j} \quad \sum_{i=1}^{n} (x_{ij} - \sum_{k>j-b} l_{jk} \varepsilon_{ik})^2.$$

b is a tuning parameter, indicating the width of the band for non-zero entries in L, which usually is determined through a standard process like cross validation. We use “Band-L” to denote this method. While this approach achieves the positive definiteness of the estimated covariance matrix, it has an issue that the value of b can not exceed the number of sample points so as to let ordinary least squares technique perform, i.e, the width of the non-zero band in L can not exceed the sample size. Because the width of the non-zero band in L would determine the non-zero band in Σ, the width of non-zero band in the estimated Σ would be correspondingly constrained by 2 times of the sample size.

3.2 Covariance Matrix Estimation via $\mathcal{L}_1$ Regularization

$\mathcal{L}_1$ regularization is another popular tool in covariance matrix estimation. Huang et al. (2006) apply $\mathcal{L}_1$ regularization in the estimation of inverse covariance matrices. Comparing with the banding L approach from Rothman et al. (2010), we believe $\mathcal{L}_1$ penalty on the Cholesky factor matrix L can be used for covariance matrix estimation in general settings. For covariance matrices without banded structures, application of $\mathcal{L}_1$ regularizatio might be able to lead to improved covariance matrix estimates as well.

The sequential regressions with $\mathcal{L}_1$ regularization can be described as follows:

$$\begin{aligned} \min_{\boldsymbol{l}_j} \quad & \sum_{i=1}^{n}(x_{ij} - \sum_{k<j} l_{jk}\varepsilon_{ik})^2 \\ s.t. \quad & \sum_{k<j} |l_{jk}| < t\,, \end{aligned} \tag{3.2}$$

where $\boldsymbol{l}_j = (l_{j1}, \ldots, l_{j,j-1})$ and t is a tuning parameter. Recall ε_{ik} is the realized E_k from i-th observation. ε_{ik}'s are obtained through model fittings sequentially. The solution of (3.2) is equivalent to obtaining

$$\hat{\boldsymbol{l}}_j = \arg\min_{\boldsymbol{l}_j} \; \{ \sum_{i=1}^{n} (x_{ij} - \sum_{k<j} l_{jk}\varepsilon_{ik})^2 + \; \eta_j \sum_{k<j} |l_{jk}| \}, \quad j = 1, \ldots, p, \tag{3.3}$$

where η_j's are tuning parameters. It is impractical to tune all η_j's, $1 < j \le p$, with a large p, so we set all η_j's the same, and denote them by η. In dealing with covariance matrices with banded structures, Rothman et al. (2010) also mentioned $\mathcal{L}_1$ penalty and the Nested LASSO penalty (Levina et al., 2008) could be applied, but they did not explore the usage of such regularization in general situations.

For $1 < j \le p$, after we obtain $\hat{\boldsymbol{l}}_j = (\hat{l}_{jk})$ from (3.3), we further have $\varepsilon_{ij} = x_{ij} - \sum_{k<j} \hat{l}_{jk}\varepsilon_{ik}$, $1 \le i \le n$, and $\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ij}^2$. Therefore, the covariance matrix estimate $\hat{\Sigma}$ is obtained from

$$\hat{L} = (\hat{\boldsymbol{l}}_1, \ldots, \hat{\boldsymbol{l}}_p)^T \quad \text{and} \quad \hat{\Sigma} = \hat{L}\ \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)\ \hat{L}^T. \tag{3.4}$$

Because of the nature of applying $\mathcal{L}_1$ regularization to the Cholesky factor L matrix, we use "L_1-on-L" to represent this approach throughout this thesis.

3.2.1 Algorithms of Implementing $\mathcal{L}_1$ Regularization

Aiming to combine the advantages from both subset selection and ridge regression method, Tibshirani (1996) propose a technique, Least Absolute Shrinkage and Selection Operator (LASSO), in regression modeling. The mechanism of fitting data $(\boldsymbol{x}_i, y_i)$, $i = (1, \ldots, n)$ where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$'s are predicting variables and y_i's are responses, is defined through casting $\mathcal{L}_1$ regularization on the coefficients as follows:

$$\begin{aligned}(\hat{\alpha}, \hat{\boldsymbol{\beta}}) &= \arg\min \Big\{ \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \Big\} \\ s.t. \quad & \sum_j |\beta_j| < t\,. \end{aligned} \tag{3.5}$$

Although the optimization for LASSO is convex, because of the indifferentiable nature of absolute value function, the general fitting with $\mathcal{L}_1$ regularization does not have a closed form solution. Only for special cases such as simple linear regression or orthonormal design case in which the vectors of predictors are orthogonal to each other, there is a simple way to express the solution

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+, \tag{3.6}$$

where $\hat{\beta}_j^0$ is the ordinary least squares estimate and parameter γ is determined by the condition $\sum_j |\hat{\beta}_j| = t$. sign$(\cdot)$ function extracts the sign of the input argument number, and $(\cdot)^+$ picks the larger of the input argument and zero.

While the appearance of (3.6) is coincident with the proposal of Donoho and Johnstone (1994) by using a soft thresholding procedure for recovering functions from noisy data, the process from Donoho and Johnstone is implemented in a marginal manner. Hence, the general form with simple functions is not available. From the traditional perspective of convex optimization, an absolute value constraint can be divided into two inequality constraints. Following that idea, the very first version of LASSO fitting is using traditional quadratical programming under linear inequality constraints. Because the function to be optimized is convex, the convergence of LASSO fitting through quadratical programming is obtainable. However, the computation in this approach is intense and exceeds the supporting resources quickly along with the increase of variable numbers.

As the first one attempting coordinate optimization techniques to solve the regression with $\mathcal{L}_1$ regularization, Fu (1998) proposes the so called “shooting” method by cyclically setting each variable to the minimum using the modified Newton-Raphson algorithm. Perhaps because Fu’s work is mainly to develop a new type of regularized regression, bridge regression, and the “shooting” method is an auxiliary tool for solving that, unfortunately the “shooting” method does not get into the highlight of the stage until recent years, even though it is developed right after the proposal of the regression $\mathcal{L}_1$ regularization.

While the idea of LASSO regression in the area of signal processing is also known as basis pursuit (Chen et al., 1998), some people use interior-point algorithms to solve equivalent optimizations.

With the modification of $\mathcal{L}_1$ constraints, Fan and Li (2001) develop a penalty function called Smoothly Clipped Absolute Deviation (SCAD). Fan and Li suggest using the locally quadratic approximation and iteratively solving the optimization through coordinate optimization. This suggestion is being viewed directly applicable in solving the regression with $\mathcal{L}_1$ regularization, with a simple approximation as follows:

$$|\beta_j^{\text{new}}| = \frac{1}{2}\frac{(\beta_j^{\text{new}})^2}{|\beta_j^{\text{old}}|} + \frac{1}{2}|\beta_j^{\text{old}}|. \tag{3.7}$$

The local quadratic approximation (3.7) has been proved useful in fitting LASSO regression or related topics. For instance, Levina et al. (2008) use this technique dealing with their Nested LASSO penalty.

Efron et al. (2004) provide another procedure called Least Angle Regression (LARS) for problems like LASSO and forward stagewise regression. Unlike other methods starting with coefficients from the ordinary least squares estimates, LARS starts with all coefficients equal to zero. Carried out in successive steps, LARS imports only one variable into the model at each step until the number of variables reaches the sample size. Although the starting point of LARS is different from other algorithms, and even the setting of (3.5) is not included, LARS produces very similar outcomes as fitting the regression with $\mathcal{L}_1$ regularization. Furthermore, with an extra modification, the LARS algorithm yields all LASSO solutions. LARS algorithm is significantly faster than quadratic programming in solving LASSO regression, and the algorithm stops nicely under the given completion condition.

In the work of Friedman et al. (2010), the idea of cyclical coordinate optimization in LASSO regression and other generalized linear models with regularization has been comprehensively revisited. Similar as "shooting" algorithm, this set of methods focus on one variable each time while keeping others constant. Moreover, for the closed form formula of updating coefficients using the coordinate descent has been developed and discussed, the algorithm for fitting the regression with $\mathcal{L}_1$ regularization becomes straightforward. When the increase of computation along with the growth of variable sizes gets under control, the algorithm can handle much larger problems with considerably higher efficiency.

While the coordinate descent becomes a routine for fitting the regression with $\mathcal{L}_1$ regularization, we also rely on this approach to perform our work.

3.2.2 Choice of Tuning Parameter

Recall that in the consideration of potentially large p, we set the η_j in (3.3) to be a common value η for practical purposes. We adopt a procedure of repeated learning-testing (Burman, 1989) to choose this tuning parameter η. Specifically, we repeatedly split the data set into a learning set and a testing set with roughly equal sizes for V times. Let $\hat{\Sigma}^{(v)}(\eta)$ be the estimated covariance matrix based on the learning data with tuning parameter η in the v-th replicate, $v = 1, \ldots, V$, and let $S^{(v)}$ be the sample covariance matrix obtained from the testing data in the v-th replicate.

The calculation would be carried out through all V replicates, and then the tuning parameter is chosen to be

$$\hat{\eta} = \underset{\eta}{\text{argmin}} \ \frac{1}{V} \sum_{v=1}^{V} ||\hat{\Sigma}^{(v)}(\eta) - S^{(v)}||_F,$$

where $\|\cdot\|_F$ is F norm defined in (1.11). Through the simulation study, we compared the results based on three different norms: the induced L_1 norm, the induced L_2 norm and the F norm, and we found that the differences are minimal. The L_1, L_2 of a matrix $A = (a_{st})$ are denoted by $||A||_1$ and $||A||_2$, respectively. They are defined as follows:

$$||A||_1 = \max_t \sum_s |a_{st}|, \quad \text{and} \quad ||A||_2 = \sqrt{\lambda_{\max}(A^T A)}\,,$$

where $\lambda_{\max}(M)$ denotes the largest eigenvalue of the square matrix M. Based on the similarities of results corresponding to different norms, we use F norm as the measure in choosing the tuning parameter throughout our numerical studies.

Other methods of choosing the tuning parameter include cross validation (see

Stone, 1974) and information based criterion such as Bayesian information criterion (Schwarz, 1978). Here we adopt a repeated learning-testing procedure because we aim at a balance between estimating the covariance matrix and calculating the sample covariance matrix. While the covariance matrix estimator needs reasonable sample size to perform, we need to take care of the sample covariance matrix as well since it functions as a bench mark and also needs modest amounts of sample points to provide a reasonable standard. Equally splitting the samples into both sides is supposed to deal with these issues better, given the limited sample points. The choice of V might be a concern in the computation cost, and we find that setting V to be 20 gives stable tuning parameter estimation in our simulations and real case studies.

3.3 Simulation Study

In this section, we would like to show simulation results of various estimates corresponding to different patterns of a covariance matrix. Scenarios according to covariance matrix structures being considered are listed below:

- Compact Banding Structure;
- Permuted Banding Structure;
- Loose Banding Structure;
- Block Diagonal Structure;
- Block Diagonal Structure with Permutation.

We compare the performance of our approach with five other methods: sample covariance matrix, LW estimate, thresholding, Banding Cholesky factor estimate and

Bien's method.

As introduced in Chapter 1, the LW estimate is proposed by Ledoit and Wolf (2004) through minimizing the squared F norm of the difference between the estimated and true covariance matrices, and the estimate is in the form of

$$\hat{\Sigma}_{LW} = (1-\rho)\, S + \rho\, \nu\, I_p,$$

where I_p is the p-dimensional identity matrix and S is the sample covariance matrix. Ledoit and Wolf give the closed form formula for ρ and ν through

$$\begin{aligned}
\hat{\nu} &= \frac{1}{p}\,\mathrm{tr}(S), \\
\hat{\rho} &= \frac{\mathrm{tr}(\Sigma^2) - \mathrm{tr}(S^2)}{\mathrm{tr}(S - \hat{\nu} I_p)^2} \approx \frac{\sum_{s=1}^{p}\sum_{t=1}^{p}\{\frac{1}{n}\sum_{i=1}^{n} x_{is}^2 x_{it}^2\} - \mathrm{tr}(S^2)}{\mathrm{tr}(S - \hat{\nu} I_p)^2},
\end{aligned}$$

where $\mathrm{tr}(\cdot)$ is the trace function for square matrices.

Another popular method in covariance matrix estimation is to use thresholding. With a data driven threshold parameter, entries of the sample covariance matrix whose absolute values are below that would be set zeros. This method is fast and intuitively, but the drawback is also clear that the structure of the covariance matrix might get destroyed. For instance, the property of positive definiteness is no longer protected.

With the assumption of multivariate normal distribution, another perspective of achieving regularized covariance matrix estimation is to add $\mathcal{L}_1$ penalty of Σ to the log-likelihood function, resulting in an estimate

$$\hat{\Sigma} = \underset{\Sigma \succ 0}{\mathrm{argmin}} \Big\{ -\log|\Sigma^{-1}| + \mathrm{tr}(\Sigma^{-1} S) + \eta \sum_{s}\sum_{t} |\sigma_{st}| \Big\}\ ,$$

where σ_{st}'s are entries of Σ and η is a tuning parameter. The main difficulty of this approach is that the function to be optimized is not convex in term of Σ and the solution might not be unique. Bien and Tibshirani (2011) develop a majorize-minimize algorithm for this optimization problem in which they iteratively solve convex approximations to the original non-convex problem. We call this method the Bien's method.

In order to evaluate the performance of different methods, we refer to some commonly used criteria. The first group of criteria are matrix norms, including the induced L_1, L_2 norm, and the F norm. The norms of the difference matrix between the estimated covariance matrix and the true one are reported as norm-based loss measures.

Defined in (1.8), the entropy loss is also widely used in measuring closeness of two square matrices. Another popular loss function is Kullback-Leibler divergence (KL divergence), and it is similar to the entropy loss except we interchange the symbols for the true covariance matrix and its estimate in the entropy loss formula. The formula is shown below,

$$\text{KL divergence} = \text{tr}(\hat{\Sigma}^{-1}\Sigma) - \log|\hat{\Sigma}^{-1}\Sigma| - p. \tag{3.8}$$

As pointed out by Levina et al. (2008), KL divergence is more suitable in measuring the inverse covariance matrix, while the entropy loss is a more appropriate measure if the primary interest is the covariance matrix itself. Therefore, we consider the entropy loss rather than KL divergence in our numerical studies.

Regarding the eigen-structure of covariance matrix estimates, we consider the accuracy in estimating the condition number. The condition number, originally used to measure how far the output value would vary for a small change in the input, equals the ratio of the largest eigenvalue to the smallest eigenvalue for a positive definite matrix. Therefor, in this study, the condition number ($\lambda_{\max}/\lambda_{\min}$), is also a rational measure. However, this measure is not a loss function or has a baseline zero; instead, the reference line is the condition number of the true covariance matrix. Taking the difference of condition numbers is not straightforward to tell the control of eigenvalues for possible estimate. Thus, we would like to keep the raw values of condition numbers. To make the presentation clear, we add one extra "estimate" using the true covariance matrix into the comparisons. The value of condition number from for this "estimate" serves as the reference. While for other measures like F norm or entropy loss, the corresponding values for this "estimate" would be zeros.

For each scenario, we generated normally distributed data with three settings: (1) $n = 50$, $p = 30$; (2) $n = 50$, $p = 50$; (3) $n = 50$, $p = 100$. Each case was repeated 200 times, and the average values from the 200 loss measures were reported as well as their corresponding standard errors.

3.3.1 Compact Banded Structure

We design the true covariance matrix Σ_1 to have an order-1 moving average (MA(1)) structure. Specifically, $\Sigma_1 = \Sigma_1(\sigma_{st})$ is a tri-diagonal and Toeplitz matrix with

$$\sigma_{st} = \begin{cases} 1, & s = t; \\ 0.4, & |s - t| = 1; \\ 0 & \text{otherwise.} \end{cases}$$

p	Measure	Sample	LW	Threshold	Band-L	Bien's	L_1-on-L
30	L_1 norm	4.85 (0.04)	2.05 (0.02)	1.58 (0.03)	0.86 (0.01)	2.47 (0.01)	1.40 (0.01)
	L_2 norm	2.04 (0.02)	0.88 (0.00)	1.04 (0.01)	0.67 (0.01)	1.21 (0.00)	0.90 (0.01)
	F norm	4.29 (0.02)	2.53 (0.00)	2.65 (0.02)	1.51 (0.01)	2.95 (0.01)	2.30 (0.01)
	entropy	12.58 (0.06)	7.23 (0.04)	-	1.42 (0.02)	9.67 (0.07)	3.90 (0.04)
50	L_1 norm	7.92(0.05)	2.47 (0.02)	1.78 (0.03)	0.93 (0.01)	3.42 (0.02)	1.50 (0.01)
	L_2 norm	2.92(0.02)	0.90 (0.00)	1.13 (0.01)	0.73 (0.01)	1.44 (0.00)	0.95 (0.00)
	F norm	7.11(0.02)	3.50 (0.00)	3.68 (0.02)	1.97 (0.01)	4.56 (0.01)	3.13 (0.01)
	entropy	-	14.70 (0.06)	-	2.39 (0.02)	32.73 (0.10)	7.98 (0.06)
100	L_1 norm	15.44 (0.14)	3.11 (0.05)	1.99 (0.06)	1.01 (0.03)	5.85 (0.05)	1.61 (0.02)
	L_2 norm	4.84 (0.05)	0.92 (0.00)	1.26 (0.02)	0.80 (0.02)	1.67 (0.00)	1.00 (0.01)
	F norm	14.13 (0.05)	5.28 (0.01)	5.56 (0.03)	2.77 (0.03)	8.21 (0.01)	4.73 (0.02)
	entropy	-	34.49 (0.18)	-	4.78 (0.07)	173.06 (0.09)	20.06 (0.17)

Table 3.1: Performance of estimates from various methods in Scenario 1 (Compact Banded Structure). Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

p	Σ	Sample	LW	Band-L	Bien's	L_1-on-L
30	8.80	115.95 (2.47)	2.90 (0.03)	14.13 (0.19)	58.77 (1.34)	6.17 (0.08)
50	8.92	-	2.55 (0.02)	15.98 (0.18)	191.58 (1.04)	5.76 (0.06)
100	8.98	-	2.28 (0.03)	18.15 (0.40)	213.10 (2.50)	5.16 (0.10)

Table 3.2: Summary of condition numbers ($\lambda_{\text{max}}/\lambda_{\text{min}}$) of different estimates in Scenario 1 (Compact Banded Structure). The values in the column for the true Σ serve as references. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

The performance of different approaches is listed in **Table 3.1** and **Table 3.2**. **Table 3.1** shows the norm-based loss and entropy loss measures. Dashed line in the table indicates the corresponding values are either not achievable or infinite. For various variable dimension choices and error measure terms, L_1-on-L outperforms other methods except Band-L method which catches the structure more precisely. **Table 3.2** shows the summary for condition numbers of different estimates. To compare the results, the condition number for the true covariance matrix is listed for reference. The condition numbers from L_1-on-L estimate are closer to true values than other estimates in comparison.

3.3.2 Permuted Banded Structure

Scenario 2 is developed to investigate the sensitivities of different approaches regarding covariance structure with permutation. We randomly permuted rows and columns of Σ_1 to generate Σ_2 and applied all the methods in dealing with Σ_2 with 200 replicates.

The performance of different approaches is listed in **Table 3.3** for loss measures and **Table 3.4** for condition number accuracy.

p	Measure	Sample	LW	Threshold	Band-L	Bien's	L_1-on-L
30	L_1 norm	4.87 (0.04)	2.07 (0.02)	1.57 (0.03)	1.27 (0.01)	2.49 (0.01)	1.45 (0.01)
	L_2 norm	2.03 (0.02)	0.88 (0.01)	1.03 (0.01)	0.99 (0.00)	1.22 (0.01)	0.91 (0.01)
	F norm	4.30 (0.02)	2.53 (0.01)	2.65 (0.02)	3.24 (0.00)	2.95 (0.01)	2.35 (0.01)
	entropy	12.58 (0.06)	7.22 (0.04)	-	12.76 (0.06)	9.65 (0.07)	4.32 (0.04)
50	L_1 norm	7.98 (0.05)	2.50 (0.02)	1.77 (0.03)	1.31 (0.01)	3.45 (0.02)	1.59 (0.01)
	L_2 norm	2.96 (0.02)	0.91 (0.00)	1.13 (0.01)	1.02 (0.00)	1.44 (0.00)	0.96 (0.00)
	F norm	7.11 (0.02)	3.50 (0.00)	3.69 (0.01)	4.21 (0.00)	4.56 (0.01)	3.21 (0.01)
	entropy	-	14.68 (0.06)	-	21.95 (0.08)	32.78 (0.10)	8.70 (0.06)
100	L_1 norm	15.36 (0.13)	3.07 (0.05)	1.99 (0.07)	1.36 (0.01)	5.86 (0.05)	1.78 (0.04)
	L_2 norm	4.83 (0.04)	0.91 (0.00)	1.26 (0.03)	1.05 (0.01)	1.67 (0.00)	1.01 (0.01)
	F norm	14.09 (0.05)	5.28 (0.01)	5.58 (0.03)	5.97 (0.01)	8.21 (0.01)	4.82 (0.02)
	entropy	-	34.49 (0.17)	-	44.42 (0.20)	173.09 (0.10)	21.19 (0.18)

Table 3.3: Performance of estimates from various methods in Scenario 2 (Permuted Banded Structure). Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

p	Σ	Sample	LW	Band-L	Bien's	L_1-on-L
30	8.80	115.54 (2.27)	2.91 (0.03)	2.31 (0.02)	58.28 (1.18)	5.74 (0.06)
50	8.92	-	2.58 (0.02)	2.57 (0.02)	192.19 (0.97)	5.57 (0.06)
100	8.98	-	2.26 (0.03)	2.79 (0.05)	214.18 (2.44)	5.22 (0.10)

Table 3.4: Summary of condition numbers ($\lambda_{\max}/\lambda_{\min}$) of different estimates in Scenario 2 (Permuted Banded Structure). The values in the column for the true Σ serve for reference. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

Comparing the performance reported in **Table 3.3** and **Table 3.4**, we could tell L_1-on-L overrides other methods including Band-L method in this scenario since Band-L is highly sensitive to the assumption of the banded structure.

Although L_1-on-L approach is not directly order invariant, comparisons of its performance between **Table 3.1** and **Table 3.3** imply L_1-on-L estimate is rather robust towards potential permutations of covariance structures.

