DOCTORAL THESIS

High-dimensional covariance matrix estimation with application to Hotelling's tests

Dong, Kai

Date of Award:
2015

High-Dimensional Covariance Matrix Estimation with Application to Hotelling's Tests

DONG Kai

A thesis submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Principal Supervisor: Dr. TONG Tiejun

Hong Kong Baptist University

August 2015

DECLARATION

I hereby declare that this thesis represents my own work which has been done after registration for the degree of PhD at Hong Kong Baptist University, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

Signature: _____

Date: August 2015

ABSTRACT

In recent years, high-dimensional data sets are widely available in many scientific areas, such as gene expression study, finance and others. Estimating the covariance matrix is a significant issue in such high-dimensional data analysis. This thesis focuses on high-dimensional covariance matrix estimation and its application.

First, this thesis focuses on the covariance matrix estimation. In Chapter 2, a new optimal shrinkage estimation of the covariance matrices is proposed. This method is motivated by the quadratic discriminant analysis where many covariance matrices need to be estimated simultaneously. We shrink the sample covariance matrix towards the pooled sample covariance matrix through a shrinkage parameter. Some properties of the optimal shrinkage parameter are investigated and we also provide how to estimate the optimal shrinkage parameter. Simulation studies and real data analysis are also conducted. In Chapter 4, we estimate the determinant of the covariance matrix using some recent proposals for estimating high-dimensional covariance matrix. Specifically, a total of nine covariance matrix estimation methods will be considered for comparison. Through extensive simulation studies, we explore and summarize some interesting comparison results among all compared methods. A few practical guidelines are also made on the sample size, the dimension, and the correlation of the data set for estimating the determinant of high-dimensional covariance matrix. Finally, from a perspective of the loss function, the comparison study in this chapter also serves as a proxy to assess the performance of the covariance matrix estimation.

Second, this thesis focuses on the application of high-dimensional covariance matrix estimation. In Chapter 3, we consider to estimate the high-dimensional covariance matrix based on the diagonal matrix of the sample covariance matrix and apply it to the Hotelling's tests. In this chapter, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. We also propose several different ways to derive the approximate null distribution under different scenarios of p and n for our proposed shrinkage-based test. Simulation studies show that the

proposed method performs comparably to existing competitors when n is moderate or large, and it is better when n is small. In addition, we analyze four gene expression data sets and they demonstrate the advantage of our proposed shrinkage-based diagonal Hotelling's test.

Apart from the covariance matrix estimation, we also develop a new classification method for a specific type of high-dimensional data, RNA-sequencing data. In Chapter 5, we propose a negative binomial linear discriminant analysis for RNA-Seq data. By Bayes' rule, we construct the classifier by fitting a negative binomial model, and propose some plug-in rules to estimate the unknown parameters in the classifier. The relationship between the negative binomial classifier and the Poisson classifier is explored, with a numerical investigation of the impact of dispersion on the discriminant score. Simulation results show the superiority of our proposed method. We also analyze four real RNA-Seq data sets to demonstrate the advantage of our method in real-world applications.

Keywords: Covariance matrix, Discriminant analysis, High-dimensional data, Hotelling's test, Log determinant, RNA-sequencing data

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who helped me during my Ph.D. study.

First and foremost, my deepest gratitude goes to my supervisor, Dr. Tiejun Tong, for his constant encouragement and guidance. During my Ph.D. study, he spent much time on discussing academic problems with me, passing his research experience and writing skills to me. Without his kind support, this thesis would not have been completed. I would also like to thank my co-supervisor, Prof. Lixing Zhu, for his helpful suggestions and comments on my work.

Second, I would like to express my deep thanks to those who helped me in my research in the past three years. Prof. Marc G. Genton from King Abdullah University of Science and Technology, Prof. Hongyu Zhao from Yale University, Dr. Xiang Wan from Hong Kong Baptist University, Dr. Heng Peng from Hong Kong Baptist University, Dr. Zhishui Hu from University of Science and Technology of China, Dr. Herbert Pang from The University of Hong Kong, Dr. Gaorong Li from Beijing University of Technology, and Dr. Cheng Wang from Shanghai Jiao Tong University provided many helpful comments and suggestions on my work. I am also grateful to Dr. Wenlin Dai from Hong Kong Baptist University Institute of Research and Continuing Education for his continuous assistance and encouragement.

I would like to thank all faculties and staffs in the Department of Mathematics at Hong Kong Baptist University for their lectures and sincere help. I thank all my classmates and friends for their help.

Finally, my sincere thanks go to my parents who have always been helping me, supporting me and encouraging me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 High-dimensional Covariance Matrix Estimation

In recent years, based on the development of modern science and technology, an increasing number of variables can be observed or measured in many scientific areas. For instance, in biomedical research, DNA microarray, which is a revolutionary technology in gene expression study, can measure thousands or even tens of thousands of genes simultaneously. Recently, the next generation sequencing technology is widely applied in transcriptome profiling and protein-DNA interaction and a large number of genes can also be measured by this new sequencing technology. One important feature of such high-dimensional data sets is that the number of variables or dimension p is often comparable with or even larger than the number of observations n. Such feature is not uncommon in gene expression study. For example, in many microarray experiments, the sample size is much smaller than the number of genes because of the cost or rare patients (Kuster et al.; 2011; Mokry et al.; 2012; Kaur et al.; 2012; Searcy et al.; 2012). Such high-dimensional data bring challenges for statistical analysis including estimating the covariance matrix. In next two sections we briefly introduce two applications of estimating high-dimensional covariance matrix.