3.3.3 Loose Banded Structure

With the existence of seasonal effects, we may encounter a loose banding structure in which there is a gap between non-zero covariances in the previous order-1 moving average (MA(1)) structure. That's why we consider another structure $\Sigma_3 = \Sigma_3(\sigma_{st})$ as follows:

$$\sigma_{st} = \begin{cases} 1, & s = t; \\ 0.4, & |s - t| = p/5; \\ 0 & \text{otherwise.} \end{cases}$$

The performance of various approaches is reported in **Table 3.5** and **Table 3.6**, and the measures suggest that the L_1-on-L estimate is superior to estimates from other methods in this scenario.

p	Measure	Sample	LW	Threshold	Band-L	Bien's	L_1-on-L
30	L_1 norm	4.88 (0.04)	1.87 (0.02)	1.39 (0.02)	2.19 (0.03)	2.39 (0.01)	1.30 (0.01)
	L_2 norm	2.02 (0.02)	0.79 (0.00)	0.95 (0.01)	1.06 (0.01)	1.18 (0.00)	0.85 (0.01)
	F norm	4.31 (0.02)	2.34 (0.00)	2.47 (0.01)	2.73 (0.01)	2.92 (0.01)	2.18 (0.01)
	entropy	12.58 (0.06)	4.84 (0.03)	-	5.61 (0.07)	9.41 (0.07)	3.09 (0.03)
50	L_1 norm	7.96 (0.05)	2.22 (0.02)	1.57 (0.03)	1.28 (0.01)	3.30 (0.01)	1.41 (0.01)
	L_2 norm	2.93 (0.02)	0.81 (0.00)	1.06 (0.01)	0.96 (0.01)	1.39 (0.00)	0.90 (0.00)
	F norm	7.13 (0.02)	3.20 (0.00)	3.41 (0.01)	3.81 (0.00)	4.50 (0.01)	2.95 (0.01)
	entropy	-	9.71 (0.04)	-	14.00 (0.05)	33.01 (0.10)	6.02 (0.05)
100	L_1 norm	15.39 (0.12)	2.74 (0.04)	1.74 (0.05)	1.32 (0.01)	5.58 (0.06)	1.59 (0.04)
	L_2 norm	4.77 (0.04)	0.82 (0.00)	1.16 (0.02)	0.99 (0.01)	1.60 (0.00)	0.95 (0.01)
	F norm	14.11 (0.05)	4.79 (0.01)	5.12 (0.02)	5.41 (0.01)	8.11 (0.01)	4.44 (0.02)
	entropy	-	22.40 (0.11)	-	28.55 (0.15)	177.16 (0.10)	14.44 (0.18)

Table 3.5: Performance of estimates from various methods in Scenario 3 (Loose Banded Structure). Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

p	Σ	Sample	LW	Band-L	Bien's	L_1-on-L
30	5.51	93.41 (1.97)	2.54 (0.03)	21.09 (0.55)	44.88 (0.97)	4.76 (0.05)
50	5.51	-	2.28 (0.02)	2.77 (0.19)	177.75 (1.07)	4.77 (0.06)
100	5.51	-	2.05 (0.03)	2.81 (0.05)	193.68 (3.22)	4.62 (0.10)

Table 3.6: Summary of condition numbers ($\lambda_{\max}/\lambda_{\min}$) of different estimates in Scenario 3 (Loose Banded Structure). The values in the column for the true Σ serve for reference. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

To better understand the behaviors of all these methods, we illustrate estimates for five methods with $p = 50$ using heat maps in **Figure 3.1** for one simulated replicate.

We purposely use the same scale of darkness for all heat maps, so that the absolute values of entries in the matrices can be directly compared by the extent of darkness. The image for the estimate from threshold method has the clearest appearance. Threshold method correctly identifies a large amount of zero entries in its estimate. However, many truly non-zero ones are also misidentified as zeros while a few zero entries are incorrectly identified as non-zeros. The mistakenly identified zero entries as well as the mistakenly identified non-zero entries invite difficulties in capturing the true structure, and that explains why its performance measures **Table 3.5** are not as competitive as it appears in **Figure 3.1**. In contrast, the two sub-diagonals for estimates from LW and Bien's method are not clear. For the estimate from LW method, because of the large number of zero entries, the entries in the main diagonal are enforced in terms of absolute values of their magnitude, and comparably the ones not in the main diagonal are weakened including the two truly non-zero

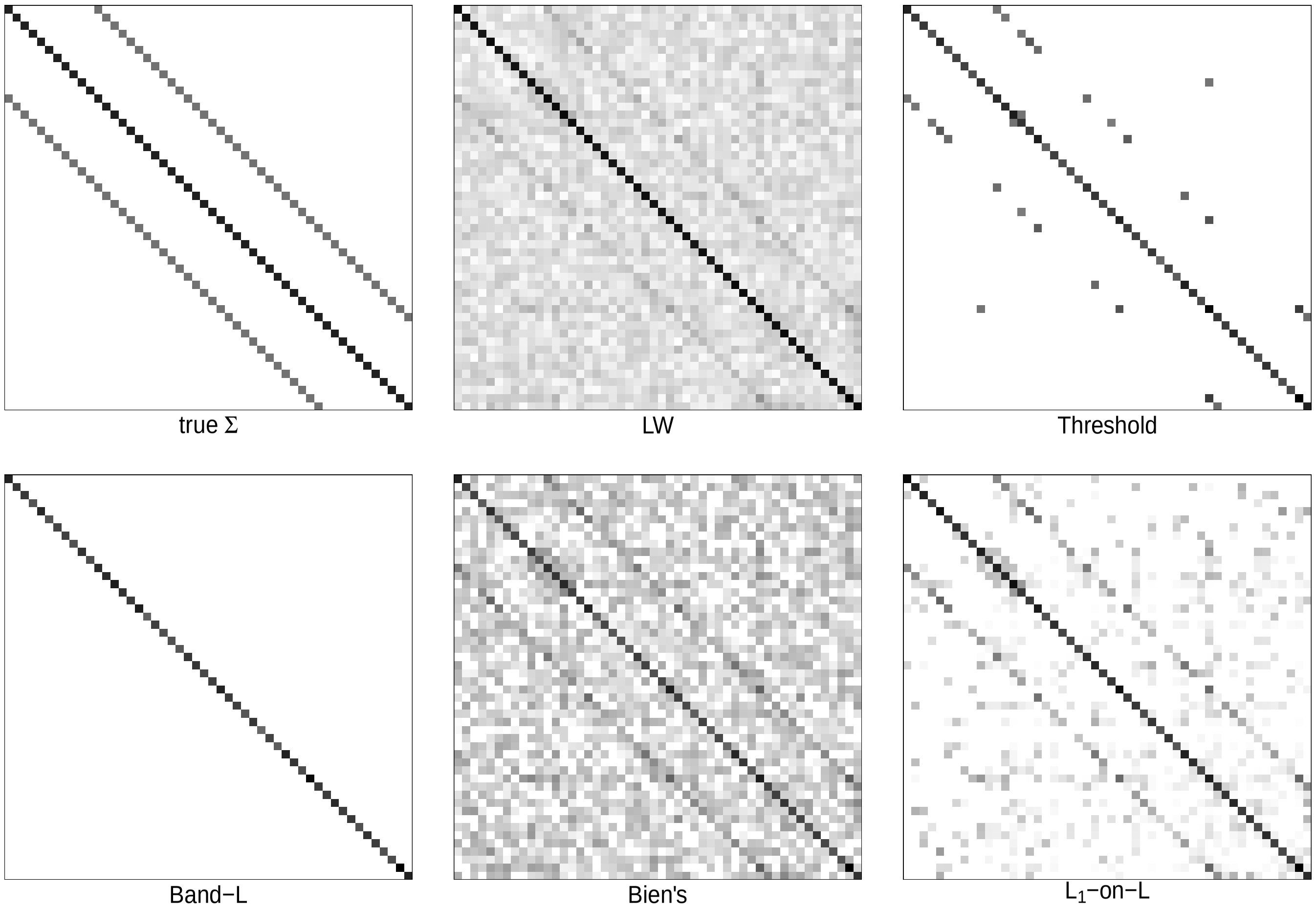

Figure 3.1: Heat maps of the true covariance matrix and various estimates from one replicate of simulations in Scenario 3 (Loose Banded Structure). Absolute values of entries are used to replace original entries. An entry of magnitude 1 or over is represented by a black square and an entry of magnitude 0 is represented by a white square.

sub-diagonals, which impairs its efficiency in covariance matrix estimation. For the estimate from Bien's method, because of the nature of the optimization, the penalty can not be very heavy, so that it has a less controlled structure and not enough zero entries are identified. For the estimate from Band-L method, for the disturbance of potential noise in the gap area, the tuning procedure may encounter great difficulty in determining the width of the band. In this estimate, the two sub-diagonal entries are abandoned, which entirely toppled down the estimation. On the other hand, even if these two sub-diagonals are detected correctly, the entries in the gap area between the two sub-diagonals and the diagonal would be filled with non-zero entries, in which the appearance of such noise wound also impair the performance. The estimate from L_1-on-L method, though could not obtain zero entries as efficient as Threshold and Band-L methods, captures better the big picture for the true covariance matrix.

3.3.4 Block Diagonal Structure

A covariance matrix with roughly a block diagonal structure is often seen, especially when several groups of variables are involved while variables in different groups may not have much interaction. To see the performance of covariance estimation methods in dealing with this type of situations, we design the scenario with covariance matrix Σ_4 in which the first 20% variables are closely correlated while others are unrelated. Specifically,

$$\Sigma_4 = \{\sigma_{st}\} \quad \text{with} \quad \sigma_{st} = \begin{cases} 1 & s = t; \\ 0.8 & s \neq t, s \leq p/5, t \leq p/5; \\ 0 & \text{otherwise.} \end{cases}$$

The performance of various approaches is listed in **Table 3.7** and **Table 3.8**.

p	Measure	Sample	LW	Threshold	Band-L	Bien's	L_1-on-L
30	L_1 norm	4.97 (0.04)	3.71 (0.03)	1.29 (0.04)	1.95 (0.03)	3.33 (0.04)	2.66 (0.05)
	L_2 norm	2.23 (0.03)	2.40 (0.04)	0.99 (0.04)	1.19 (0.03)	1.92 (0.04)	2.05 (0.04)
	F norm	4.31 (0.02)	3.19 (0.02)	1.48 (0.03)	2.49 (0.02)	3.23 (0.03)	2.60 (0.04)
	entropy	12.58 (0.06)	6.45 (0.04)	-	3.59 (0.03)	10.15 (0.06)	2.69 (0.04)
50	L_1 norm	8.40 (0.08)	6.32 (0.04)	2.30 (0.08)	3.27 (0.05)	4.93 (0.08)	4.67 (0.08)
	L_2 norm	3.62 (0.06)	4.02 (0.07)	1.65 (0.07)	1.85 (0.06)	2.80 (0.07)	3.56 (0.07)
	F norm	7.17 (0.03)	5.35 (0.04)	2.26 (0.06)	4.10 (0.03)	4.81 (0.05)	4.33 (0.07)
	entropy	-	12.85 (0.04)	-	10.36 (0.05)	33.05 (0.12)	6.17 (0.18)
100	L_1 norm	17.00 (0.26)	12.98 (0.18)	5.00 (0.33)	6.32 (0.22)	9.78 (0.36)	9.96 (0.31)
	L_2 norm	7.04 (0.24)	8.16 (0.30)	3.25 (0.27)	3.40 (0.25)	5.64 (0.31)	7.03 (0.25)
	F norm	14.29 (0.12)	10.81 (0.14)	4.13 (0.26)	7.78 (0.15)	9.07 (0.23)	8.66 (0.28)
	entropy	-	34.30 (0.38)	-	47.04 (0.24)	170.12 (0.55)	30.40 (3.13)

Table 3.7: Performance of estimates from various methods in Scenario 4 (Block Diagonal Structure). Σ serve as the reference. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

p	Σ	Sample	LW	Band-L	Bien's	L_1-on-L
30	25	195.92 (5.09)	6.70 (0.21)	52.42 (1.04)	130.25 (3.43)	27.64 (0.59)
50	41	-	11.87 (0.39)	145.90 (3.20)	607.19 (7.55)	66.82 (1.71)
100	81	-	24.06 (1.67)	2429 (153)	1188 (33)	435 (61)

Table 3.8: Summary of condition numbers ($\lambda_{\text{max}}/\lambda_{\text{min}}$) of different estimates in Scenario 4 (Block Diagonal Structure). The values in the column for the true Σ serve for reference. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

With high correlations among first 20% variables, Threshold and Band-L, outperform others in terms of norm measures. However, in term of entropy loss, their performance is not as desirable as norm measures. Threshold method could not guarantee the positive definiteness of its estimate even though the difference between zero and non-zero entries is set rather large. The estimate from Band-L method contains much noise in the matrix band because the length of the band is determined by the block in the upper left corner, even if majority of the band entries should actually be zeros. The behaviors of two $\mathcal{L}_1$ regularization based methods, Bien's method and L_1-on-L, produce estimates with similar patterns, but in term of the entropy loss and condition number, the estimate from Bien's method is not as good as the one from L_1-on-L.

3.3.5 Block Diagonal Structure with Permutation

A general situation more close to realistic problems is that the covariance matrix has a structure equivalent to block diagonal, but without the information of group relationships among variables, we are unable to tell the structure directly. We explore such a situation by simulating data with covariance matrix Σ_5. Σ_5 is generated by randomly permute the rows of Σ_4 and the corresponding columns.

The performance of various approaches is reported in **Table 3.9** and **Table 3.10**. Similar as scenario 3, we also present the set of estimates for one replicate in **Figure 3.2** to better explain the results. Threshold method produces the cleanest estimate, and thus wins much with norm measures. The drawback is that the positive definite structure is still under threat even if just a few truly zero entries are mistakenly set non-zeros or vice versa. For the true covariance structure is no longer banded, Band-L struggles to operate, and in the case, only the main diagonal remains in the estimate. Both Bien's and L_1-on-L method import certain amount of incorrect non-zero entries into their estimates, but between them, it seems the noise level for the estimate from L_1-on-L is lower, which explains why the L_1-on-L estimate suffers less entropy loss and has better control of the condition number.

p	Σ	Sample	LW	Band-L	Bien's	L_1-on-L
30	25	190.92 (4.35)	7.10 (0.21)	146.86 (5.20)	125.51 (2.96)	28.42 (0.62)
50	41	-	12.36 (0.41)	42793 (5459)	606.31 (7.23)	66.39 (1.57)
100	81	-	23.56 (1.56)	-	1185 (31.74)	259 (17.44)

Table 3.9: Summary of condition numbers ($\lambda_{\max}/\lambda_{\min}$) of different estimates in Scenario 5 (Permuted Block Diagonal Structure). The values in the column for the true Σ serve for reference. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

The performance comparisons between **Table 3.7** and **Table 3.10** for L_1-on-L, suggests its robustness towards variable permutations.

p	Measure	Sample	LW	Threshold	Band-L	Bien's	L_1-on-L
30	L_1 norm	5.00 (0.05)	3.64 (0.03)	1.29 (0.05)	4.77 (0.04)	3.24 (0.04)	2.58 (0.05)
	L_2 norm	2.24 (0.03)	2.29 (0.04)	0.98 (0.04)	2.31 (0.05)	1.86 (0.04)	2.02 (0.04)
	F norm	4.32 (0.02)	3.15 (0.02)	1.47 (0.03)	4.24 (0.02)	3.19 (0.03)	2.54 (0.04)
	entropy	12.58 (0.06)	6.39 (0.04)	-	12.84 (0.16)	10.13 (0.06)	2.56 (0.03)
50	L_1 norm	8.50 (0.08)	6.31 (0.05)	2.33 (0.08)	8.18 (0.05)	4.84 (0.08)	4.65 (0.08)
	L_2 norm	3.63 (0.06)	3.95 (0.07)	1.67 (0.07)	3.63 (0.07)	2.77 (0.07)	3.60 (0.08)
	F norm	7.20 (0.04)	5.33 (0.03)	2.27 (0.06)	7.07 (0.03)	4.78 (0.05)	4.31 (0.07)
	entropy	-	12.85 (0.04)	-	48.34 (0.37)	32.91 (0.12)	5.52 (0.07)
100	L_1 norm	17.11 (0.26)	12.97 (0.17)	4.94 (0.33)	15.70 (0.13)	9.90 (0.35)	10.24 (0.30)
	L_2 norm	7.00 (0.22)	8.21 (0.29)	3.16 (0.25)	9.79 (0.19)	5.71 (0.31)	7.61 (0.28)
	F norm	14.26 (0.11)	10.83 (0.13)	4.05 (0.23)	15.01 (0.04)	9.13 (0.23)	9.00 (0.29)
	entropy	-	34.17 (0.38)	-	232.09(3.23)	170.30 (0.56)	22.96 (1.73)

Table 3.10: Performance of estimates from various methods in Scenario 5 (Permuted Block Diagonal Structure). Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

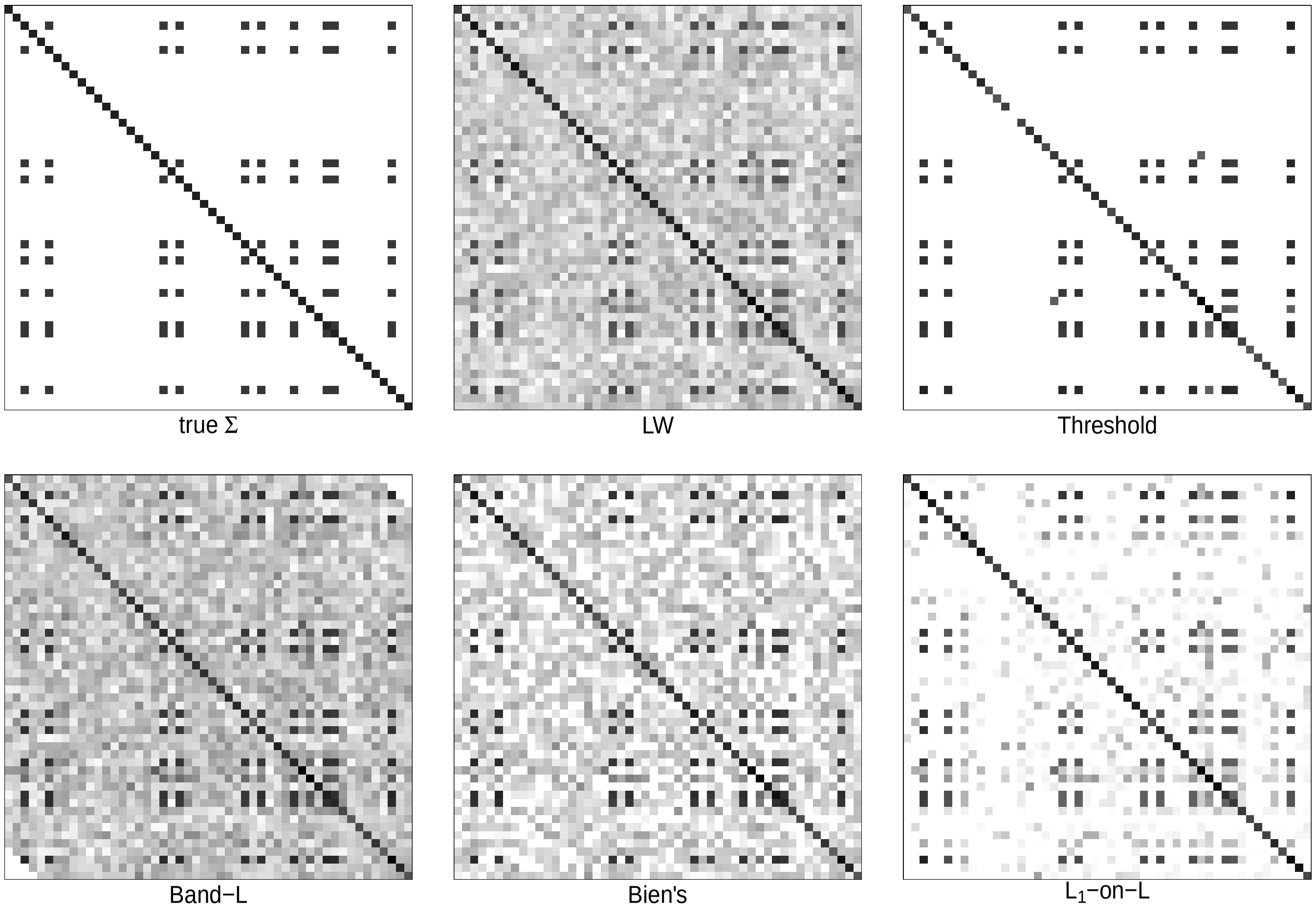

Figure 3.2: Heat maps of the true covariance matrix and various estimates from one replicate of simulations in Scenario 5 (Permuted Block Diagonal Structure). Absolute values of entries are used to replace original entries. An entry of magnitude 1 or over is represented by a black square and an entry of magnitude 0 is represented by a white square.

3.4 Real Data Analysis

In this section, two real data cases are presented to show the performance of estimates from the L_1-on-L approach.

3.4.1 A Blue-cell Tumor Data Study

Analyzing connections among genes using their covariance structure is a widely used technique in gene expression data analysis. We applied L_1-on-L to a data set from a small round blue-cell tumor microarray experiment (Khan et al., 2001). The original data set included 6,567 genes and was filtered down to 2,308 according to their intensities. This data set contains 63 tissue samples, and there are four types of tumors in the sample.

Following the work by Rothman et al. (2009), we ranked the genes by calculating the F-statistics which is used to compare means of sub-groups in one way-ANOVA analysis,

$$F = \frac{\frac{1}{k-1} n_m (\bar{x}_m - \bar{x})^2}{\frac{1}{n-k} \sum_{m=1}^{k} (n_m - 1) \hat{\sigma}_m^2} \ . \tag{3.9}$$

$n = 63$ is the total sample size; $k = 4$ is the number of sub-groups; $n_m, \bar{x}_m, \hat{\sigma}_m^2$ are the sample sizes, sample means and sample covariances for different sub-groups, with m indexing the sub-groups, $1 \leq m \leq 4$; $\bar{x}$ is the grand mean. The 40 genes with the largest F values from (3.9) and bottom 160 genes with smallest F values were chosen. Since F values measure the extents of differential expression for genes, the selection of the 40+160 genes has an intention that the genes in first group are closely correlated while the correlation between the first and second groups as well as within the second group is not quite significant.

To select the tuning parameter associated with $\mathcal{L}_1$ regularization in L_1-on-L, we repeatedly divided the 63 observations into learning and testing groups of roughly equal sizes. Because there are four types of tumors in the sample, we made the proportions of four types of tumors in each group be nearly the same, so that both the learning and testing data are good representatives of the whole.

We show the heat maps for the sample covariance matrix and the estimate from L_1-on-L in **Figure 3.3** in the same scale of darkness. Absolute values of entries are used in replace of original entries, and the darkness for the cells directly indicate the magnitude of the corresponding entries. For the top-left corners of both maps, which are corresponding to the 40 most differentially expressed genes, the appearances are rather alike, while the remaining area of the maps has evident differences. The absolute values of entries outside the top-left corner have been greatly reduced by L_1-on-L compared with the ones from the sample covariance matrix, which represents the improvement brought by L_1-on-L since those values are expected to be tiny.

3.4.2 A Parkinson Data Study

To further explore the performance of the L_1-on-L approach, we carried out another real data analysis. Scherzer et al. (2007) performed a transcriptome-wide scan to investigate the the molecular processes perturbed in cellular blood of patients with early-stage Parkinson's disease (PD). They probed RNA extracted from whole blood of 50 PD patients at early disease stages, and 55 age-matched controls using microarrays.

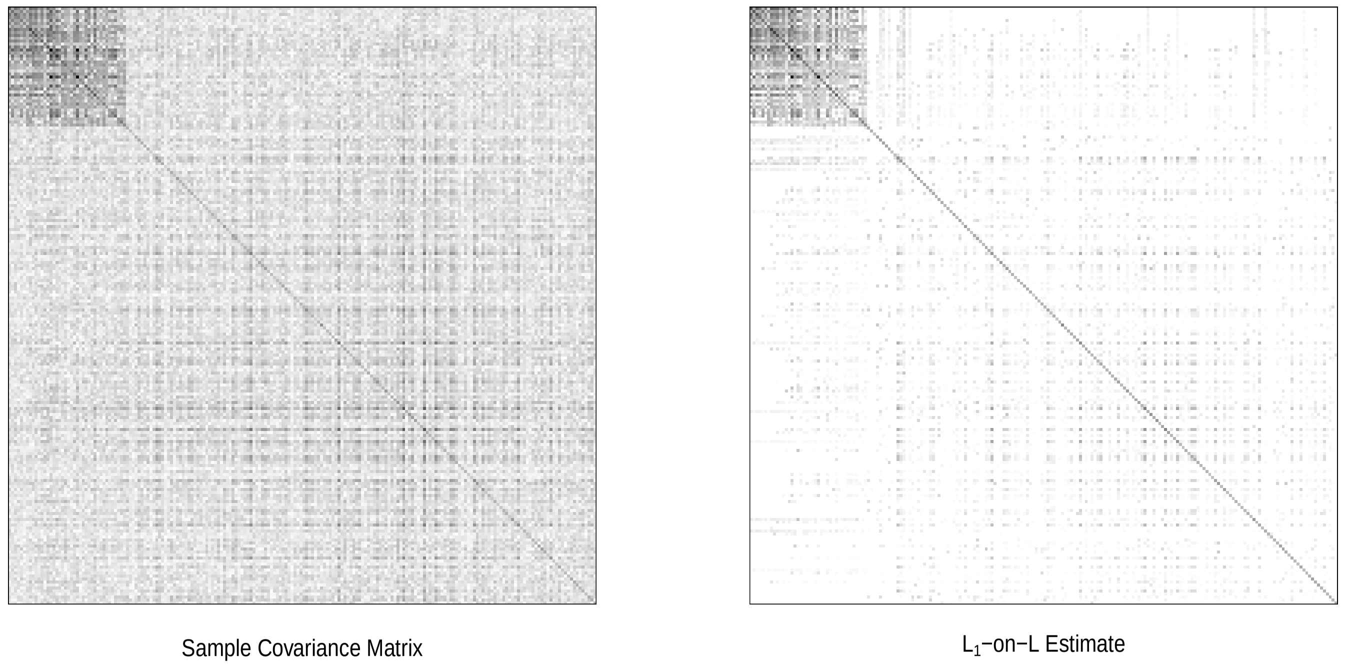

Figure 3.3: Heat maps of the estimated covariance matrices for blue-cell gene expression data. The left panel shows sample covariance matrix, and the right one shows the estimate from L_1-on-L. Absolute values of entries are used in replace of original entries. An entry of magnitude 1 or over is represented by a black square and an entry of magnitude 0 is represented by a white square.

The data is available at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser with the identification number GDS2519. In spite of a total size over 23,000 expression variables, we narrowed down the variable size into a smaller number. By measuring the extents of differential expression from two-sample t-tests, we sorted the expression variables in accordance with their significance levels and chose the first P ones of our interest.

Application of Principal Component Regression

We investigated the performance of various covariance matrix estimation methods by comparing the behaviors of their first three principal components (PCs) of the estimated covariance matrices. With the phenotype information as response, we fit the data of all 105 individuals to three logistic regression models. These models used only the 1st PC, the 1st and 2nd PCs, and all first three PCs as the covariates, respectively. To achieve systematical comparisons, we took the variable size P in 150, 200, 300, 500 and 1000. Residual deviances of these logistic regressions are reported in **Table 3.11**.

From **Table 3.11**, the results show that the estimate from L_1-on-L provides consistent improvement in terms of residual deviances compared with the sample covariance matrix. The gained improvement further increases as the the number of variables P increases. Note that the sample covariance matrix and the LW estimate share the same eigenvectors and thus their residual deviances are identical in **Table 3.11**. Interestingly, the residual deviances from the Threshold estimate are identical to the ones from the sample covariance matrix up to the second digital points. The

Using the first principal component					
P	Sample/LW	Threshold	Band-L	Bien's	L_1-on-L
150	65.64	65.64	96.88	67.65	62.06
200	67.43	67.43	96.17	76.00	66.22
300	75.27	75.27	98.29	83.58	73.95
500	78.34	78.34	105.45	-	74.75
1000	87.09	87.09	116.97	-	82.71

Using first two principal components					
P	Sample/LW	Threshold	Band-L	Bien's	L_1-on-L
150	65.51	65.51	54.27	65.35	61.54
200	65.93	65.93	54.39	66.11	62.59
300	73.29	73.29	97.48	73.51	69.73
500	78.34	78.34	105.35	-	74.64
1000	85.20	85.20	116.89	-	80.86

Using first three principal components					
P	Sample/LW	Threshold	Band-L	Bien's	L_1-on-L
150	48.84	48.84	54.22	53.50	49.69
200	46.61	46.61	53.70	49.81	47.58
300	57.99	57.99	97.13	61.45	55.75
500	69.04	69.04	104.62	-	60.63
1000	67.10	67.10	112.01	-	62.31

Table 3.11: Residual deviances for three sets of logistic regressions. The first one, the first two, and the first three principal components from covariance matrix estimates are used, respectively.

reason is that majority of entries in the sample covariance matrices are very small and close to each other. Consequently, only a very small portion of entries are set zeros by the Threshold method, and thus, the impact of thresholding is minimal. For the performance of the Band-L estimate, the pattern of deviance values is inconsistent in different conditions. Although there are two situations, $P = 150$ and 200 with first two PCs, in which the performance of Band-L is better than others, the commonly appeared large gaps between the deviances from using Band-L estimates and ones from using other methods suggest its inaccuracy in majority circumstances considered in this investigation. We would like to remark that the Bien's method requires heavy computation capacity in this study. Using the existing R-package (Bien and Tibshirani, 2011), it needs several days to finish the computation for $P = 300$ with a reasonably adequate tuning set using a 3.0 GHz CPU. Hence, we skipped the situations corresponding to bigger P's for the Bien's estimate. For the cases of $P \leq 300$, the comparisons of deviance values show that using Bien's method does not improve the covariance matrix estimation compared with the sample covariance matrix in this study.

Comparisons between ROC curves

We further explored the performance of these covariance matrix estimates in term of prediction accuracies. In this section, we only dealt with the case with $P = 150$. Training and validation groups were generated by randomly splitting the data into two sub-groups of roughly equal sizes. We applied different covariance matrix estimation methods to both the training and validation data, producing covariance matrix estimates, and further, obtained the first PC from these estimates. We used these first PC and the phenotype information from the training data to build a logistic regression model, and plugged the first PC from the estimated covariance matrices using validation data into the built logistic model.

Phenotype predictions were conducted for individuals in the validation group. By comparing the predicted results with real class labels, Receiver operating characteristic (ROC) curves are produced for different covariance matrix estimation methods. These ROC curves are shown in **Figure 3.4** as well as their Area Under the Curve (AUC) values. Such a prediction procedure was repeated for 50 times with different data splitting in determining the case and control groups. Bien's method is excluded in this comparison, because of its computation cost. With 50 values available, the distributions of AUC for different methods are drawn, as shown in **Figure 3.5**. The distribution for Band-L falls behind, while other three have large overlaps. The distribution of AUC corresponding to L_1-on-L appears to have higher median and smaller variability compared with the distributions for Threshold and LW method.

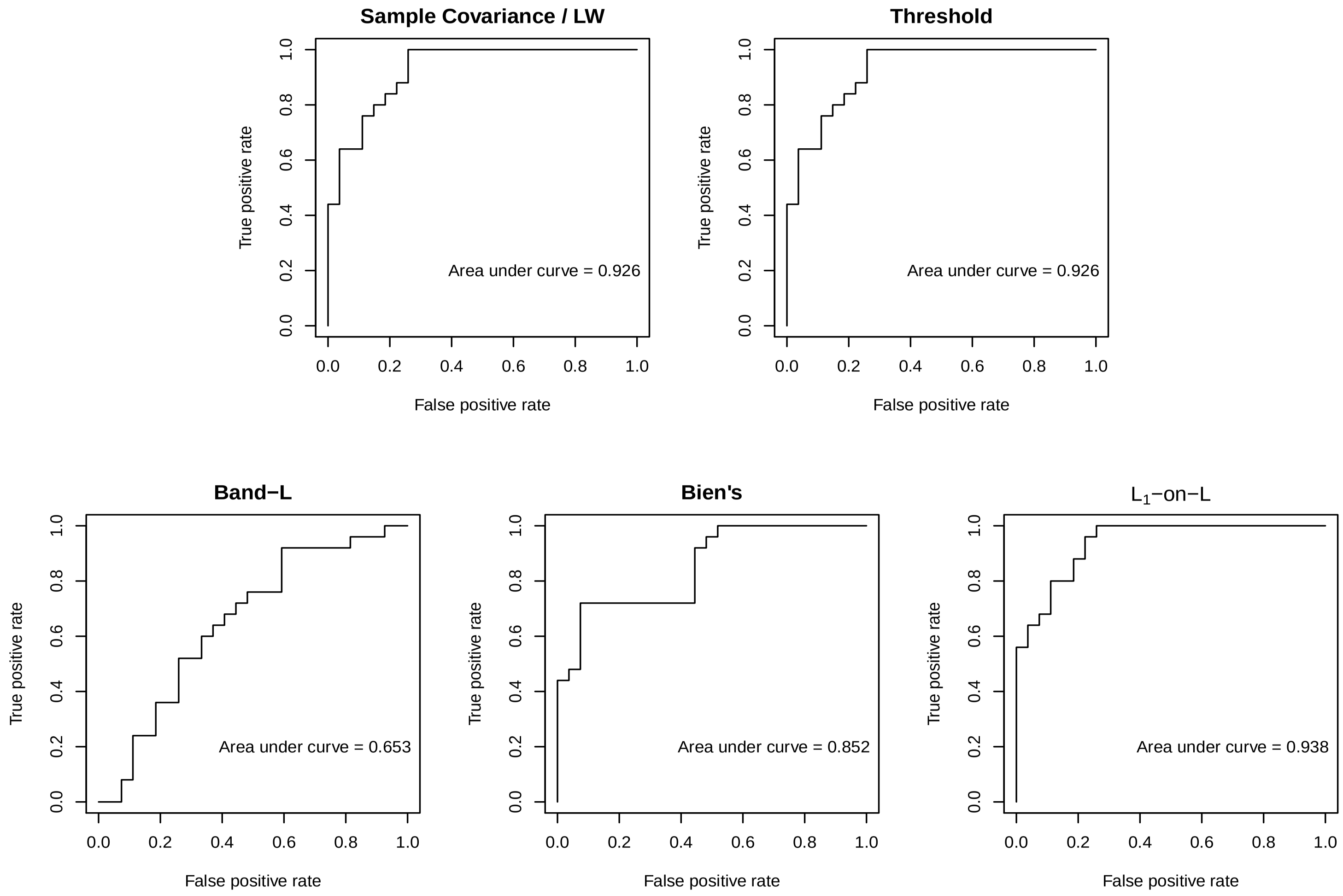

Figure 3.4: Receiver operating characteristic curves based on logistics regressions with the first principal component from various approaches.

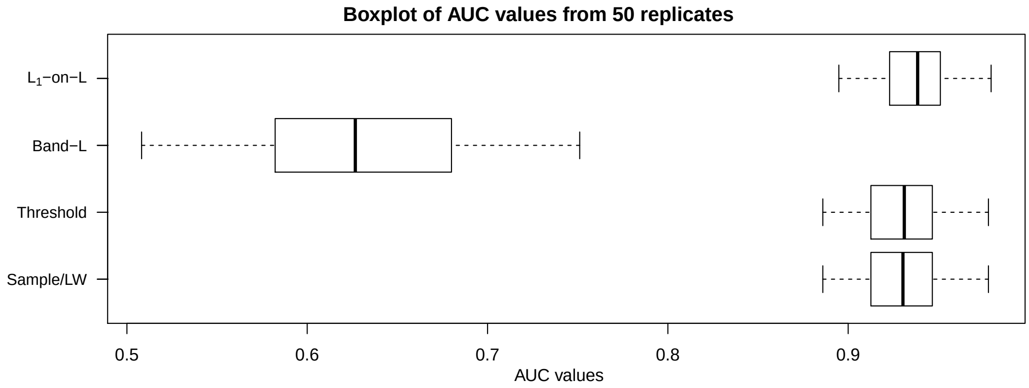

Figure 3.5: Distribution of AUC values from 50 replicates of classification based on logistics regressions with the first principal component from various approaches.

4

Order Issue of Cholesky-based Estimation with $\mathcal{L}_1$ Regularization

The assumption of the covariance matrix having a banded structure limits the usage of covariance matrix estimation through regularizing the Cholesky factor matrix. In Chapter 3, we consider covariance matrix estimation without assuming particular structures, and choose to impose $\mathcal{L}_1$ regularization on the Cholesky factor matrices. The corresponding estimates, are parsimonious, in the sense that much fewer parameters are involved to produce outcomes. Another essential advantage of such parsimonious estimation is that the resultant estimates are not sensitive towards the orders of variables. In Chapter 2, we have shown that the covariance matrix estimate from sequential regressions without any regularization is equal to the sample covariance matrix, which is order-invariant. Such a property does not remain if the banding regularization is imposed, since changing the order of variables permutes the rows and columns of the covariance matrix, resulting in a no-longer-banded Cholesky factor matrix even if the original one is banded. However, $\mathcal{L}_1$ regularization is not that sensitive to the change of variable orders. The main role of $\mathcal{L}_1$ penalty on regression coefficients is to shrink unimportant ones, leaving the significant ones less impacted. Such a procedure is not vulnerable towards the order of sequential regressions. As repeated by Pourahmadi (2011), the very important requirement of the modified Cholesky decomposition is that a specific order of variables should be provided beforehand. Often, such order information is not available in practice, or can not be rationally assumed, which constrains the application of covariance matrix estimation through regularizing Cholesky factor matrices. The idea of parsimonious estimation from $\mathcal{L}_1$ regularization on Cholesky factor matrices casts a light of providing order-invariant estimates associated with the modified Cholesky decomposition. We use an illustrative example to begin our explanation.

4.1 An Illustrative Example

Although we point out that the property of estimates from $\mathcal{L}_1$ regularizing Cholesky factor matrices is parsimony, rather than sparsity, in this illustrative example, we present a sparse case so that the message is more straightforward. An 8 × 8 sparse covariance matrix shown below is used in this illustration.

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0 \\ 0 & 1 & 0 & 0.9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.9 \\ 0 & 0.9 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.9 & 1 & 0 & 0 \\ 0.9 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

As shown on the next page, the true covariance matrices with respect to different variable orders are shown on the left, while their Cholesky factor matrices L's are shown on the right. If the structure of L can be successfully identified, the true covariance structure would be restored. In the case, the structure of the Cholesky factor matrices remains regardless under possible orders of variables. Because of that, in terms of identifying Cholesky factor matrix L using $\mathcal{L}_1$ penalty, there appears no substantial differences between processes of achieving Cholesky factor matrices under different variable orders in this case.

Covariance Matrix Σ Cholesky Factor Matrix L

$$\begin{array}{c} x_1\\ x_2\\ x_3\\ x_4\\ x_5\\ x_6\\ x_7\\ x_8 \end{array}\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0\\ 0 & 1 & 0 & 0.9 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.9\\ 0 & 0.9 & 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0.9 & 0 & 0\\ 0 & 0 & 0 & 0 & 0.9 & 1 & 0 & 0\\ 0.9 & 0 & 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \Leftarrow \begin{array}{c} x_1\\ x_2\\ x_3\\ x_4\\ x_5\\ x_6\\ x_7\\ x_8 \end{array}\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0.9 & 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0.9 & 1 & 0 & 0\\ 0.9 & 0 & 0 & 0 & 0 & 0 & 1 & 0\\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$\Updownarrow$ Permutation

$$\begin{array}{c} x_3\\ x_4\\ x_1\\ x_6\\ x_2\\ x_8\\ x_5\\ x_7 \end{array}\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0.9 & 0 & 0\\ 0 & 1 & 0 & 0 & 0.9 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.9\\ 0 & 0 & 0 & 1 & 0 & 0 & 0.9 & 0\\ 0 & 0.9 & 0 & 0 & 1 & 0 & 0 & 0\\ 0.9 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0.9 & 0 & 0 & 1 & 0\\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \Leftarrow \begin{array}{c} x_3\\ x_4\\ x_1\\ x_6\\ x_2\\ x_8\\ x_5\\ x_7 \end{array}\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0.9 & 0 & 0 & 1 & 0 & 0 & 0\\ 0.9 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0.9 & 0 & 0 & 1 & 0\\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$\Updownarrow$ Permutation

$$\begin{array}{c} x_6\\ x_5\\ x_4\\ x_1\\ x_7\\ x_3\\ x_8\\ x_2 \end{array}\begin{pmatrix} 1 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0\\ 0.9 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0.9\\ 0 & 0 & 0 & 1 & 0.9 & 0 & 0 & 0\\ 0 & 0 & 0 & 0.9 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 0.9 & 0\\ 0 & 0 & 0 & 0 & 0 & 0.9 & 1 & 0\\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \Leftarrow \begin{array}{c} x_6\\ x_5\\ x_4\\ x_1\\ x_7\\ x_3\\ x_8\\ x_2 \end{array}\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0.9 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0.9 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0.9 & 1 & 0\\ 0 & 0 & 0.9 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Given this 8 × 8 covariance matrix under the original order of variables, $n = 20$ independent observations from a multivariate normal distribution were generated and centered. The procedure was repeated for multiple times. To show the evolvement, we first take one replicate of simulation for example. The sample covariance matrix in this replicate is calculated as follows:

$$S = \begin{pmatrix} 1.095 & & & & & & & \\ -0.824 & 1.268 & & & & & & \\ 0.155 & 0.151 & 0.878 & & & & & \\ -0.712 & 1.121 & 0.129 & 1.181 & & & & \\ 0.184 & -0.261 & -0.223 & -0.234 & 0.868 & & & \\ 0.250 & -0.217 & -0.226 & -0.126 & 0.772 & 0.933 & & \\ 1.127 & -0.891 & 0.270 & -0.802 & 0.315 & 0.331 & 1.359 & \\ 0.112 & 0.235 & 0.764 & 0.224 & -0.220 & -0.170 & 0.212 & 0.804 \end{pmatrix}.$$

Based on the L_1-on-L approach described in Chapter 3, the estimate $\hat{\Sigma}$ is produced and presented below,

$$\hat{\Sigma} = \begin{pmatrix} 1.153 & & & & & & & \\ -0.562 & 1.038 & & & & & & \\ 0 & 0 & 0.924 & & & & & \\ -0.269 & 0.569 & 0 & 0.668 & & & & \\ 0 & 0 & 0 & 0 & 0.914 & & & \\ 0 & 0 & 0 & 0 & 0.507 & 0.644 & & \\ 0.881 & -0.429 & 0 & -0.206 & 0 & 0 & 0.964 & \\ 0 & 0 & 0.499 & 0 & 0 & 0 & 0 & 0.516 \end{pmatrix}.$$

Randomly permuting the order of variables will not change the structure of the covariance matrix, while the corresponding Cholesky factor matrices could be different. Thus, the estimates for the covariance matrix based on different order of variables are not necessarily the same. This example is to illustrate that, estimates corresponding to various orders of variables are not far away from each other.

To make this illustration complete, we applied L_1-on-L to get all covariance matrix estimates for $8! = 40320$ possible permutations. The average of all these 40320 estimated covariance matrices based on different orders of variables is shown as $\hat{\Sigma}$ as below. We also calculated the standard deviations for all covariance entries, and presents them in parentheses beside the corresponding entries. The standard deviations are small, indicating the estimated structure is stable.

$$\hat{\Sigma} = \begin{pmatrix} 0.864(0.199) & & & & & & & \\ -0.393(0.118) & 0.998(0.230) & & & & & & \\ 0.001(0.004) & 0.001(0.005) & 0.742(0.179) & & & & & \\ -0.331(0.076) & 0.729(0.145) & 0.000(0.002) & 0.941(0.212) & & & & \\ 0.001(0.004) & -0.001(0.003) & 0.000(0.000) & 0.000(0.002) & 0.756(0.157) & & & \\ 0.002(0.007) & -0.001(0.005) & 0.000(0.000) & -0.001(0.004) & 0.506(0.003) & 0.812(0.168) & & \\ 0.735(0.145) & -0.445(0.132) & 0.007(0.016) & 0.383(0.108) & 0.004(0.009) & 0.006(0.143) & 1.071(0.245) & \\ 0.001(0.003) & 0.003(0.010) & 0.497(0.005) & 0.001(0.005) & 0.000(0.000) & 0.000(0.000) & 0.006(0.011) & 0.680(0.163) \end{pmatrix}$$

To compare the variability of matrix estimates across permutations within individual replicates to the variability across different replicates, we repeated the previous simulation for 200 times. **Table 4.1** lists the summaries of accuracy measures and their variabilities for 200 replicates. The L_1, L_2 and F norms of the difference matrix between the estimated and the true covariance matrix are used as well as the entropy loss. To better explain the result, we take the cell 1.394(0.134) in the line of "replicate 1" for instance. This cell tells that the averaged L_1 norm of 40320 difference matrices between the estimated and true covariance matrices in replicate 1 is 1.394, and the standard deviation of these 40320 values is 0.134.

If we use 1.394 as the L_1 norm measure for L_1-on-L in replicate 1, then 200 such averages, which are $1.394, \cdots, 1.154$, have a mean value 1.158 and a standard deviation 0.240. This standard deviation 0.240 is larger than the standard deviation of corresponding measure for estimates across permutations in replicate 1, as well as in other replicates. Such mean and standard deviation values are listed in the line of "mean (s.d.)". The claim that the variability across simulation replicates is greater than the variability across permutations within individual replicates, still holds when the measure is L_2 norm, F norm or entropy loss.

	L_1 norm	L_2 norm	F norm	entropy loss
replicate 1	1.394 (0.134)	1.022 (0.075)	1.656 (0.098)	3.062 (0.305)
$\vdots$				
replicate 200	1.154 (0.062)	1.085 (0.055)	1.818 (0.076)	1.944 (0.392)
mean (s.d.)	1.158 (0.240)	1.016 (0.226)	1.483 (0.325)	2.075 (0.834)

Table 4.1: Performance of L_1-on-L estimates across permutations in individual replicates and across 200 simulation replicates in the illustrative example.

4.2 Refinement of Covariance Matrix Estimate from $\mathcal{L}_1$ Regularization

In many cases like the illustrative example, we find that the variability of performance across simulation replicates is greater than the variability of performance across permutations within individual replicates. Those cases give us an expression that the L_1-on-L estimate described in Chapter 3 is not sensitive towards the order of variables. One understanding about this is: the application of $\mathcal{L}_1$ regularization aims to improve the estimate from the the sample covariance matrix towards the true covariance matrix; marching along this track with both ends are order-invariant matrices, estimation using different orders of variables may implement the improvement differently, but is not expected to deviate far away from the track.

Note that the prerequisite of applying the modified Cholesky decomposition technique is the availability of the order of variables. The property that the estimate resulted from $\mathcal{L}_1$ regularization is not sensitive towards the order serves well to eliminate this prerequisite. It motivates us to propose an order-invariant covariance matrix estimate from refining estimates under random permutations of the order of variables. Such refinement from individual estimates can be analog to building a random forest (Breiman, 2001) from individual decision trees. In random forest theory, an individual tree is a classifier which is able to fulfill the job, just as an individual estimate from (3.3) and (3.4) is a desirable estimate for the covariance matrix. While one tree may overfit the data, a particular-order-based estimate may be more preferred under certain order. Combining representative estimates corresponding to different orders is like building a random forest so that the performance becomes more stable.

As shown in Chapter 2 Section (2.1.2), when the sample covariance matrix S is non-singular, the covariance matrix estimate from sequential regressions is exactly equal to S. This treatment is equivalent to setting tuning parameter η zero while applying L_1-on-L. A fairly chosen η will probably lead to a better estimate. Further refining estimates in accordance with different orders of variables is expected to produce a stable estimate. We illustrate our consideration in **Table 4.2**.

In more detail, we define a permutation mapping $\pi : \{1, \ldots, p\} \rightarrow \{1, \ldots, p\}$,

		Covariance Matrix Σ			
$\swarrow$		$\downarrow$			$\searrow$
Covariance Matrix Σ	permutation	Covariance Matrix Σ	$\cdots$	permutation	Covariance Matrix Σ
under Variable Order π_1	$\Longleftrightarrow$	under variable order π_2		$\Longleftrightarrow$	under variable order π_K
$\Downarrow$		$\Downarrow$			$\Downarrow$
Cholesky Factor L_{π_1}		Cholesky Factor L_{π_2}			Cholesky Factor L_{π_K}
$\Downarrow$		$\Downarrow$			$\Downarrow$
Estimated $\hat{L}_{\pi_1}$		Estimated $\hat{L}_{\pi_2}$			Estimated $\hat{L}_{\pi_K}$
$\Downarrow$		$\Downarrow$			$\Downarrow$
Estimate $\hat{A}_{\pi_1}$ for Permuted		Estimate $\hat{A}_{\pi_2}$ for Permuted	$\cdots$		Estimate $\hat{A}_{\pi_K}$ for Permuted
Covariance Matrix		Covariance Matrix			Covariance Matrix
$\Downarrow$		$\Downarrow$			$\Downarrow$
Estimate $\hat{\Sigma}_{\pi_1}$		Estimate $\hat{\Sigma}_{\pi_2}$	$\cdots$		Estimate $\hat{\Sigma}_{\pi_K}$
Restored in Original Order		Restored in Original Order			Restored in Original Order
$\searrow$		$\downarrow$			$\swarrow$
		Refined Estimate $\hat{\Sigma}$			

Table 4.2: Work flow of generating the refined estimate from individual estimates under permutations.

which represents a rearrangement of the orders, $1, 2, \ldots, p$,

$$\big(\pi(1),\ \ldots,\ \pi(p)\big).$$

If we use e_t to denote a p-dimensional vector with only the t-th element one and all others zeros, we can further define a permutation matrix

$$\boldsymbol{P}_\pi = \big(e_{\pi(1)},\ \ldots,\ e_{\pi(p)}\big).$$

Thus, the columns of data matrix $\mathscr{X}$ could be permuted by right multiplying $\boldsymbol{P}_\pi$ as follows:

$$\mathscr{X}_\pi = \mathscr{X}\boldsymbol{P}_\pi = (\boldsymbol{X}_{\pi(1)}, \ldots, \boldsymbol{X}_{\pi(p)}),$$

where $\boldsymbol{X}_{\pi(t)}$ is the $\pi(t)$-th column of $\mathscr{X}$.

Under the permutation π, we simply replace the $\boldsymbol{x}_t$ with $\boldsymbol{x}_{\pi(t)}$ in (3.3) and (3.4) for $t = 1, \ldots, p$, and get $\hat{A}_\pi$ as the covariance matrix estimate under the order π. We further restore the estimated covariance matrix according to the original order by applying the inverse of the mapping π as follows:

$$\hat{\Sigma}_\pi = \boldsymbol{P}_{\boldsymbol{\pi}}\ \hat{A}_\pi\ \big(\boldsymbol{P}_\pi\big)^{-1}.$$

By incorporating possible permutations π's, one can have a pool of covariance matrix estimates. Reasonably refining them would lead to an order-invariant estimate, and such an refinement could be taking average of the estimates over all permutations. In practice, a modest number of permutations is able to roughly fulfill our purpose of pursuing the order-invariant estimate. Therefore, we randomly select a moderate size subset of permutations, denoted as $\mathcal{C} = \{\pi_1, \ldots, \pi_K\}$. and the refined covariance matrix estimate $\hat{\Sigma}_*$ is obtained by

$$\hat{\Sigma}_* = \frac{1}{K} \sum_{\pi_k \in \mathcal{C}} \hat{\Sigma}_{\pi_k}. \tag{4.1}$$

From the finite population sampling survey theory (see Cochran, 1977), the selection of permutation set $\mathcal{C}$ is not essential if we have a reasonable size K. Choosing a larger K would further reduce the variability of the refined estimate $\hat{\Sigma}_*$, while a modest number is capable to lead to stable results.

4.3 Previous Simulation Revisit

If the knowledge of variable orders is not available, we refer to refined L_1-on-L to produce the covariance matrix estimate. To illustrate the performance of the refinement approach, we carried out the procedure to all previous simulations in Chapter 3. Following the exact same settings including the covariance matrix assumption, the measure criteria of accuracy and the method in the tuning procedure, we investigated the performance of estimates from our refined L_1-on-L. We repeatedly randomly permuted the variable orders while the true arrangement information was not taken into consideration. A refined L_1-on-L estimate followed from this group of estimates. The main concern was the size of this group, and it was from a tradeoff between efficacy and computation consumption. We took two values for the group size, 30 and 100, so the corresponding refined 30-L_1-on-L and refined 100-L_1-on-L estimates were generated. The results of loss measures are listed in **Table 4.3** and **Table 4.4**, and the summary for condition numbers is reported in **Table 4.5**. The same as in the presentations in Chapter 3, average values from 200 replicates as well as the standard errors (in parentheses) are reported. To emphasize the consistency of the refined L_1-on-L with the original approach when the order information is available, we combine the original results corresponding to L_1-on-L estimate in both **Table 4.3** and **Table 4.4**. In **Table 4.5**, besides the original results corresponding to L_1-on-L estimate, the true condition numbers from the simulated covariance matrices are listed as references. Both estimates from refined 30-L_1-on-L and refined 100-L_1-on-L catch the true covariance structure as nicely as the one from original L_1-on-L, which implies 30 is a reasonable number of the group size in producing the order invariant estimate.

Figure 4.1 corresponds to **Figure 3.1**. In the exact same replicate of simulation study in Scenario 3, the heat map for the 30-L_1-on-L covariance matrix estimate is added, while the one for the true covariance matrix as reference is removed. The appearance of 30-L_1-on-L estimate is very alike to the image of L_1-on-L estimate. Similar claim stands for **Figure 4.2** in comparison to **Figure 3.2**.

In summary, we conclude that for all simulation scenarios, the performance of refined L_1-on-L estimate, despite the permutation size 30 or 100, is consistently close to the original L_1-on-L estimate with prior knowledge of the order of variables.

4.4 Real Data Analysis

4.4.1 Parkinson Data Revisit

We revisit the Parkinson data analysis we presented in Chapter 3. The purpose of this revisit is, through comparisons with other methods, to show that the refined estimate behavior similarly to the original estimate that uses the order information of variables. For certain cases, there might be a gap, and the refined estimate might

$\mathbf{\Sigma}_1$	$p = 30$			$p = 50$			$p = 100$		
	original	refined	refined	original	refined	refined	original	refined	refined
	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L
L_1 norm	1.40 (0.01)	1.38 (0.01)	1.38 (0.01)	1.50 (0.01)	1.48 (0.01)	1.48 (0.01)	1.61 (0.02)	1.57 (0.01)	1.56 (0.01)
L_2 norm	0.90 (0.01)	0.89 (0.01)	0.89 (0.01)	0.95 (0.00)	0.94 (0.00)	0.94 (0.00)	1.00 (0.01)	0.99 (0.00)	0.99 (0.00)
F norm	2.30 (0.01)	2.25 (0.01)	2.25 (0.01)	3.13 (0.01)	3.07 (0.01)	3.07 (0.01)	4.73 (0.02)	4.66 (0.01)	4.66 (0.01)
entropy	3.90 (0.04)	4.02 (0.04)	4.00 (0.04)	7.98 (0.06)	8.15 (0.05)	8.15 (0.05)	20.06 (0.17)	20.28 (0.08)	20.26 (0.08)
$\mathbf{\Sigma}_2$	$p = 30$			$p = 50$			$p = 100$		
	original	refined	refined	original	refined	refined	original	refined	refined
	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L
L_1 norm	1.45 (0.01)	1.37 (0.01)	1.37 (0.01)	1.59 (0.01)	1.47 (0.01)	1.46 (0.01)	1.78 (0.04)	1.57 (0.01)	1.56 (0.01)
L_2 norm	0.91 (0.01)	0.88 (0.01)	0.88 (0.01)	0.96 (0.00)	0.94 (0.00)	0.94 (0.00)	1.01 (0.01)	1.00 (0.00)	1.00 (0.00)
F norm	2.35 (0.01)	2.25 (0.01)	2.25 (0.01)	3.21 (0.01)	3.08 (0.01)	3.08 (0.01)	4.82 (0.02)	4.67 (0.01)	4.67 (0.01)
entropy	4.32 (0.04)	4.05 (0.04)	4.04 (0.04)	8.70 (0.06)	8.20 (0.05)	8.20 (0.05)	21.19 (0.18)	20.35 (0.09)	20.34 (0.09)
$\mathbf{\Sigma}_3$	$p = 30$			$p = 50$			$p = 100$		
	original	refined	refined	original	refined	refined	original	refined	refined
	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L
L_1 norm	1.30 (0.01)	1.26 (0.01)	1.27 (0.01)	1.41 (0.01)	1.35 (0.01)	1.35 (0.01)	1.59 (0.04)	1.47 (0.01)	1.46 (0.01)
L_2 norm	0.85 (0.01)	0.84 (0.01)	0.84 (0.01)	0.90 (0.00)	0.89 (0.00)	0.89 (0.00)	0.95 (0.01)	0.93 (0.00)	0.94 (0.00)
F norm	2.18 (0.01)	2.12 (0.01)	2.12 (0.01)	2.95 (0.01)	2.88 (0.01)	2.88 (0.01)	4.44 (0.02)	4.34 (0.01)	4.34 (0.01)
entropy	3.09 (0.03)	3.04 (0.03)	3.03 (0.03)	6.02 (0.05)	5.93 (0.04)	5.94 (0.04)	14.44 (0.18)	14.38 (0.07)	14.37 (0.07)

Table 4.3: Comparisons of performance between the original L_1-on-L estimate and the refined ones based on 30 or 100 permutations in simulation scenarios with covariance matrix Σ_1, Σ_2 and Σ_3. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

$\mathbf{\Sigma}_4$	$p = 30$			$p = 50$			$p = 100$		
	original	refined	refined	original	refined	refined	original	refined	refined
	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L
L_1 norm	2.66 (0.05)	2.54 (0.05)	2.52 (0.05)	4.67 (0.08)	4.53 (0.08)	4.50 (0.08)	9.96 (0.31)	9.78 (0.15)	9.74 (0.15)
L_2 norm	2.05 (0.04)	2.04 (0.05)	2.02 (0.05)	3.56 (0.07)	3.58 (0.08)	3.56 (0.08)	7.03 (0.25)	7.43 (0.15)	7.37 (0.14)
F norm	2.60 (0.04)	2.52 (0.04)	2.51 (0.04)	4.33 (0.07)	4.21 (0.08)	4.20 (0.08)	8.66 (0.28)	8.50 (0.15)	8.45 (0.15)
entropy	2.69 (0.04)	1.99 (0.03)	1.99 (0.03)	6.17 (0.18)	4.28 (0.06)	4.27 (0.06)	30.40 (3.13)	15.02 (0.37)	15.16 (0.37)
$\mathbf{\Sigma}_5$	$p = 30$			$p = 50$			$p = 100$		
	original	refined	refined	original	refined	refined	original	refined	refined
	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L	L_1-on-L	30-L_1-on-L	100-L_1-on-L
L_1 norm	2.58 (0.05)	2.41 (0.05)	2.41 (0.05)	4.65 (0.08)	4.41 (0.09)	4.41 (0.08)	10.24 (0.30)	9.85 (0.15)	9.81 (0.15)
L_2 norm	2.02 (0.04)	1.92 (0.05)	1.91 (0.05)	3.60 (0.08)	3.47 (0.08)	3.49 (0.08)	7.61 (0.28)	7.49 (0.14)	7.41 (0.14)
F norm	2.54 (0.04)	2.41 (0.04)	2.41 (0.04)	4.31 (0.07)	4.12 (0.08)	4.13 (0.08)	9.00 (0.29)	8.58 (0.14)	8.52 (0.14)
entropy	2.56 (0.03)	1.95 (0.03)	1.94 (0.03)	5.52 (0.07)	4.27 (0.06)	4.24 (0.05)	22.96 (1.73)	15.52 (0.37)	15.65 (0.37)

Table 4.4: Comparisons of performance between the original L_1-on-L estimate and the refined ones based on 30 or 100 permutations in simulation scenarios with covariance matrix Σ_4 and Σ_5. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

Scenario	p	Σ	original L_1-on-L	refined 30-L_1-on-L	refined 100-L_1-on-L
	30	8.80	6.17 (0.08)	5.26 (0.05)	5.26 (0.05)
$\mathbf{\Sigma}_1$	50	8.92	5.76 (0.06)	5.00 (0.04)	4.98 (0.04)
	100	8.98	5.16 (0.10)	4.56 (0.03)	4.56 (0.03)
	30	8.80	5.74 (0.06)	5.19 (0.05)	5.18 (0.05)
$\mathbf{\Sigma}_2$	50	8.92	5.57 (0.06)	4.98 (0.04)	4.97 (0.04)
	100	8.98	5.22 (0.10)	4.59 (0.04)	4.58 (0.04)
	30	5.51	4.76 (0.05)	4.35 (0.04)	4.35 (0.04)
$\mathbf{\Sigma}_3$	50	5.51	4.77 (0.06)	4.32 (0.04)	4.30 (0.04)
	100	5.51	4.62 (0.10)	4.09 (0.03)	4.08 (0.03)
	30	25	27.64 (0.59)	13.63 (0.17)	13.60 (0.16)
$\mathbf{\Sigma}_4$	50	41	66.82 (1.71)	22.30 (0.27)	22.08 (0.25)
	100	81	435 (61)	50.86 (0.48)	51.18 (0.47)
	30	25	28.42 (0.62)	13.86 (0.16)	13.80 (0.16)
$\mathbf{\Sigma}_5$	50	41	66.39 (1.57)	22.63 (0.26)	22.33 (0.25)
	100	81	259 (17.44)	50.93 (0.46)	51.65 (0.46)

Table 4.5: Summary of condition numbers ($\lambda_{\max}/\lambda_{\min}$) from the original L_1-on-L estimate and the refined ones based on 30 or 100 permutations in all five simulation scenarios The values in the column for the true Σ serve for reference. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

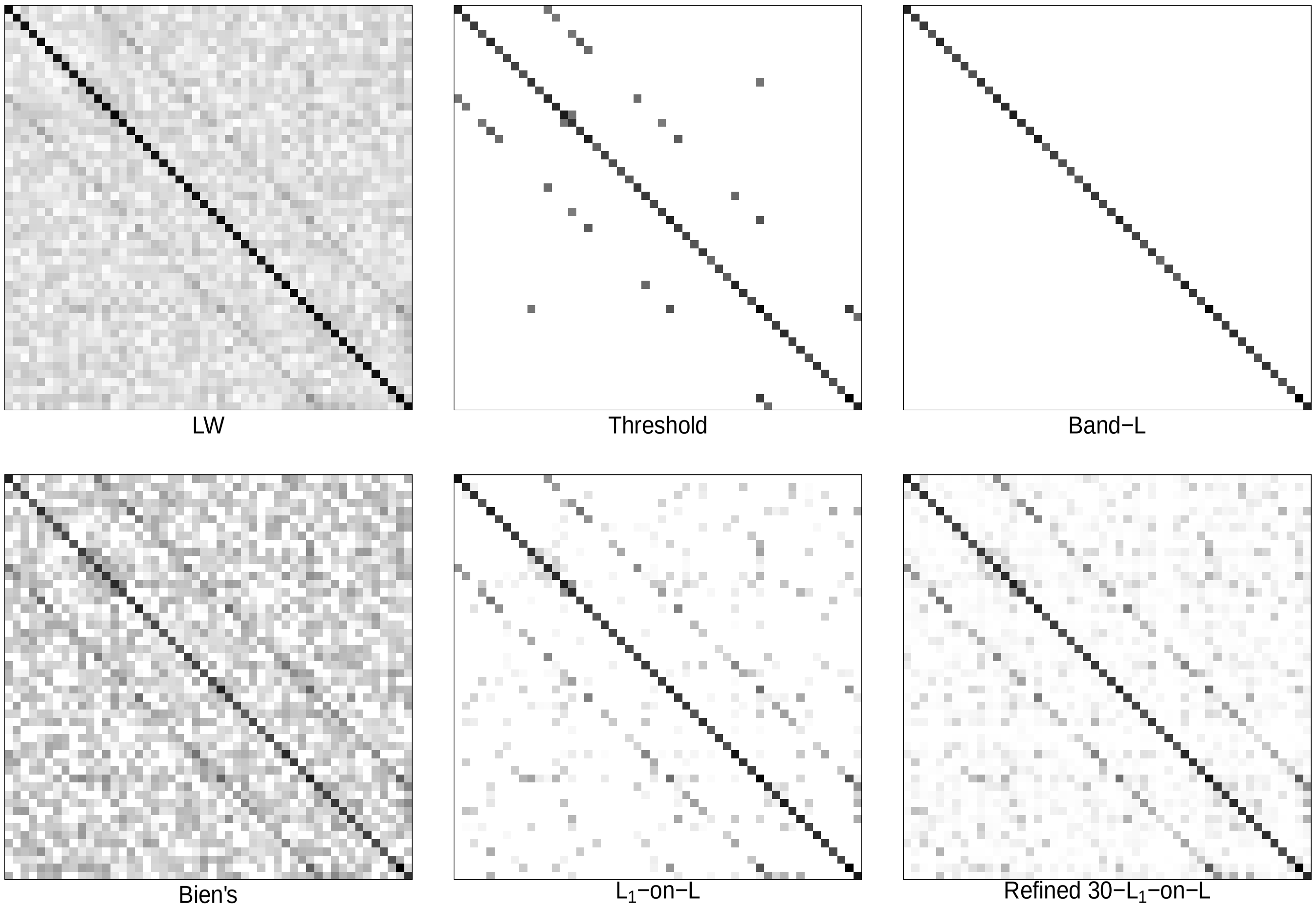

Figure 4.1: Heat maps of the true covariance matrix and various estimates from one replicate of simulations in Scenario 3 (Loose Banded Structure). Absolute values of entries are used to replace original entries. An entry of magnitude 1 or over is represented by a black square and an entry of magnitude 0 is represented by a white square.

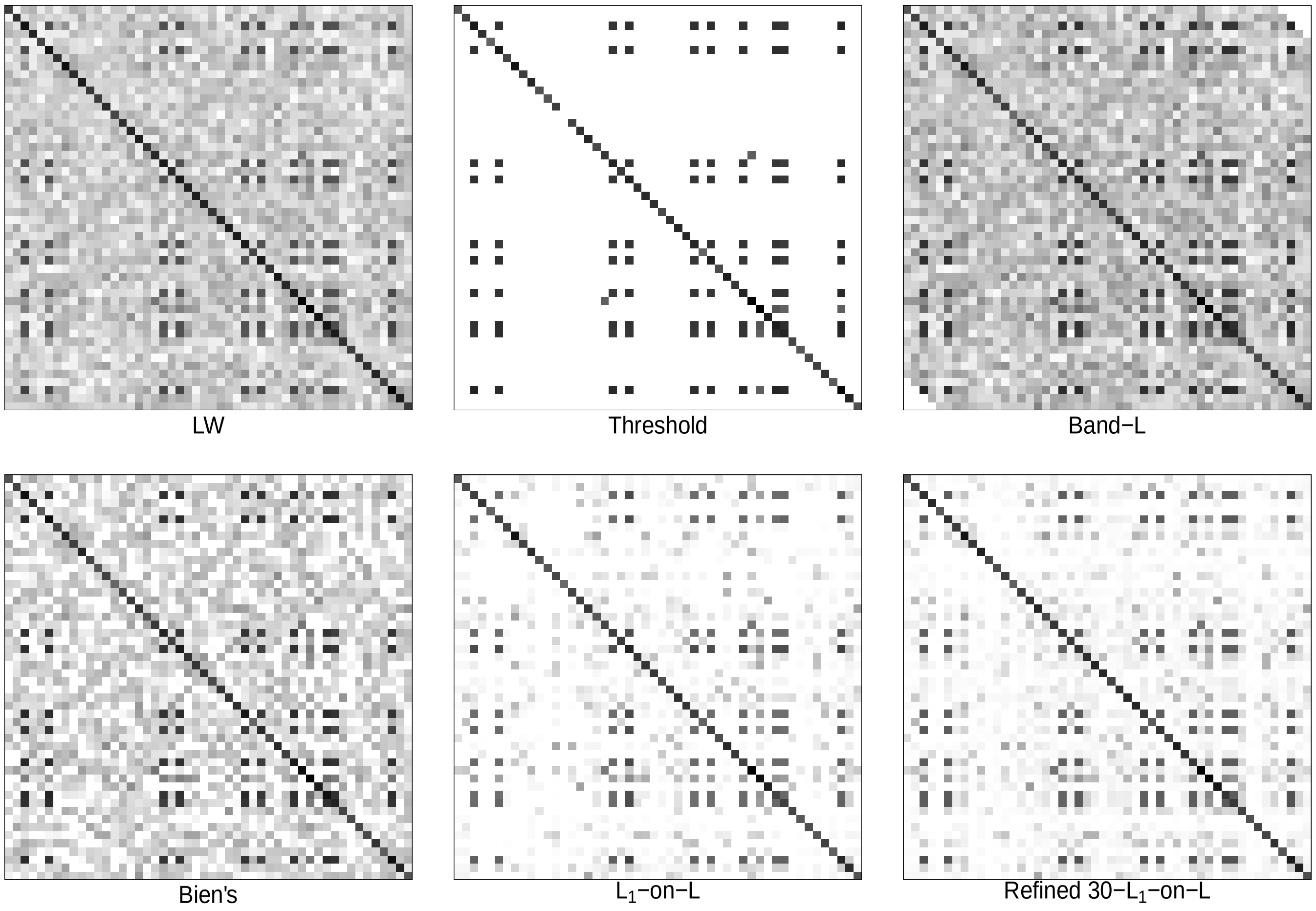

Figure 4.2: Heat maps of the true covariance matrix and various estimates from one replicate of simulations in Scenario 5 (Permuted Block Diagonal Structure). Absolute values of entries are used to replace original entries. An entry of magnitude 1 or over is represented by a black square and an entry of magnitude 0 is represented by a white square.

be less capable. However, the performance of the refined estimate is still expected to be competent relatively to the estimates from other methods.

Recall the Parkinson data contain the blood samples of 50 parkinson disease patients, and 55 age-matched health participants. By measuring the extents of differential expression from two-sample t-tests, we sorted the overall 23,000 expression variables in accordance with their significance levels and chose the first P ones as our focus. With the phenotype information as response, we fit the data to three logistic regression models. These models used only the 1st PC, the 1st and 2nd PCs, and all first three PCs as the covariates, respectively.

The results for refined L_1-on-L based on a permutation set with size $K = 30$ are attached to the previous outcomes, as shown in **Table 4.6**. The values match the previous expectation. In general, the refined 30-L_1-on-L estimate brings improvement to logistic regression fittings compared with the sample covariance matrix, and the improvement is systematical regardless the choices of matrix dimension P and how many PCs involved. The improvement from 30-L_1-on-L estimate, is systematically not as good as the one from original L_1-on-L estimate, but the gap in terms of differences of deviance values is quite small.

To show that the refined 30-L_1-on-L estimate is not sensitive to the choice of the permutation set, we generated 200 estimates by using 200 different sets of $K = 30$ randomly selected permutations. We computed the individual deviance values of logistic regressions correspondingly. Because of the 200 different permutation sets, we are not only able to show the performance of the refined 30-L_1-on-L estimate using the averages from the corresponding 200 values, but also able to present the values

Using the first principal component

p	Sample/LW	Threshold	Band-L	Bien's	L_1-on-L	30-L_1-on-L
150	65.64	65.64	96.88	67.65	62.06	65.15 (0.01)
200	67.43	67.43	96.17	76.00	66.22	69.17 (0.03)
300	75.27	75.27	98.29	83.58	73.95	76.91 (0.04)
500	78.34	78.34	105.45	-	74.75	78.05 (0.02)
1000	87.09	87.09	116.97	-	82.71	86.12 (0.02)

Using first two principal components

p	Sample/LW	Threshold	Band-L	Bien's	L_1-on-L	30-L_1-on-L
150	65.51	65.51	54.27	65.35	61.54	64.40 (0.02)
200	65.93	65.93	54.39	66.11	62.59	64.76 (0.01)
300	73.29	73.29	97.48	73.51	69.73	72.46 (0.01)
500	78.34	78.34	105.35	-	74.64	77.95 (0.02)
1000	85.20	85.20	116.89	-	80.86	83.96 (0.02)

Using first three principal components

p	Sample/LW	Threshold	Band-L	Bien's	L_1-on-L	30-L_1-on-L
150	48.84	48.84	54.22	53.50	49.69	48.67 (0.02)
200	46.61	46.61	53.70	49.81	47.58	46.07 (0.03)
300	57.99	57.99	97.13	61.45	55.75	56.16 (0.08)
500	69.04	69.04	104.62	-	60.63	63.25 (0.30)
1000	67.10	67.10	112.01	-	62.31	64.26 (0.14)

Table 4.6: Residual deviances for three sets of logistic regressions. The first one, the first two, and the first three principal components from covariance matrix estimates are used, respectively.

of standard errors, which are listed in the parentheses beside the average values in **Table 4.6**. The negligible quantities of standard errors confirm the minimal impact of the selection of permutation set.

Comparisons between ROC curves

Similarly, we also revisited the prediction analysis as in Chapter 3 corresponding to different estimates, and the case is $P = 150$ using only the first PC. ROC curves for different approaches including the refined 30-L_1-on-L are shown in **Figure 4.4** as well as their AUC values. The same procedure was repeated for 50 times with different ways of data splitting for various methods except Bien's method for the computing consideration. The distributions of AUC values are shown in **Figure 4.3**. The performance for the refined 30-L_1-on-L estimate is almost identical to the original L_1-on-L estimate with order information of variables.

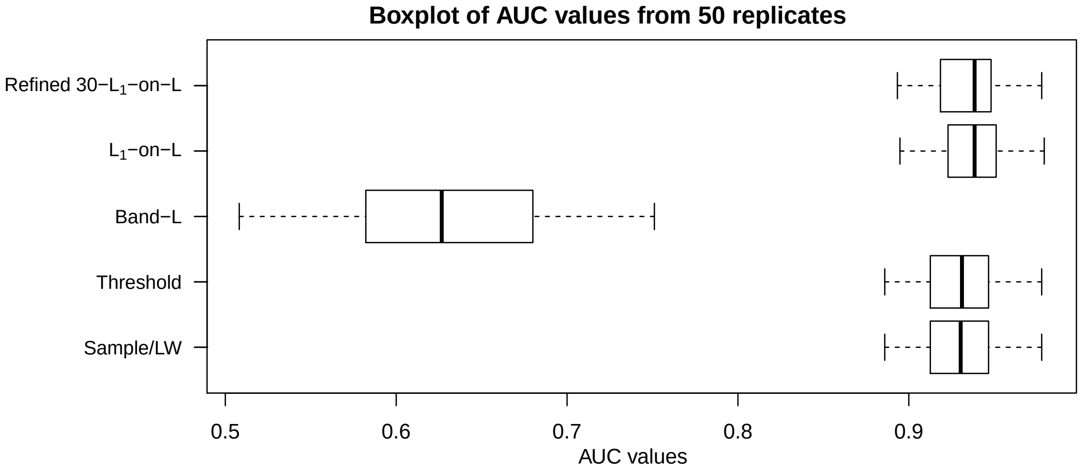

Figure 4.3: Distribution of AUC values from 50 replicates of classification based on logistics regressions with the first principal component from various approaches.

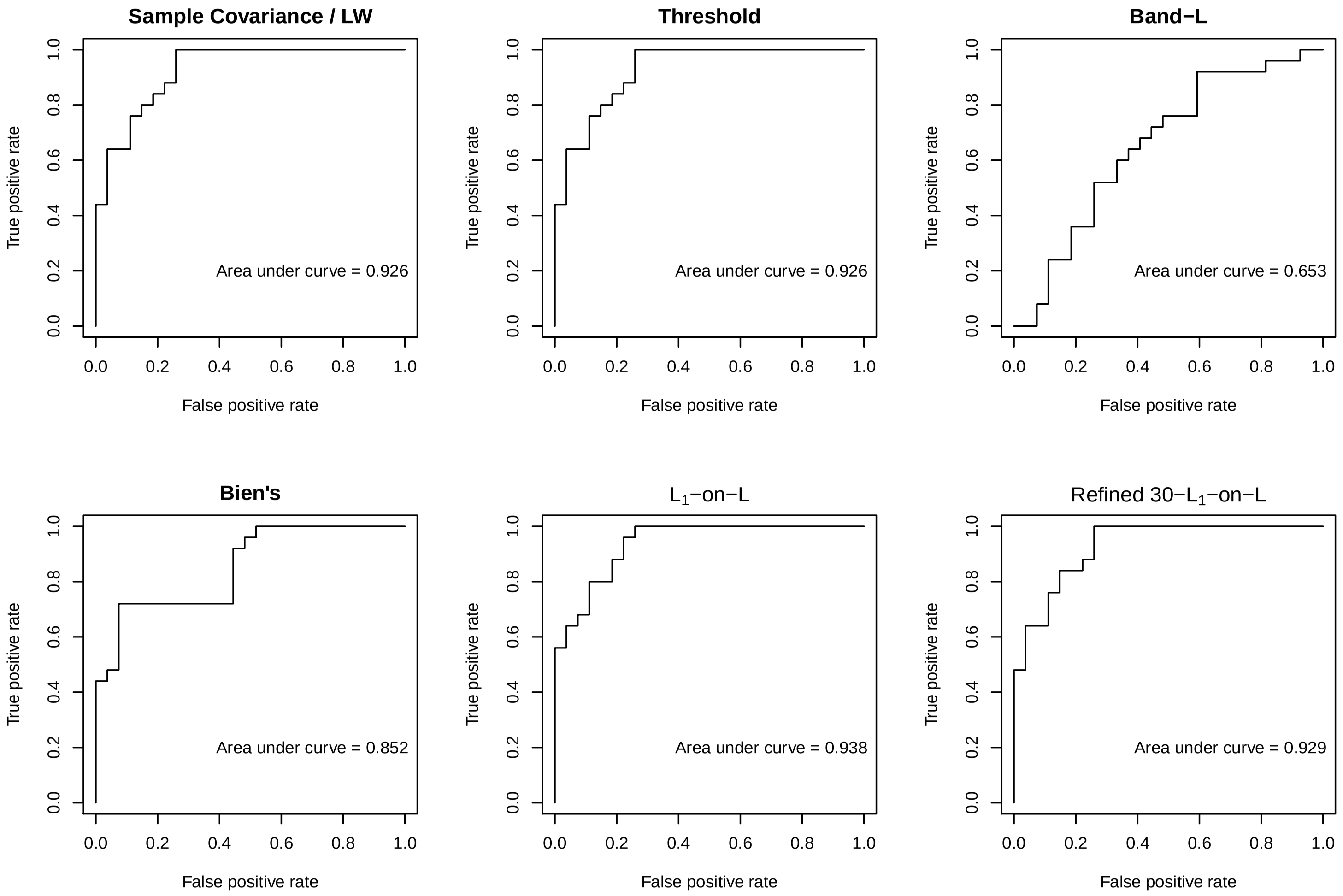

Figure 4.4: Receiver operating characteristic curves based on logistics regressions with the first principal component from various approaches.

4.4.2 Portfolio Allocation with Stock Data

Markowitz (1952) builds the foundation for modern portfolio theory. The work attempts to maximize portfolio expected return for a given amount of portfolio risk, or equivalently minimize risk for a given level of expected return by choosing the proportions of various assets. The risk is generally measured by the variance or standard deviation of the portfolio returns. The optimal mean-variance efficient portfolio is constructed by solving the following quadratic optimization

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \boldsymbol{w}^T \Sigma \, \boldsymbol{w} \\ \text{subject to} \quad & \boldsymbol{w}^T \boldsymbol{R} = R_p, \\ & \boldsymbol{w}^T \boldsymbol{e} = 1. \end{aligned}$$

$\boldsymbol{w}$ represents the proportions of various assets in the portfolio; Σ is the volatility matrix of returns in the asset pool; $\boldsymbol{R}$ consists from the returns for individual assets; r_P denotes the required return for the portfolio and $\boldsymbol{e}$ is a vector with entries equal to 1. Generally, components of $\boldsymbol{w}$ can be either positive or negative, corresponding to long or short certain assets. In common situations, $\boldsymbol{w}$ is set to be positive, i.e., the portfolio is constructed under no-shortsale constraint. Then, the optimal portfolio is obtained by solving optimization problem as follows:

$$\begin{aligned} \underset{\boldsymbol{w}}{\text{minimize}} \quad & \boldsymbol{w}^T \Sigma \, \boldsymbol{w} \\ \text{subject to} \quad & \boldsymbol{w}^T \boldsymbol{e} = 1, \\ & \boldsymbol{w} \geq 0. \end{aligned}$$

Traditionally, the sample covariance matrix is used to replace Σ above. However, in many cases, the length of the asset return series used is not big enough compared to the number of assets considered. As pointed by Michaud (1989), since the optimization problem requires the inversion of a covariance matrix, an ill-conditioned matrix results in unstable solutions and greatly amplifies the estimation error. The estimate for Σ has to be positive definite, otherwise the quadratic programming would be ill-defined. Besides, the assets do not have an order among them. Because of these issues, only methods producing positive definite and order-invariant estimates are used in comparison, including LW, Bien's and refined 30-L_1-on-L estimate. We considered the stock return data of companies included in the Standard & Poor's 100 index. Because of financial crisis in 2008, we used time zone before that period. We used weekly return data in 2006 as the training set to build portfolios, and used weekly return data in 2007 to test their performance. To comprehensively compare the differences, the performance of the portfolios in 2006 is also included. Since Mastercard, Visa and Philip Morris International are not listed throughout this time zone, these companies were excluded from the equity pool and only the remaining 97 stocks were used.

We built portfolio 1, 2 and 3 according to estimates from LW, Bien's and refined 30-L_1-on-L method, respectively. **Figure 4.5** displays the realized returns of these portfolios for all 52 weeks in both 2006 and 2007. Because the differences between the realized returns of three portfolios are dominated by the volatility caused by time effect, the patterns for the realized return look similar. To deliberately compare the performance of these portfolios, we summarize the realized returns in **Table 4.7**. The averages of realized weekly returns as well as their standard deviations are presented for both 2006 and 2007. Annualized returns are so available from synthesizing individual weekly returns. Because portfolios are achieved solving optimization with

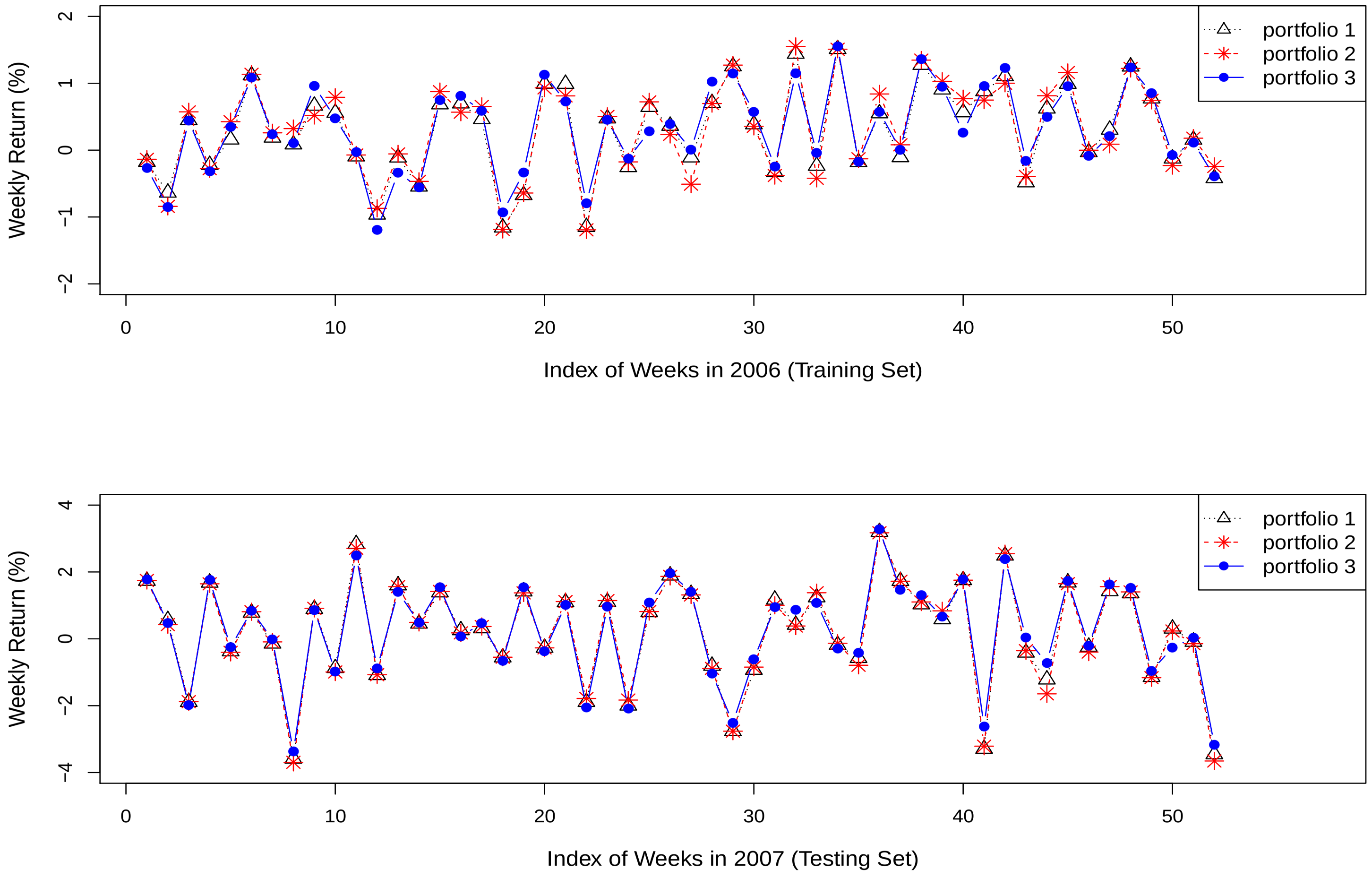

Figure 4.5: Realized weekly returns using portfolios from various approaches in 2006 and 2007. Portfolio 1 is derived from using LW estimate; portfolio 2 is derived from using Bien's estimate and portfolio 3 is derived from using 30-L_1-on-L estimate.

Year 2006 (Training Set)	Weekly Return		Annualized Return
	Arithmetic Average	Standard Deviation	
Portfolio 1 (based on LW estimate)	0.29 %	0.67 %	16.24 %
Portfolio 2 (based on Bien's estimate)	0.32 %	0.69 %	17.78 %
Portfolio 3 (based on 30-L_1-on-L estimate)	0.32 %	0.65 %	17.77 %
Year 2007 (Testing Set)	Weekly Return		Annualized Return
	Arithmetic Average	Standard Deviation	
Portfolio 1 (based on LW estimate)	0.21 %	1.56 %	10.62 %
Portfolio 2 (based on Bien's estimate)	0.19 %	1.58 %	9.89 %
Portfolio 3 (based on 30-L_1-on-L estimate)	0.26 %	1.50 %	13.58 %

Table 4.7: Summary of returns of portfolios built from using different covariance matrix estimates.

2006 weekly return data, returns from portfolios in 2006 are higher than ones in 2007, and variabilities of these portfolios in 2006 are less than ones in 2007. Regards the performance in 2006, portfolio 1 is not as good as the other two in terms of both return and volatility. Portfolio 1 and 3 produce similarly better results. When portfolios are applied to the 2007 data, portfolio 3 outperforms others. Not only its realized returns are higher than the other two, but its volatility is also less than the counterparts.

To eliminate any confusion about the choice of permutation set in this case, we adopt the similar strategy in the deviance calculation using refined 30-L_1-on-L estimate in the Parkin data study. We took 200 different sets of randomly selected permutations to obtain the covariance matrix estimates, and further, generated 200 portfolios. The behaviors of these 200 portfolios are consistent. The standard deviation of annualized returns of these 200 portfolios is 0.24% for 2006, and 0.50% for 2007. With a moderate size of permutation set, $K = 30$, the impact of permutation set selection has already been reduced to an acceptable level.

4.5 Refining Cholesky-based Inverse Covariance Matrix Estimate

The idea of refining estimates corresponding to different orders of variables can be useful in other techniques related to the modified Cholesky decomposition. Huang et al. (2006) consider estimating inverse covariance matrices through penalizing the likelihood with data from multivariate normal distributions. The implementation of their algorithm also depends on the modified Cholesky decomposition. In this section, we would like to show that the strategy of refining estimates also applies in such an approach, so that the pre-specified order information is no longer required.

4.5.1 Cholesky-based Inverse Covariance Matrix Estimation

Recall the *i.i.d* observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, and the data matrix is $\mathscr{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, a $n \times p$ matrix. When the data are from multivariate normal population with mean zero and covariance matrix Σ, the likelihood function is as follows:

$$\text{likelihood} = \prod_{i=1}^{n} f(\boldsymbol{x}_i) = \Big\{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\Big\}^n \exp\{-\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{x}_i\}.$$

Thus, the log-likelihood function is,

$$\text{log-likelihood} = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{x}_i^T\Sigma^{-1}\boldsymbol{x}_i.$$

Recall equation (2.3), (2.4) and (2.5), we have

$$\Sigma = (I_p - \Phi)^{-1}\, D^2\, \{(I_p - \Phi)^{-1}\}^T \qquad \text{and} \qquad \Sigma^{-1} = (I_p - \Phi)^T D^{-2}(I_p - \Phi),$$

where Φ is a lower triangular matrix with diagonal entries zeros. $(I_p - \Phi)$ is a unit lower triangular matrix, and so is its inverse. Hence, $|(I_p - \Phi)| = 1$ and $|\Sigma| = |D^2|$.

$$\begin{aligned} -2 \times \log(\text{likelihood}) &= np\log(2\pi) + n\log|D^2| + \sum_{i=1}^{n}\Big((I_p - \Phi)\boldsymbol{x}_i\Big)^T D^{-2}\Big((I_p - \Phi)\boldsymbol{x}_i\Big) \\ &= np\log(2\pi) + n\sum_{j=1}^{p}\log\sigma_j^2 + \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(x_{ij} - \sum_{k<j}\phi_{jk}x_{ik})^2}{\sigma_j^2}. \end{aligned}$$

In the work of Huang et al. (2006), terms of $\mathcal{L}_1$ or $\mathcal{L}_2$ penalty on (ϕ_{jk})'s are added to the $-2 \times \log(\text{likelihood})$. Therefore, the function to be minimize is as follows,

$$\begin{aligned} &-2 \times \log(\text{likelihood}) + \lambda\sum_{j=2}^{p}\sum_{k<j}|\phi_{jk}|^q \\ &= np\log(2\pi) + n\sum_{j=1}^{p}\log\sigma_j^2 + \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(x_{ij} - \sum_{k<j}\phi_{jk}x_{ik})^2}{\sigma_j^2} + \lambda\sum_{j=2}^{p}\sum_{k<j}|\phi_{jk}|^q, \quad (4.2) \end{aligned}$$

where $q = 1$ for $\mathcal{L}_1$ penalty or 2 for $\mathcal{L}_2$ penalty.

The minimization of (4.2) appears rather complicated. Fortunately, the parameters corresponding to different j's do not interact, so the whole penalized likelihood can be broken down into p pieces. The overall minimizer consists of individual minimizers from these p pieces. In details, the working flow is shown as follows:

- $j = 1$

$$\min \ \big\{ n \log \sigma_1^2 + \textstyle\sum_{i=1}^n \dfrac{x_{i1}^2}{\sigma_1^2} \big\}$$

$$\Longrightarrow \quad \hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^n x_{i1}^2$$

- $j = 2, \ldots, p$

$$\min \ \big\{ n \log \sigma_j^2 + \textstyle\sum_{i=1}^n \dfrac{(x_{ij} - \sum_{k<j} \phi_{jk} x_{ik})^2}{\sigma_j^2} + \lambda \sum_{k<j} |\phi_{jk}|^q \big\}$$

(1) if ϕ_{jk}'s are known,

$$\Longrightarrow \quad \hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^n \big(x_{ij} - \sum_{k<j} \phi_{jk} x_{ik}\big)^2$$

(2) if σ_j is known,

$$\Longrightarrow \quad \hat{\phi}_{jk} = \text{argmin} \sum_{i=1}^n \big(\frac{x_{ij}}{\sigma_j} - \sum_{k<j} \phi_{jk} \frac{x_{ik}}{\sigma_j}\big)^2 + \lambda \sum_{k<j} |\phi_{jk}|^q$$

With $q = 2$, ordinary least squares can be used to implement step (2), as long as there are enough sample points. When $q = 1$, the optimization of step (2) needs $\mathcal{L}_1$ penalty algorithms described in Chapter 3 Section (3.2.1). Iterative optimizations from step (1) and (2) produce expected estimates for a fixed j. Repeating such procedures along with $j = 2, \ldots, p$ completes the whole optimization.

With $\hat{\Phi} = (\hat{\phi}_{jk})$ and $\hat{D} = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$ available, the estimate for the inverse covariance matrix is achieved as $(I_p - \hat{\Phi})^T \hat{D}^{-2} (I_p - \hat{\Phi})$.

4.5.2 Refinement of Inverse Covariance Matrix Estimates from $\mathcal{L}_1$ Regularization

We stick to $\mathcal{L}_1$ penalty with $q = 1$ in (4.2) to make our illustration. Through minimizing this penalized log-likelihood, the inverse covariance matrix estimate is obtained. Recall in the refinement of covariance matrix estimates, we define a permutation mapping $\pi : \{1, \ldots, p\} \to \big(\pi(1),\ \ldots,\ \pi(p)\big)$ and the permutation matrix $\boldsymbol{P}_\pi = \big(e_{\pi(1)},\ \ldots,\ e_{\pi(p)}\big)$, so the columns of data matrix $\mathscr{X}$ could be permuted by right multiplying $\boldsymbol{P}_\pi$. For that circumstance, we simply replace the data matrix with $\mathscr{X}_\pi = \mathscr{X}\boldsymbol{P}_\pi = (\boldsymbol{X}_{\pi(1)}, \ldots, \boldsymbol{X}_{\pi(p)})$, so as to get the covariance matrix estimate. In a parallel manner, we carry all the definitions but use data matrix $\mathscr{X}_\pi$ in the minimization of (4.2). Suppose $\hat{B}_\pi^{-1}$ is produced as the inverse covariance matrix estimate under the order π. By applying the inverse of the mapping π, we restore the estimated inverse covariance matrix according to the original order

$$\{\hat{\Sigma}_\pi\}^{-1} = \boldsymbol{P}_{\boldsymbol{\pi}}\ \hat{B}_\pi^{-1}\ \big(\boldsymbol{P}_\pi\big)^{-1}.$$

Refining possible $\{\hat{\Sigma}_\pi\}^{-1}$'s under different π's leads to an order-invariant estimate for the inverse covariance matrix. Similar as we refine the covariance matrix estimates, we also randomly select a moderate size subset of permutations, and take the average for resultant outcomes.

Four scenarios were investigated in the simulation study of Huang et al. (2006), and the generation mechanisms for their inverse covariance matrices are listed below. To distinct these matrices from ones used elsewhere, the four matrices are marked with a star on the shoulder.

- $\{\Sigma_1^*\}^{-1} = I_p$. The identity matrix.

- $\{\Sigma_2^*\}^{-1} = \mathrm{diag}(\frac{1}{p}, \frac{1}{p-1}, \ldots, 1)$.

- $\{\Sigma_3^*\}^{-1} = \Gamma^T D^{-1} \Gamma$, where $D = 0.01 \times I_p$, and $\Gamma = (-\phi_{t,s})$, with $\phi_{t,t} = 1$,

$\phi_{t+1,t} = 0.8$, and $\phi_{t,s} = 0$ otherwise. The AR(1) model.

- $\{\Sigma_4^*\}^{-1} = \Gamma^T D^{-1} \Gamma$, where $D = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with $\sigma_t^2 = \sigma^2\{1 - \frac{(t-1)\rho^2}{1+(t-1)\rho}\}$, $1 \leq t \leq p$. and $\Gamma = (-\phi_{t,s})$ with $\phi_{t,t} = 1, \phi_{t,j} = \frac{\rho}{1+(t-1)\rho}, t \geq 2, 1 \leq j \leq t-1$, $\sigma = 1$, and $\rho = 0.5$.

The same settings used in the work of Huang et al. (2006) were adopted here. Simulation data were generated from the multivariate normal distribution with zero mean and the covariance matrix of interest. The sample size is 100 versus the number of variables 30. Each simulation was replicated for 100 times. Permutation sets with size $K = 30$ and $K = 100$ were chosen to implement the refinement strategy, where the permutations in the sets were randomly selected. The comparisons of performance for three estimates, the original estimate given the order of variables, the refined estimate based on a $K = 30$ permutation set, and the refined estimate based on a $K = 100$ permutation set, are presented in **Table 4.8**. The accuracy measures include: the L_1, L_2 and F norms of the difference between the inverse covariance matrix estimate and the true inverse covariance matrices; KL divergence of the inverse covariance matrix estimate. The average values for the accuracy measures from 100 replicates for three sets of estimates corresponding to the four covariance matrices are reported, and numbers in the parentheses indicate the standard errors for the averages.

The performance of estimates from permutation set $K = 30$ and $K = 100$ is very consistent. In comparing them to the original estimate, we found for $\{\Sigma_3^*\}^{-1}$, the original estimate using order information of variables outperforms the two refined estimates. The gap based on the L_1 norm measure is roughly 20%, while the counterparts are around 10% for the measures L_2 norm, F norm and KL divergence. The interesting part is, for $\{\Sigma_1^*\}^{-1}$, $\{\Sigma_2^*\}^{-1}$ and $\{\Sigma_4^*\}^{-1}$, the averages of accuracy measures imply better performance of the refined estimates compared with the original one. Especially for $\{\Sigma_4^*\}^{-1}$, the differences of average values for L_1 norm, L_2 norm, F norm and KL divergence, are greater than twice the corresponding standard errors, which suggests the refined estimates be superior in this scenario even if they do not take account of the order information. These comparisons show that the permutation and refinement strategy works well for the numerical experiments in the work of Huang et al., so the order information of variables might not be indispensable in their approach either. From this example, we believe that the refinement idea can be used to improve other methods that are related to the modified Cholesky decomposition.

Scenario	Measure	original estimate	refined estimate ($K = 30$)	refined estimate ($K = 100$)
$\{\Sigma_1^*\}^{-1}$	L_1 norm	0.414 (0.014)	0.408 (0.011)	0.408 (0.011)
	L_2 norm	0.398 (0.012)	0.394 (0.011)	0.394 (0.011)
	F norm	0.855 (0.014)	0.848 (0.013)	0.848 (0.013)
	KL divergence	0.326 (0.008)	0.320 (0.008)	0.321 (0.008)
$\{\Sigma_2^*\}^{-1}$	L_1 norm	0.142 (0.010)	0.132 (0.008)	0.132 (0.008)
	L_2 norm	0.132 (0.008)	0.126 (0.007)	0.125 (0.007)
	F norm	0.171 (0.007)	0.163 (0.007)	0.163 (0.007)
	KL divergence	0.324 (0.008)	0.317 (0.008)	0.316 (0.008)
$\{\Sigma_3^*\}^{-1}$	L_1 norm	144.242 (2.487)	178.799 (3.358)	177.794 (3.385)
	L_2 norm	88.449 (1.777)	99.472 (2.259)	99.239 (2.242)
	F norm	189.929 (2.213)	208.397 (2.753)	207.844 (2.701)
	KL divergence	1.218 (0.019)	1.293 (0.017)	1.284 (0.017)
$\{\Sigma_4^*\}^{-1}$	L_1 norm	6.893 (0.016)	6.840 (0.017)	6.840 (0.017)
	L_2 norm	4.765 (0.009)	4.745 (0.009)	4.745 (0.008)
	F norm	5.108 (0.009)	5.082 (0.009)	5.081 (0.009)
	KL divergence	2.406 (0.021)	2.330 (0.021)	2.329 (0.021)

Table 4.8: Performance of inverse covariance matrix estimates in different scenarios. L_1, L_2, F norm and KL divergence are used to measure the accuracy. Average values from 100 replicates are reported, and the corresponding standard errors are listed in parentheses.

5

Regularized Estimation using Matrix Exponential

This chapter follows another matrix technique, matrix exponential, in the estimation of large covariance matrices. Because the matrices involved are positive definite matrices, this technique is closely related to the spectral decomposition. The application falls in the framework of decomposition-based covariance matrix estimation.

Unlike the regularized estimation using the modified Cholesky decomposition, which doesn't require the distribution of the random vector in consideration, the application of regularized estimation using matrix exponential assumes that observations are from a multivariate normal distribution.

5.1 Motivation of Applying Matrix Exponential

The most important reason of applying matrix exponential technique is to circumvent the positive definiteness constraint of covariance matrices. As explained in Chapter 2, the only requirement of a matrix logarithm for a covariance matrix is symmetry. In other words, half of the parameters are free to vary. Thus, regularization on the parameters is easy to apply.

Recall the Bien's method described in Chapter 3, $\mathcal{L}_1$ penalty on the entries of Σ is added to the log-likelihood function directly, resulting in an estimate

$$\hat{\Sigma} = \underset{\Sigma \succ 0}{\operatorname{argmin}} \Big\{ -\log|\Sigma^{-1}| + \operatorname{tr}(\Sigma^{-1}S) + \eta \sum_s \sum_t |\sigma_{st}| \Big\}. \tag{5.1}$$

where σ_{st}'s are entries of Σ and η is the tuning parameter. The main difficulty of this approach is that the function to be optimized is not convex in term of Σ entries, so the typical optimization problems such as non-convergence, boundary constraints and multiple solutions from different initializations may occur.

Besides these issues, the minimization of function

$$-\log|\Sigma^{-1}| + \operatorname{tr}(\Sigma^{-1}S) + \eta \sum_s \sum_t |\sigma_{st}|. \tag{5.2}$$

is actually not well-defined. To see the contradiction, let η go large, and σ_{st}'s would go small and many of them might be shrank to zeros in order to minimize (5.2). However, along with the increase of η and decrease of σ_{st}'s, $|\Sigma|$ goes tiny fast, especially when the dimension is high. $|\Sigma^{-1}|$ would diverge towards positive infinity, and $-\log|\Sigma^{-1}|$ would diverge towards to negative infinity, so that the function (5.2) can not be minimized. In other words, Bien's estimate works only when the tuning parameter is allowed to vary in certain domain.

Instead, if the penalty in (5.2) is placed on the entries of the matrix logarithm of the covariance matrix, $A = \log(\Sigma)$ based on (2.10), the ill-defined optimization would be changed into the one as follows:

$$\min\Big\{ -\log|\Sigma^{-1}| + \mathrm{tr}(\Sigma^{-1}S) + \eta\sum_s\sum_t |a_{st}|^q\Big\}, \tag{5.3}$$

where a_{st}'s are entries of A, and $q > 0$ so that the penalty term is convex. This optimization is well-defined. When η goes large, and a_{st}'s would go small, $|\Sigma|$ would not necessarily go small. Even if all of a_{st}'s become zeros, $\Sigma = e^A$ is the identity matrix, which might result in a low log-likelihood. This gives another important reason of applying matrix exponential technique, because directly penalizing log-likelihood by constraining norm of covariance entries presents an ill-defined optimization problem.

One may argue that the reason of applying penalty on covariance entries is such a treatment could shape the estimated covariance matrix under some specific directions, and placing penalties on the entries of matrix logarithm of the covariance matrix may not fulfill this purpose well. As repeated throughout this thesis, directly shaping the estimated covariance matrix while guaranteeing its positive definiteness is usually not easy, especially when there is no special assumption about the covariance matrix structure. From another perspective, an desirable covariance matrix estimate is mainly supposed to better explain the data, and placing penalty on the entries of matrix logarithm serves this purpose well through shaping the estimated covariance matrix indirectly. There are many evidences to support this statement. For instance, when $q = 2$ in (5.3), $\sum_s\sum_t |a_{st}|^2$ corresponds to the sum of logarithms of eigenvalues of the covariance matrix. $\mathcal{L}_2$ regularization on entries of $\log(\Sigma)$ helps controlling the eigen-structure of the estimated covariance matrix.

There are even more direct structure connections between Σ and $\log(\Sigma)$. **Figure**

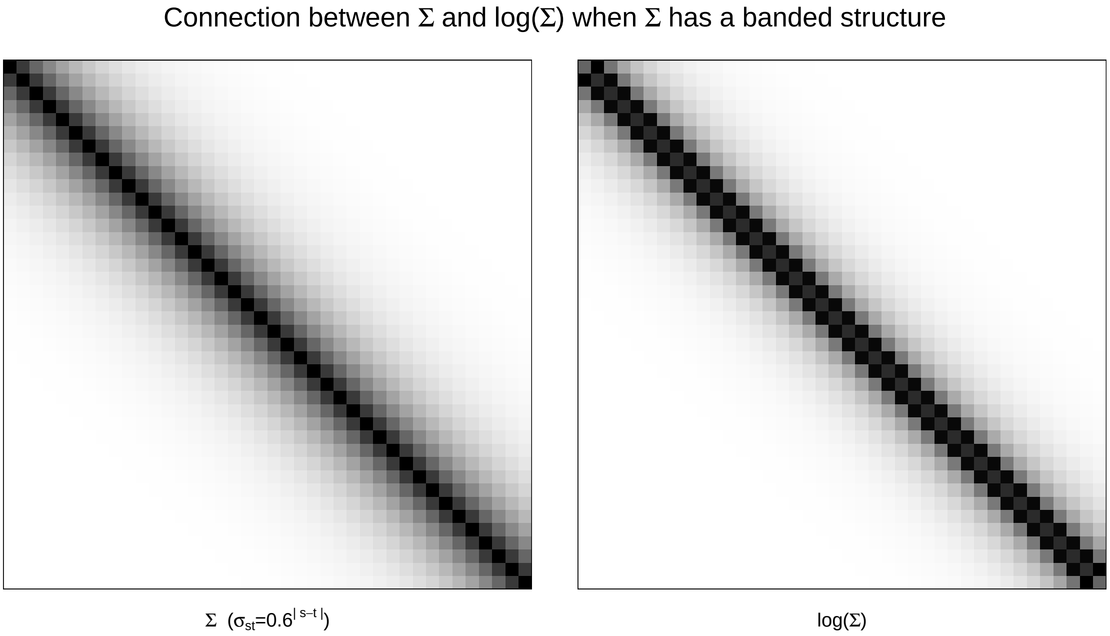

Figure 5.1: Heat maps of Σ and $\log(\Sigma)$ to illustrate their structure connection when Σ has a banded structure. $\Sigma = (\sigma_{st})$ where $\sigma_{st} = 0.6^{|s-t|}$.

5.1 and **Figure 5.2** show two situations of such connections. In these plotted heat maps, absolute values of entries are used in replace of original entries. An entry of magnitude one or over is represented by a black square and an entry of magnitude zero is represented by a white square, so the darkness indicates the extent of values. When Σ has a banded structure, typically $\log(\Sigma)$ also has a banded structure. $\Sigma = (\sigma_{st})$ with $\sigma_{st} = 0.6^{|s-t|}$ is shown on the left panel of **Figure 5.1**, so the darkness of Σ entries shades from the central diagonals. Correspondingly, the appearance of $\log(\Sigma)$ also has a shading pattern as shown on the right panel. As illustrated in **Figure 5.2**, the matrix logarithm of a positive definite covariance matrix with a diagonal block structure also has a diagonal block structure. This result is theoretical valid. Suppose $\Sigma = \text{diag}(\Sigma_1, \ldots, \Sigma_K)$ is a positive definite covariance matrix with diagonal

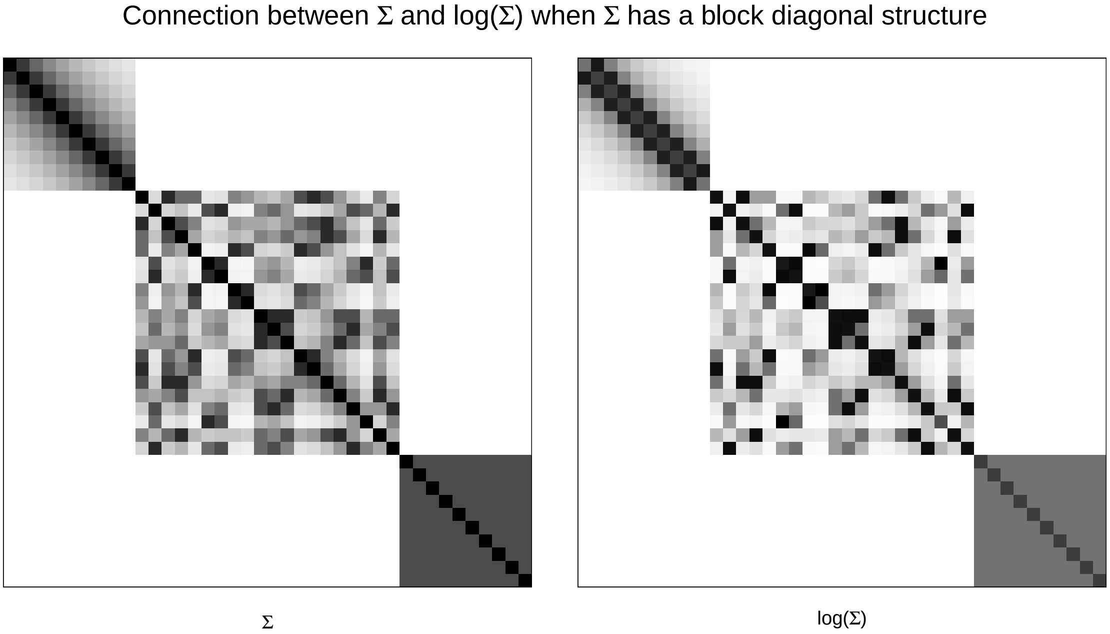

Figure 5.2: Heat maps of Σ and $\log(\Sigma)$ to illustrate their structure connection when Σ has a block diagonal structure. The block in the top left corner of Σ has $\sigma_{st} = 0.6^{|s-t|}$. The block in the middle of Σ permutes the matrix produced using $\sigma_{st} = 0.7^{|s-t|}$. The block in the bottom left of Σ has $\sigma_{st} = 0.5 + 0.5I_{\{s=t\}}$.

blocks $\Sigma_1, \ldots, \Sigma_K$. Hence, for $k = 1, \ldots, K$, Σ_k has to be positive definite and thus is accompanied by a spectral decomposition as $\Sigma_k = \Gamma_k \, \Lambda_k \, \Gamma_k^T$.

Therefore,

$$\begin{aligned}\Sigma &= \text{diag}(\Sigma_1, \ldots, \Sigma_K) = \text{diag}(\Gamma_1\Lambda_1\Gamma_1^T, \ldots, \Gamma_K\Lambda_K\Gamma_K^T) \\ &= \text{diag}(\Gamma_1, \ldots, \Gamma_K) \cdot \text{diag}(\Lambda_1, \ldots, \Lambda_K) \cdot \text{diag}(\Gamma_1, \ldots, \Gamma_K)^T \\ &= \Gamma \, \Lambda \, \Gamma^T,\end{aligned}$$

where $\Gamma = \text{diag}(\Gamma_1, \ldots, \Gamma_K)$ and $\Lambda = \text{diag}(\Lambda_1, \ldots, \Lambda_K)$. All Λ_k's are diagonal matrices, $k = 1, \ldots, K$, and thus, Λ is a diagonal matrix.

Based on (2.10) and the properties of matrix exponential listed at the end of

Chapter 2, we have

$$\begin{aligned}\log(\Sigma) &= \Gamma\ \log(\Lambda)\ \Gamma^T = \Gamma\ \mathrm{diag}(\Lambda_1, \ldots, \Lambda_K)\ \Gamma^T \\ &= \mathrm{diag}(\Gamma_1, \ldots, \Gamma_K) \cdot \mathrm{diag}(\Lambda_1, \ldots, \Lambda_K) \cdot \mathrm{diag}(\Gamma_1, \ldots, \Gamma_K)^T \\ &= \mathrm{diag}(\Gamma_1\Lambda_1\Gamma_1^T, \ldots, \Gamma_K\Lambda_K\Gamma_K^T) = \mathrm{diag}\big(\log(\Sigma_1), \ldots, \log(\Sigma_K)\big),\end{aligned}$$

Thus, $\log(\Sigma)$ also has a block diagonal structure.

Because of the structure connection between Σ and $\log(\Sigma)$, if we can restore the structure of $\log(\Sigma)$, the structure of Σ then, can be restored correspondingly.

5.2 Approximation of Log-likelihood with Matrix Exponential $\Sigma = e^{\log(\Sigma)}$

Suppose random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ follows multivariate normal distribution with mean zero and positive definite covariance matrix Σ. $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are *i.i.d* observations for $\boldsymbol{X}$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T, 1 \le i \le n$. The log-likelihood for Σ is as follows:

$$\text{log-likelihood} = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{x}_i^T\Sigma^{-1}\boldsymbol{x}_i\ .$$

Maximizing log-likelihood is equivalent to minimizing

$$\log|\Sigma| + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^T\Sigma^{-1}\boldsymbol{x}_i = \log|\Sigma| + \mathrm{tr}(\Sigma^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T) = \log|\Sigma| + \mathrm{tr}(\Sigma^{-1}S).$$

where $S = \dfrac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T$ is the sample covariance matrix.

Matrix exponential $\Sigma = e^A$ is adopted for reparameterizing the covariance matrix.

Recall the spectral decompositions for Σ and $\log(\Sigma)$, which are

$$\Sigma = \Gamma \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_p \end{pmatrix} \Gamma^T \quad \text{and} \quad A = \log(\Sigma) = \Gamma \begin{pmatrix} \log(d_1) & & \\ & \ddots & \\ & & \log(d_p) \end{pmatrix} \Gamma^T,$$

where Γ is a square matrix whose columns are the normalized eigenvectors of Σ, $d_1 \geq \ldots \geq d_p \geq 0$ are the eigenvalues of Σ. Thus,

$$\log(|\Sigma|) = \log \prod_{i=1}^{p} d_i = \sum_{i=1}^{p} \log d_i = \mathrm{tr}(A).$$

The penalized log-likelihood with L_q penalty on A can be rewritten as

$$\min \left\{ \mathrm{tr}(A) + \mathrm{tr}(e^{-A}S) + \eta \sum_s \sum_t |a_{st}|^q \right\}. \tag{5.4}$$

We mainly consider the scenarios for $q = 1$ or 2. In (5.4), it is relatively straightforward to deal with the terms $\mathrm{tr}(A) = \sum_{t=1}^{p} a_{tt}$ and $\eta \sum_s \sum_t |a_{st}|^q$ with $q = 2$. For $\eta \sum_s \sum_t |a_{st}|^q$ with q=1, the approaches described in Section (3.2.1) are able to handle the issue. Thus, the emphasis is to implement the computation of $\mathrm{tr}(e^{-A}S)$. Despite its simple appearance, the calculation of this term using A is rather complicated. Alternatively, we use linear and quadratical terms of A to approximate it in high accuracy. The approximation starts with a very useful equation — we call it "matrix exponential equality".

5.2.1 Matrix Exponential Equality

The "matrix exponential equality" comes from one of the original purposes of using matrix exponential—solving linear systems of ordinary differential equations. For a exponential matrix function $X(t)$ based on matrix A which is defined as $X(t) = e^{tA}$, it has

$$\frac{dX(t)}{dt} = A\,X(t).$$

From the perspective of approximation, we have $A = A_0 + (A - A_0)$, where A_0 is some initial setting to approach A, and thus,

$$A\,X(t) = A_0\,X(t) + (A - A_0)\,X(t).$$

Consider the following derivation operation,

$$\begin{aligned}
\frac{d}{dt}\Big(e^{-tA_0}\,X(t)\Big) &= -e^{-tA_0}\,A_0\,X(t) + e^{-tA_0}\,\frac{dX(t)}{dt} \\
&= -e^{-tA_0}\,A_0\,X(t) + e^{-tA_0}\,A\,X(t) \\
&= -e^{-tA_0}\,A_0\,X(t) + e^{-tA_0}\,A_0\,X(t) + e^{-tA_0}\,(A - A_0)\,X(t). \\
&= e^{-tA_0}\,(A - A_0)\,X(t). \qquad (5.5)
\end{aligned}$$

Integrating both sides of (5.5), we have

$$e^{-tA_0}\,X(t) = I_p + \int_0^t e^{-sA_0}(A - A_0)\,X(s)\,ds. \qquad (5.6)$$

Based on (5.6), the expected “matrix exponential equality” is shown as follows:

$$e^{tA} = e^{tA_0} + e^{tA_0}\int_0^t e^{-sA_0}(A - A_0)\,e^{sA}\,ds. \qquad (5.7)$$

5.2.2 Quadratic Approximation of Matrix Exponential e^{tA}

Matrix exponential equality (5.7) is the essential tool to approximate matrix exponential function e^{tA}. As one may expect, the computation of e^{tA} leads us to the main concern of $\mathrm{tr}(e^{-A}S)$ in (5.4). However, the actual calculation of e^{tA} requires applying spectral decomposition to A, which is not desirable if entries of A are assumed to be parameters. A more straightforward connection between e^{tA} and A is preferred so that the optimization involving e^{tA} is feasible with respect to A, and matrix exponential equality (5.7) helps to build such a connection. Particularly, we are interested in developing quadratic approximation of e^{tA} in terms of A. Before that, we would like to show the linear approximations for e^{tA}.

Linear Approximation of e^{tA}

Replacing the e^{sA} inside the integral of (5.7) with e^{sA_0}, we get the linear approximation of e^{tA} in terms of A.

$$e^{tA} \approx e^{tA_0} + e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)\, e^{sA_0}\, ds. \tag{5.8}$$

The difference between the real e^{tA} and this approximation would be

$$\text{bias}_1(t) = e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)(e^{sA} - e^{sA_0})\, ds. \tag{5.9}$$

To better understand this linear approximation, we analyze the approximation in one particular scenario, in which both A and A_0 are diagonalizable, and they commute, i.e.

$$A\, A_0 = A_0\, A.$$

Under the assumptions of two diagonalizable matrices which commute, A and A_0 are simultaneously diagonalizable (see Horn and Johnson, 2012). In other words, A and A_0 share the same eigen-space. Therefore, the same eigenvectors can be chosen to apply the spectral decomposition for both A and A_0.

$$A = \Gamma\, \log(D)\, \Gamma^T \quad \text{and} \quad A_0 = \Gamma_0\, \log(D_0)\, \Gamma_0^T \text{ with } \Gamma_0 = \Gamma. \tag{5.10}$$

Γ is an orthogonal matrix consisted of eigenvectors of A. D and D_0 are diagonal matrices whose diagonal entries are corresponding eigenvalues of e^A and e^{A_0}. Plugging (5.10) into (5.9), we have

$$\begin{aligned}
\text{bias}_1(t) &= \Gamma\, D_0^t\, \Gamma^T \int_0^t \Gamma D_0^{-s}\, \Gamma^T \big(\Gamma\big(\log(D) - \log(D_0)\big)\Gamma^T\big)(\Gamma D^s \Gamma^T - \Gamma\, D_0^s\, \Gamma^T) ds \\
&= \Gamma\, D_0^t \int_0^t D_0^{-s}\big(\log(D) - \log(D_0)\big)(D^s - D_0^s) ds\; \Gamma^T \\
&= \Gamma\, D_0^t \int_0^t \big(\log(D) - \log(D_0)\big)(D_0^{-1}D)^s - I_p) ds\; \Gamma^T \\
&= \Gamma\, D_0^t \Big((D_0^{-1}D)^t - I_p - t\big(\log(D) - \log(D_0)\big)\Big)\; \Gamma^T \\
&= \Gamma\, D^t\, \Gamma^T - \Gamma\, D_0^t\, \Gamma^T - t\Gamma D_0^t \Gamma^T \big(\Gamma \log(D)\Gamma^T - \Gamma \log(D_0)\Gamma^T\big) \\
&= e^{tA} - e^{tA_0} - t(A - A_0)\, e^{tA_0}
\end{aligned}$$

Considering the Taylor's expansion for e^A at A_0, we can tell that in this special setting of A and A_0 commuting, the linear approximation (5.8) is the same as the 1st order of Taylor's expansion. When A and A_0 can not commute, we have some empirical experience that bias (5.9) is smaller than its counterpart using Taylor's expansion, i.e. the linear approximation (5.8) is tighter than $e^{tA_0} + t(A - A_0)\, e^{tA_0}$.

Quadratic Approximation of e^{tA}

To obtain the quadratic approximation of e^{tA} in terms of A, matrix exponential equality (5.7) has to be applied for another round. Specifically, we replace e^{sA} inside the integral of (5.7) with another use of matrix exponentially equality as follows:

$$e^{sA} = e^{sA_0} + e^{sA_0} \int_0^s e^{-uA_0}(A - A_0)e^{uA}\, du.$$

Then we get another matrix equality below, which contains a double integral,

$$\begin{aligned} e^{tA} &= e^{tA_0} + e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)\big(e^{sA_0} + e^{sA_0} \int_0^s e^{-uA_0}(A - A_0)e^{uA} du\big) ds \\ &= e^{tA_0} + e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)e^{sA_0} ds \qquad (5.11) \\ &\quad + e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)e^{sA_0}\big(\int_0^s e^{-uA_0}(A - A_0)e^{uA} du\big) ds. \end{aligned}$$

Replacing the e^{uA} inside the second integral (w.r.t u) with e^{uA_0}, we get the quadratic approximation of e^{tA}, viz

$$\begin{aligned} e^{tA} &\approx e^{tA_0} + e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)e^{sA_0} ds \\ &\quad + e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)e^{sA_0}\big(\int_0^s e^{-uA_0}(A - A_0)e^{uA_0} du\big) ds. \end{aligned} \qquad (5.12)$$

If A and A_0 commute, the difference between (5.11) and (5.12) is

$$\begin{aligned} \text{bias}_2(t) &= e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)e^{sA_0}\big(\int_0^s e^{-uA_0}(A - A_0)(e^{uA} - e^{uA_0}) du\big) ds \\ &= e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)\ \text{bias}_1(s)\ ds \\ &= e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)\big(e^{sA} - e^{sA_0} - s(A - A_0)e^{sA_0}\big) ds \\ &= e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)(e^{sA} - e^{sA_0}) ds - e^{tA_0} \int_0^t e^{-sA_0}(A - A_0)s(A - A_0)e^{sA_0} ds \\ &= e^{tA} - e^{tA_0} - t(A - A_0)e^{tA_0} - \frac{t^2}{2}(A - A_0)(A - A_0)e^{tA_0}. \end{aligned}$$

Under the assumption that diagonalizable matrices A and A_0 commute, the difference between true e^{tA} and its quadratic approximation (5.12) equals the sum of cubic and higher order terms from Taylor's expansion. Consistent with our experience with the linear approximation (5.8), when A and A_0 do not commute, (5.12) approaches e^{tA} better comparing the quadratic approximation from Taylor's expansion.

5.2.3 Quadratic Approximation of $e^{-\log(\Sigma)}$

Recall our concern is $\mathrm{tr}(e^{-A}S)$ with $A = \log(\Sigma)$. The development of quadratic approximation of e^{tA} from the previous section enables us to approximate $e^{-\log(\Sigma)}$ using a quadratic function of $\log(\Sigma)$. We still use A in the derivation.

Using $t = -1$ in (5.12), we have

$$\begin{aligned}
e^{-A} &\approx e^{-A_0} - e^{-A_0}\int_0^1 e^{sA_0}(A-A_0)e^{-sA_0}ds \\
&\quad + e^{-A_0}\int_0^1 e^{sA_0}(A-A_0)e^{-sA_0}\big(\int_0^s e^{uA_0}(A-A_0)e^{-uA_0}du\big)ds \\
&= e^{-A_0} - e^{-A_0}\int_0^1 e^{sA_0}Ae^{-sA_0}ds + A_0e^{-A_0} \\
&\quad + e^{-A_0}\int_0^1 e^{sA_0}(A-A_0)e^{-sA_0}\big(\int_0^s e^{uA_0}Ae^{-uA_0}du - sA_0\big)ds \\
&= e^{-A_0} + A_0\, e^{-A_0} + \frac{1}{2}A_0^2\, e^{-A_0} \\
&\quad - e^{-A_0}\left(\int_0^1 e^{sA_0}Ae^{-sA_0}ds + \int_0^1 se^{sA_0}Ae^{-sA_0}A_0ds + A_0\int_0^1\int_0^s e^{uA_0}Ae^{-uA_0}duds\right) \\
&\quad + e^{-A_0}\int_0^1 e^{sA_0}Ae^{-sA_0}(\int_0^s e^{uA_0}Ae^{-uA_0}du)ds\ .
\end{aligned}$$

To simplify the notations, we further decompose the quadratic approximation into three parts as follows:

$$e^{-A} \approx \text{constant term of } A + \text{linear term of } A + \text{quadratic term of } A.$$

Constant Term of A

$$\text{constant term of } A = e^{-A_0}(I_p + A_0 + \frac{1}{2}A_0^2)$$

Linear Term of A

The derivation of linear term of A uses the spectral decomposition of $A_0 = \Gamma_0 \ \log(D_0) \ \Gamma_0^T$. Hence, we have

$$\int_0^s e^{uA_0} \, A \, e^{-uA_0} \, du = \Gamma_0 (\int_0^s D_0^u \, \Gamma_0^T \, A \, \Gamma_0 D_0^{-u} \, du) \, \Gamma_0^T .$$

Denote $B = \Gamma_0^T A \Gamma_0$. The above term can be rewritten as

$$\int_0^s e^{uA_0} A e^{-uA_0} du = \ \Gamma_0 (\int_0^s D_0^u B D_0^{-u} du) \ \Gamma_0^T \ = \ \Gamma_0 \int_0^s \left[b_{ij} (\frac{d_i^o}{d_j^o})^u \right]_{i,j} du \ \Gamma_0^T$$
$$= \Gamma_0 \left[b_{ij} \int_0^s (\frac{d_i^o}{d_j^o})^u du \right]_{i,j} \ \Gamma_0^T \ = \ \Gamma_0 \left[b_{ij} \Delta_{ij}(s) \right]_{i,j} \Gamma_0^T \ ,$$

with

$$\Delta_{ij}(s) = \int_0^s (\frac{d_i^o}{d_j^o})^u du = \begin{cases} s \ , & \text{if} \quad d_s^o = d_t^o \ ; \\ \frac{(d_i^o / d_j^o)^s - 1}{\log{(d_i^o / d_j^o)}} \ , & \text{if} \quad d_i^o \neq d_j^o \ . \end{cases}$$

Thus, we have the three component of the linear term of A as follows:

(1) $-e^{-A_0} \int_0^1 e^{sA_0} A \ e^{-sA_0} ds = -\Gamma_0 D^{-1} \int_0^1 \left[b_{ij} (\frac{d_i^o}{d_j^o})^s \right]_{i,j} ds \ \Gamma_0^T = -\Gamma_0 \left[b_{ij} \Delta_{ij}^{(1)} \right]_{i,j} \Gamma_0^T$;

(2) $-e^{-A_0} \int_0^1 s \ e^{sA_0} A \ e^{-sA_0} A_0 ds = -\Gamma_0 \left[b_{ij} \Delta_{ij}^{(2)} \right]_{i,j} \Gamma_0^T$;

(3) $-e^{-A_0} A_0 \int_0^1 \int_0^s e^{uA_0} A \ e^{-uA_0} du ds = -\Gamma_0 \left[b_{ij} \Delta_{ij}^{(3)} \right]_{i,j} \Gamma_0^T$;

where $\Delta_{ij}^{(1)}$, $\Delta_{ij}^{(2)}$ and $\Delta_{ij}^{(3)}$ are

$$\Delta_{ij}^{(1)} = \int_0^1 \frac{1}{d_i^o} (\frac{d_i^o}{d_j^o})^s ds = \begin{cases} 1/d_i^o \ , & \text{if} \quad d_i^o = d_j^o \ ; \\ \frac{1/d_j^o - 1/d_i^o}{\log{(d_i^o / d_j^o)}} \ , & \text{if} \quad d_i^o \neq d_j^o \ ; \end{cases}$$

$$\Delta_{ij}^{(2)} = \int_0^1 s \ \frac{1}{d_i^o} (\frac{d_i^o}{d_j^o})^s \log d_j^o \ ds = \begin{cases} \frac{\log d_j^o}{2 d_i^o} \ , & \text{if} \quad d_i^o = d_j^o \ ; \\ \frac{(1/d_j^o) \log{(d_i^o / d_j^o)} + 1/d_i^o - 1/d_j^o}{\log^2{(d_i^o / d_j^o)}} \log d_j^o \ , & \text{if} \quad d_i^o \neq d_j^o \ ; \end{cases}$$

$$\Delta_{ij}^{(3)} = \int_0^1 \frac{\log d_i^o}{d_i^o}\ \Delta_{ij}(s)ds = \begin{cases} \frac{\log d_i^o}{2d_i^o}\ , & \text{if}\quad d_i^o = d_j^o\ ; \\ \frac{1/d_j^o - 1/d_i^o - \log{(d_i^o/d_j^o)} 1/d_i^o}{\log^2{(d_i^o/d_j^o)}} \log d_i^o\ , & \text{if}\quad d_i^o \neq d_j^o\ . \end{cases}$$

Therefore, the linear term of A is $-\Gamma_0 \left[b_{ij} \Delta_{ij}^* \right]_{i,j} \Gamma_0^T$ with $\Delta_{ij}^* = \Delta_{ij}^{(1)} + \Delta_{ij}^{(2)} + \Delta_{ij}^{(3)}$.

Quadratic Term of A

$$\begin{aligned} & e^{-A_0} \int_0^1 e^{sA_0} A e^{-sA_0} (\int_0^s e^{uA_0} A\ e^{-uA_0} du)\, ds \\ &= \Gamma_0\, D_0^{-1} \int_0^1 D_0^s\, \Gamma_0^T\ A\, \Gamma_0\ D_0^{-s} (\int_0^s D_0^u\, \Gamma_0^T\ A\, \Gamma_0\ D_0^{-u}\ du)\ ds\ \Gamma_0^T\ . \end{aligned}$$

The part in the middle can be rewritten as

$$\begin{aligned} & D^{-1} \int_0^1 D_0^s \Gamma_0^T A\, \Gamma_0 D_0^{-s} (\int_0^s D_0^u \Gamma_0^T A\, \Gamma_0 D_0^{-u} du)\, ds \\ &= D^{-1} \int_0^1 D_0^s\ B\ D_0^{-s} (\int_0^s D_0^u\ B\ D_0^{-u} du)\, ds \\ &= D^{-1} \int_0^1 \left[b_{ij} (\frac{d_i^o}{d_j^o})^s \right]_{i,j} \left[b_{ij} \Delta_{ij}(s) \right]_{i,j} ds \\ &= D^{-1} \int_0^1 \left[\sum_k b_{ik} (\frac{d_i^o}{d_k^o})^s \cdot b_{kj} \Delta_{kj}(s) \right]_{i,j} ds \\ &= \left[\sum_k b_{ik} b_{kj} \frac{1}{d_i^o} \int_0^1 (\frac{d_i^o}{d_k^o})^s \cdot \Delta_{kj}(s) ds \right]_{i,j} = \left[\sum_k b_{ik} b_{kj} \Delta_{ikj}^* \right]_{i,j} \end{aligned}$$

with

$$\Delta_{ikj}^* = \frac{1}{d_i^o} \int_0^1 (\frac{d_i^o}{d_k^o})^s \cdot \Delta_{kj}(s)\, ds = \begin{cases} \frac{1}{2d_i^o}\ , & d_i^o = d_k^o = d_j^o\ ; \\ \frac{(d_i^o/d_j^o) - 1 - \log{(d_i^o/d_j^o)}}{\log^2{(d_i^o/d_j^o)}} \frac{1}{d_i^o}\ , & d_i^o = d_k^o, d_k^o \neq d_j^o\ ; \\ \frac{(d_i^o/d_k^o) - 1 - \log{(d_i^o/d_k^o)}}{\log^2{(d_i^o/d_k^o)}} \frac{1}{d_i^o}\ , & d_i^o \neq d_k^o, d_j^o = d_i^o\ ; \\ \frac{(d_i^o/d_j^o) \log{(d_i^o/d_j^o)} - (d_i^o/d_j^o) + 1}{\log^2{(d_i^o/d_j^o)}} \frac{1}{d_i^o}\ , & d_i^o \neq d_k^o, d_j^o = d_k^o\ ; \\ \frac{\frac{(d_i^o/d_j^o) - 1}{\log{(d_i^o/d_j^o)}} - \frac{(d_i^o/d_k^o) - 1}{\log{(d_i^o/d_k^o)}}}{\log{(d_k^o/d_j^o)}} \frac{1}{d_i^o} = \frac{\Delta_{ij}^{(1)} - \Delta_{ik}^{(1)}}{\log{(d_k^o/d_j^o)}} \frac{1}{d_i^o}\ , & d_i^o \neq d_k^o, d_j^o \neq d_i^o, d_j^o \neq d_k^o\ . \end{cases}$$

Therefore, the quadratic term of A is $-\Gamma_0 \left[b_{ij}\Delta^*_{ikj}\right]_{i,j} \Gamma_0^T$.

Finally, we get

$$\begin{aligned}
e^{-A} &\approx e^{-A_0} + A_0 e^{-A_0} + \frac{1}{2}A_0^2 e^{-A_0} \\
&\quad - e^{-A_0}\Big(\int_0^1 e^{tA_0}Ae^{-tA_0}dt + \int_0^1 te^{tA_0}Ae^{-tA_0}A_0 dt + A_0\int_0^1\int_0^t e^{sA_0}Ae^{-sA_0}dsdt\Big) \\
&\quad + e^{-A_0}\int_0^1 e^{tA_0}Ae^{-tA_0}(\int_0^t e^{sA_0}Ae^{-sA_0}ds)dt \\
&= e^{-A_0}(I_p + A_0 + \frac{1}{2}A_0^2) - \Gamma_0\left[b_{ij}\Delta^*_{ij}\right]_{i,j}\Gamma_0^T + \Gamma_0\left[\sum_k b_{ik}b_{kj}\Delta^*_{ikj}\right]_{i,j}\Gamma_0^T.
\end{aligned} \tag{5.13}$$

Although (5.13) still seems complex, the implementation of quadratic approximation for $e^{-\log\Sigma}$ in terms of $\log\Sigma$ is routine. Based on that, we move back to the penalized log-likelihood.

5.3 Penalized Log-likelihood with Matrix Exponential Reparameterization

Plugging the approximation of e^{-A} from (5.13) into the penalized log-likelihood of (5.4), we have

$$\begin{aligned}
&\mathrm{tr}(A) + \mathrm{tr}(e^{-A}S) + \eta\sum_s\sum_t |a_{st}|^q \\
&\approx \mathrm{tr}(A) + \mathrm{tr}\big(e^{-A_0}(I_p + A_0 + \frac{1}{2}A_0^2)S\big) - \mathrm{tr}\big(\Gamma_0\left[b_{ij}\Delta^*_{ij}\right]_{i,j}\Gamma_0^T S\big) \\
&\quad + \mathrm{tr}\big(\Gamma_0\left[\sum_k b_{ik}b_{kj}\Delta^*_{ikj}\right]_{i,j}\Gamma_0^T S\big) + \eta\sum_s\sum_t |a_{st}|^q \\
&= \text{constance term} + \mathrm{tr}(A) - \mathrm{tr}\big(\left[b_{ij}\Delta^*_{ij}\right]_{i,j}\tilde{S}\big) \\
&\quad + \mathrm{tr}\big(\left[\sum_k b_{ik}b_{kj}\Delta^*_{ikj}\right]_{i,j}\tilde{S}\big) + \eta\sum_s\sum_t |a_{st}|^q ,
\end{aligned}$$

where $\tilde{S} = \Gamma_0^T S \Gamma_0$. In addition to the results $B = (b_{st})$ and $B = \Gamma_0^T A \Gamma_0$, the minimization of the approximated penalized log-likelihood is more like a standard convex optimization. For the $\mathcal{L}_2$ penalty, i.e. q=2, based on (5.10), it is clear that

$$\sum_{s,t} a_{st}^2 = \text{tr}(A^2) = \text{tr}(\Gamma\ \log D\ \Gamma^T\ \Gamma \log D\ \Gamma^T) = \text{tr}\big((\log D)^2\big) = \sum_{i=1}^{p} \{\log(d_i)\}^2\ ,$$

so optimizing the $\mathcal{L}_2$ penalized log-likelihood produces a covariance matrix estimate with eigenvalues under control. Following this direction, Deng and Tsui (2013) analyze the behavior of such an estimate through numerical investigations. As explained at the beginning of this chapter, there are other connections between covariance matrices and their logarithm matrices including the structure relationships. Estimates from minimizing $\mathcal{L}_1$ (q=1) penalized log-likelihood are expected to have their own advantages. We call such an approach "L_1-on-A", and carried out numerical studies for its performance.

5.4 Numerical Studies of Regularized Estimation with $\mathcal{L}_1$ Penalty on $\log(\Sigma)$

5.4.1 Simulation Study

We consider using Σ_2 and Σ_3 from Chapter 3 to conduct the simulation study. Recall Σ_2 is produced by permuting the banded covariance matrix $\Sigma_1 = (\sigma_{st})$ with $\sigma_{st} = I_{\{s=t\}} + 0.4 \times I_{\{|s-t|=1\}}$ and $\Sigma_3 = (\sigma_{st})$ with $\sigma_{st} = I_{\{s=t\}} + 0.4 \times I_{\{|s-t|=p/5\}}$, where p is the number of variables. For each scenario, we generated normally distributed data with three settings: (1) $n = 50$, $p = 30$; (2) $n = 50$, $p = 50$; (3) $n = 50$, $p = 100$. Each case was repeated 200 times, and the average values from the 200 accuracy measures were reported as well as their corresponding standard errors.

Regarding accuracy measures, the L_1, L_2 and F norms of the difference matrix between the estimated covariance matrix and the true one were calculated. Entropy loss of estimates is included as well.

With respect to the choice of tuning parameter η, we also adopt the repeated learning-testing to seek a balance between the quality of trained estimate and the quality of reference covariance matrix. Unlike using F norm for the difference between the estimated covariance matrix and the sample covariance matrix in Chapter 3, in this implementation, we assume multivariate normal data, so a likelihood-based criterion should be more appropriate. In details, We repeatedly split the data into learning and testing sets with roughly equal sizes for V times. Let $\hat{\Sigma}^{(v)}(\eta)$ be the estimated covariance matrix based on the learning data with tuning parameter η in the v-th replicate, $v = 1, \ldots, V$. Denote I_v the index set of the data in the learning set, $s^{(v)}$ the size of I_v, $v = 1, \ldots, V$. The measure function is chosen to be the likelihood, so that the tuning parameter is selected using

$$\hat{\eta} = \underset{\eta}{\operatorname{argmin}} \frac{1}{V} \sum_{v=1}^{V} \Big\{ s^{(v)} \log |\hat{\Sigma}^{(v)}(\eta)| - \sum_{i \in I_v} y_i^T (\hat{\Sigma}^{(v)}(\eta))^{-1} y_i \Big\}.$$

Results of simulation study are reported in **Table 5.1**, following the same way of presentations in the tables of Chapter 3 and 4. The performance for the sample covariance matrix, LW estimate and the refined L_1-on-L estimate based on a permutation set of size $K = 30$, is listed in comparison.

From **Table 5.1**, the L_1-on-A estimate has the best performance for both Σ_2 and Σ_3 in terms of L_1 and L_2 norm for the the difference between the estimated covariance

		$\Sigma = \Sigma_2$				$\Sigma = \Sigma_3$			
p	measure	Sample	LW	30-L_1-on-L	L_1-on-A	Sample	LW	30-L_1-on-L	L_1-on-A
30	L_1 norm	4.87 (0.04)	2.07 (0.02)	1.37 (0.01)	1.31 (0.01)	4.88 (0.04)	1.87 (0.02)	1.26 (0.01)	1.23 (0.01)
	L_2 norm	2.03 (0.02)	0.88 (0.01)	0.88 (0.01)	0.75 (0.01)	2.02 (0.02)	0.79 (0.00)	0.84 (0.01)	0.71 (0.01)
	F norm	4.30 (0.02)	2.53 (0.01)	2.25 (0.01)	2.01 (0.01)	4.31 (0.02)	2.34 (0.00)	2.12 (0.01)	1.91 (0.01)
	entropy	12.58 (0.06)	7.22 (0.04)	4.05 (0.04)	4.78 (0.04)	12.58 (0.06)	4.84 (0.03)	3.04 (0.03)	3.36 (0.03)
50	L_1 norm	7.98 (0.05)	2.50 (0.02)	1.47 (0.01)	1.42 (0.01)	7.96 (0.05)	2.22 (0.02)	1.35 (0.01)	1.30 (0.01)
	L_2 norm	2.96 (0.02)	0.91 (0.00)	0.94 (0.00)	0.91 (0.00)	2.93 (0.02)	0.81 (0.00)	0.89 (0.00)	0.83 (0.00)
	F norm	7.11 (0.02)	3.50 (0.00)	3.08 (0.01)	3.13 (0.01)	7.13 (0.02)	3.20 (0.00)	2.88 (0.01)	2.88 (0.01)
	entropy	-	14.68 (0.06)	8.20 (0.05)	10.08 (0.05)	-	9.71 (0.04)	5.93 (0.04)	7.06 (0.04)
100	L_1 norm	15.36 (0.13)	3.07 (0.05)	1.57 (0.01)	1.30 (0.01)	15.39 (0.12)	2.74 (0.04)	1.47 (0.01)	1.20 (0.01)
	L_2 norm	4.83 (0.04)	0.91 (0.00)	1.00 (0.00)	0.89 (0.00)	4.77 (0.04)	0.82 (0.00)	0.93 (0.00)	0.80 (0.00)
	F norm	14.09 (0.05)	5.28 (0.01)	4.67 (0.01)	5.10 (0.01)	14.11 (0.05)	4.79 (0.01)	4.34 (0.01)	4.59 (0.00)
	entropy	-	34.49 (0.17)	20.35 (0.09)	29.56 (0.07)	-	22.40 (0.11)	14.38 (0.07)	19.64 (0.05)

Table 5.1: Comparisons of performance between the L_1-on-A estimate and other estimates in Σ_2 and Σ_3. Averages of measures from 200 replicates are listed, and numbers in the parentheses indicate their standard errors.

matrix and the sample covariance matrix. With respect to F norm, the performance of L_1-on-A estimate is similar as the one of refined 30-L_1-on-L estimate, and is better than the performance of the sample covariance matrix and LW estimate. The entropy loss of L_1-on-A estimate is larger than the loss of refined 30-L_1-on-L estimate, but smaller than the one of LW estimate.

5.4.2 Portfolio Allocation Revisit

We revisit the portfolio allocation problem in Chapter 3. In addition to the previous 3 portfolios according to LW, BT and the 30-L_1-on-L estimate, we further built portfolio 4 based on L_1-on-A using the 52 weekly returns in 2006 of 97 stocks.

We summarize the realized returns in **Table 5.2** to compare the performance. Portfolio 4 does not perform so well in the training set of stock data in 2006. The returns as well as standard deviation of the returns are close to portfolio 1 and they are not as good as portfolio 2 and 3. For the testing set of stock data in 2007, the performance of portfolio 4 is satisfactory. The standard deviation is the smallest while the annualized return is the second highest.

Year 2006 (Training Set)	Weekly Return		Annualized Return
	Arithmetic Average	Standard Deviation	
Portfolio 1 (based on LW estimate)	0.29 %	0.67 %	16.24 %
Portfolio 2 (based on Bien's estimate)	0.32 %	0.69 %	17.78 %
Portfolio 3 (based on 30-L_1-on-L estimate)	0.32 %	0.65 %	17.77 %
Portfolio 4 (based on L_1-on-A estimate)	0.29 %	0.66 %	16.20 %
Year 2007 (Testing Set)	Weekly Return		Annualized Return
	Arithmetic Average	Standard Deviation	
Portfolio 1 (based on LW estimate)	0.21 %	1.56 %	10.62 %
Portfolio 2 (based on Bien's estimate)	0.19 %	1.58 %	9.89 %
Portfolio 3 (based on 30-L_1-on-L estimate)	0.26 %	1.50 %	13.58 %
Portfolio 4 (based on L_1-on-A estimate)	0.25 %	1.49 %	12.96 %

Table 5.2: Summary of returns of portfolios derived from using different covariance matrix estimates.

6

Conclusion and Discussion

Estimation of population covariance matrices from samples of multivariate data is of great importance. The sample covariance matrix estimate becomes less attractive with the increase of the number of variables. In many applications such as gene expression, fMRI, spectroscopic imaging, weather forecasting and others, the number of variables largely overrides the sample size. In this situation, the sample covariance matrix becomes degenerate with a distorted eigen-structure (Johnstone, 2001). Therefore, it is important to explore appropriate covariance matrix estimation in large dimensions.

A natural thinking of improving covariance matrix estimation is to modify the sample covariance matrix. Like the estimate from Ledoit and Wolf (2004), it is a Stein-type estimate that shrinks the sample covariance matrix towards the identity matrix. A different group of methods focus on regularizing the sample covariance matrix. Bickel and Levina (2008b) consider thresholding small entries of the sample covariance matrix to zeros. Dealing with covariance matrices with banded structures, Bickel and Levina (2008a) consider banding the sample covariance matrix through only keeping entries in the diagonal and certain sub-diagonals non-zeros. This group of methods can not guarantee the positive definiteness of the estimated covariance matrix.

To pursue improved covariance matrix estimate with guaranteed positive definiteness, one perspective is to apply regularization on the covariance entries while treating them as parameters. This strategy usually requires sophisticated optimization techniques in order to meet the positive definiteness. As an example, the estimate from Bien and Tibshirani (2011) is obtained through optimizing the $\mathcal{L}_1$ penalized log-likelihood using a majorization-minimization technique. Such sophisticated optimization often involves intensive computation and convergence issue.

Another perspective of improving covariance matrix estimates with guaranteed positive definiteness is not to directly regularizing the covariance entries. Rather, through appropriate matrix decomposition, the regularization could be placed on the entries of the factor matrices instead of on the original covariance entries. Therefore, the constraint of positive definiteness is circumvented. Two ways of reparameterizations based on matrix decomposition are considered. One is associated with the modified Cholesky decomposition, and the other one, using matrix exponential, is associated with the spectral decomposition.

The modified Cholesky decomposition from Pourahmadi (1999) is a widely used tool in dealing with covariance matrices. The sequential regressions in accordance with the modified Cholesky decomposition provide an unconstrained reparameterization of the covariance matrix, and regularization can be easily applied to the Cholesky factor matrix for it is equivalent to regularizing the coefficients of the linear regressions. Incorporating the advantages of Bickel and Levina's banding idea, Rothman et al. (2010) propose to band the Cholesky factor matrix of the covariance matrix so that the estimated covariance matrix is always positive definite. The covariance matrix estimation through regularizing the Cholesky factor matrix is not necessarily limited to the scenarios in which the covariance matrices are banded.

That is the reason we choose to employ $\mathcal{L}_1$ regularization on the Cholesky factor matrix to estimate the covariance matrix in a more general situation where particular assumption of the matrix structure is not necessary. Besides that, the covariance matrix estimation through employing $\mathcal{L}_1$ regularization on the Cholesky factor matrix does not suffer the constraint of insufficient sample points as much as the approach of banding the Cholesky factor matrix. More importantly, we find that the estimate from $\mathcal{L}_1$ regularization is not sensitive towards the order of variables. One prerequisite condition of using the modified Cholesky decomposition is the order information of variables. Often, such information is not available, or can not be reasonably assumed. Weakening or even getting rid of this requirement can greatly broaden the usage of this technique.

In application of $\mathcal{L}_1$ regularization, it is true one may encounter certain disagreement between estimates using different orders of variables, since the penalized linear regressions are not necessarily the same. However, such disagreement should not be considerable. We showed that when the tuning parameter is set to be zero, covariance matrix estimate based on any order of variables is exactly equal to the sample covariance matrix, which is order-invariant. The application of $\mathcal{L}_1$ regularization aims to improve the estimate from the the sample covariance matrix towards the true covariance matrix. Marching along this track with both ends are order-invariant matrices, estimation using different orders of variables may implement the improvement differently, but is not expected to deviate far away from the track. The property that the estimate resulted from $\mathcal{L}_1$ regularization is not sensitive towards the order serves well to eliminate this prerequisite. It motivates us to propose an order-invariant covariance matrix estimate from refining estimates under random permutations of the order of variables. Such refinement from individual estimates can be analog to building a random forest from individual decision trees. While one individual tree may overfit the data, a particular-order-based estimate could be more preferred under certain order. Combining representative estimates corresponding to different orders is like building a random forest so that the performance becomes more stable.

The theoretically order-invariant estimate is approximated from taking average of many representative ones. This refinement strategy enables the refined estimate to inherit positive definiteness and parsimony from individual estimates. Besides taking average for maintaining these properties, other strategies may have similar effects. For example, one may choose to refine Cholesky factor matrices L's and standard deviation matrices D's from individual estimates. The averages of L's and D's can be used to form an refined covariance matrix estimate with positive definiteness and parsimonious property.

The choice of this representative group is not essential since the mechanism of random selection achieves representativeness. For instance, in our simulation study, the proposed estimates under refinement with permutation groups size 30 and 100 presented similar performance. In the real data analysis, the standard deviation measures for estimates from different selection of permutation groups showed the minimal extents of variabilities for the results. Therefore, the guideline for the group size is to seek balance between computation convenience and estimation accuracy. A moderate number, like 30, should provide adequate performance in practice.

The idea of refining estimates corresponding to different orders of variables could also be useful in the applications of other techniques based on the modified Cholesky decomposition. For instance, Huang et al. (2006) propose methods of inverse covariance matrix estimation by maximizing the log-likelihood with $\mathcal{L}_1$ and $\mathcal{L}_2$ penalty on the entries of inverse covariance matrices. The implementation of their algorithms is also based on the modified Cholesky decomposition. To examine the suitability of our refinement strategy in their approach, we carried out their simulation studies, and applied our refinement treatment to their estimates. In summary, the refining strategy works as well in the estimates from Huang et al.. As an example, such an investigation helps eliciting further usage of the refinement idea in other techniques related to the modified Cholesky decomposition.

The matrix exponential technique is also able to circumvent the positive definiteness constraint in covariance matrix estimation. Based on the spectral decomposition, the matrix logarithm of a covariance matrix is closely related to the original covariance matrix. Since the only restrict of the matrix logarithm is symmetry, when the parameters to be estimated are the entries in the lower part of the matrix logarithm, these parameters are free to vary, and regularization on the parameters is easy to apply.

One may argue that the reason of applying penalty on covariance entries is, through such treatments, the estimated covariance matrix could be shaped towards some targets, and employ regularization on the entries of matrix logarithm of the covariance matrix may not fulfill this purpose well. Directly shaping the estimated covariance matrix while guaranteeing its positive definiteness is usually not easy, especially when there is no special assumption about the covariance matrix structure. From another perspective, a desirable covariance matrix estimate helps better explaining the data, and placing penalty on the entries of matrix logarithm serves this purpose well through shaping the estimated covariance matrix indirectly. There are many evidences to support this statement. In terms of eigenvalues, the sum of logarithms of eigenvalues from the original covariance matrix, is equal to the squared F norm of matrix logarithm of the covariance matrix. In terms of matrix structures, when a covariance matrix is banded, its matrix logarithm tends to be banded as well; when a covariance is block diagonal, its matrix logarithm is also block diagonal. Because of these properties, $\mathcal{L}_2$ penalty on entries of the matrix logarithm helps controlling the extreme eigenvalues of the estimated covariance matrix, and $\mathcal{L}_1$ penalty on entries of the matrix logarithm helps restoring the structure of covariance matrix through restoring the structure of the matrix logarithm.

The magnitude of penalties, measured by tuning parameters, needs to be determined by procedures such as cross validation or repeated learning-testing. One thing worth mentioning about the tuning procedure for L_1-on-A is, the scale is correspondent to the logarithm operation. Maybe not comparably in a strict sense, the selected tuning parameter in L_1-on-A is often significantly smaller than the selected one in L_1-on-L in the same scenario of simulation study. While the tuning set includes all candidate choices for the tuning parameter, the variability of the penalized function values will be much greater if the same tuning set is used in L_1-on-A than in L_1-on-L. Such a result is not unnatural, however, in practice, it does imply the tuning procedure for L_1-on-A should be more delicate. Without any prior information, one may tend to use the same scale of tuning sets for different approaches. While one tuning set is delicate enough for L_1-on-L, it could be too loose in selecting the tuning parameter for L_1-on-A.

References

Z. Bai and Y. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993.

Z. D. Bai, J. W. Silverstein, and Y. Q. Yin. A note on the largest eigenvalue of a large dimensional sample covariance matrix. *Journal of Multivariate Analysis*, 26 (2):166–168, 1988.

P. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008a.

P. Bickel and E. Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008b.

J. Bien and R. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98 (4):807–820, 2011.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM: Journal on Scientific Computing*, 20(1):33–61, 1998.

T. Y. Chiu, T. Leonard, and K.-W. Tsui. The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210, 1996.

W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.

A. Constantine. Some non-central distribution problems in multivariate analysis. *The Annals of Mathematical Statistics*, 34(4):1270–1285, 1963.

X. Deng and K.-W. Tsui. Penalized covariance matrix estimation using a matrix-logarithm transformation. *Journal of Computational and Graphical Statistics*, 22 (2):494–512, 2013.

D. K. Dey and C. Srinivasan. Estimation of a covariance matrix under stein's loss. *The Annals of Statistics*, 13(4):1581–1591, 1985.

D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

B. Efron and C. Morris. Multivariate empirical bayes and estimation of covariance matrices. *The Annals of Statistics*, pages 22–32, 1976.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

J. H. Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84:165–175, 1989.

W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

S. Geman. A limit theorem for the norm of random matrices. *The Annals of Probability*, 8(2):252–261, 1980.

L. Haff. Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, 8(3):586–597, 1980.

L. Haff. The variational form of certain bayes estimators. *The Annals of Statistics*, 19(3):1163–1190, 1991.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 2012.

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

D. R. Hunter and K. Lange. Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

D. Jonsson. Some limit theorems for the eigenvalues of a sample covariance matrix. *Journal of Multivariate Analysis*, 12(1):1–38, 1982.

J. Khan, J. Wei, M. Ringnér, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7(6):673–679, 2001.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

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

S. Lin and M. D. Perlman. A monte carlo comparison of four estimators of a covariance matrix. *Multivariate Analysis*, 6:411–429, 1985.

H. Liu, L. Wang, and T. Zhao. Sparse covariance matrix estimation with eigenvalue constraints. *Journal of Computational and Graphical Statistics*, (in press), 2013.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4):457–483, 1967.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

R. O. Michaud. The markowitz optimization enigma: is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42, 1989.

P. Painleve. Sur les équations différentielles du second ordre et d'ordre supérieur dont l'intégrale générale est uniforme. *Acta mathematica*, 25(1):1–85, 1902.

M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

M. Pourahmadi. Covariance estimation: The glm and regularization perspectives. *Statistical Science*, 26(3):369–387, 2011.

A. Rothman, E. Levina, and J. Zhu. A new approach to cholesky-based covariance regularization in high dimensions. *Biometrika*, 97(3):539–550, 2010.

A. J. Rothman, E. Levina, and J. Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

C. Scherzer, A. Eklund, L. Morse, Z. Liao, J. Locascio, D. Fefer, M. Schwarzschild, M. Schlossmacher, M. Hauser, J. Vance, et al. Molecular markers of early parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences*, 104(3):955–960, 2007.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978.

J. W. Silverstein. The smallest eigenvalue of a large dimensional Wishart matrix. *The Annals of Probability*, 13(4):1364–1368, 1985.

J. W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2): 331–339, 1995.

J. W. Silverstein and S.-I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

C. Stein. Estimation of a covariance matrix. *Rietz Lecture, 39th Annual Meeting IMS. Atlanta, Georgia.*, 1975.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

C. A. Tracy and H. Widom. Level-spacing distributions and the airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, 1994.

C. A. Tracy and H. Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177(3):727–754, 1996.

K. W. Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *The Annals of Probability*, 6(1):1–18, 1978.

J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20(1/2):32–52, 1928.

W. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.

Y. Yin, Z. Bai, and P. Krishnaiah. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields*, 78 (4):509–521, 1988.

Y. Q. Yin. Limiting spectral distribution for a class of random matrices. *Journal of Multivariate Analysis*, 20(1):50–68, 1986.