Let $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})^T$, $i = 1, \ldots, n$, be independent and identically distributed

(i.i.d.) random vectors with the covariance matrix Σ. The sample covariance matrix

$$\mathbf{S}_n = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T, \tag{1.1}$$

where $\bar{\mathbf{X}} = (1/n)\sum_{i=1}^{n}\mathbf{X}_i$, is a common estimator of the covariance matrix. However, in such high-dimensional settings, the sample covariance matrix is singular and is no longer a valid estimate of Σ. To overcome the singularity problem, various covariance matrix estimation methods are proposed in recent literatures. One type of methods is to shrink the sample covariance matrix towards a target matrix through a shrinkage parameter, and some properties of the optimal shrinkage parameter are investigated (Ledoit and Wolf; 2003; Schäfer and Strimmer; 2005; Fisher and Sun; 2011). Another type of methods is to add some structure assumptions to the covariance matrix. One assumption is the sparsity structure which assumes that most of elements of the covariance matrix are zeros. Consistent estimators of the covariance matrix can be obtained under such matrix structure (Bickel and Levina; 2008; Cai and Yuan; 2012; Rothman; 2012). Recently, Fan et al. (2013) pointed out that the sparsity assumption may not be realistic and considered a conditional sparsity structure. In their paper, the authors introduced a principle orthogonal complement thresholding method using the factor model to estimate high-dimensional covariance matrix. Additionally, Tong et al. (2014) reviewed various existing covariance matrix estimation methods.

1.2 Hotelling's Tests

One application of high-dimensional covariance matrix estimation is Hotelling's T^2 test (Hotelling; 1931). In biomedical research, scientists need to know whether a gene set or pathway is significantly differentially expressed in two experiments. In statistics this is essentially a two-sample multivariate hypothesis testing problem. Hotelling's T^2 test is a classical tool to solve such multivariate testing problems including one-sample test and two-sample test problems, which are described as follows:

- *One-sample test.* Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T,\ i = 1, \ldots, n$, be i.i.d. random vectors from a multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and

covariance matrix Σ. The hypothesis of one-sample test is

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \tag{1.2}$$

where $\boldsymbol{\mu}_0$ is a fixed vector. And the respective Hotelling's T^2 statistic is

$$T_1^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T \mathbf{S}_n^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0). \tag{1.3}$$

Under the null hypothesis, $\{(n-p)/[p(n-1)]\}T_1^2$ follows an $F_{p,n-p}$ distribution with p and $n-p$ degrees of freedom when $p \leq n-1$.

- *Two-sample test.* Let $\mathbf{X}_{ki} = (X_{ki1}, \ldots, X_{kip})^T$, $i = 1, \ldots, n_k$, be i.i.d. random vectors from a multivariate normal distribution with the mean vector $\boldsymbol{\mu}_k$ and the common covariance matrix Σ, for $k = 1$ and 2, respectively. The hypothesis of two-sample test is

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \tag{1.4}$$

and the respective Hotelling's T^2 statistic is

$$T_2^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \mathbf{S}_{pool}^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2), \tag{1.5}$$

where $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ are the sample means and $\mathbf{S}_{pool} = \{(n_1 - 1)\mathbf{S}_{n1} + (n_2 - 1)\mathbf{S}_{n2}\}/(n_1 + n_2 - 2)$ is the pooled sample covariance matrix. Under the null hypothesis, $\{(n_1 + n_2 - p - 1)/[p(n_1 + n_2 - 2)]\}T_2^2$ follows an F_{p,n_1+n_2-p-1} distribution with p and $n_1 + n_2 - p - 1$ degrees of freedom when $p \leq n_1 + n_2 - 2$.

Obviously, for both one-sample and two-sample tests, we need to estimate the covariance matrix. The sample covariance matrix is used as the estimator of the covariance matrix in Hotelling's T^2 tests. However, it suffers from the singularity problem in high-dimensional settings because it is no longer invertible. Therefore, high-dimensional covariance matrix estimation methods are necessary when Hotelling's T^2 tests are applied.

1.3 Discriminant Analysis

Another application of high-dimensional covariance matrix estimation is discriminant analysis. In biomedical research, scientists measure the gene expression levels of biopsy or serum sample from an individual and identify whether this individual has a disease and/or a specific type of disease. In statistics this is essentially a classification problem. Discriminant analysis, including quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA), is one type of classical method to solve such classification problems. In particular, let K be the number of distinct classes. The individual or sample from class k, where $k = 1, \cdots, K$, follows a multivariate normal distribution $N_p(\mu_k, \Sigma_k)$, where μ_k and Σ_k are the mean vector and covariance matrix, respectively. QDA allocates the new individual to the class which makes the discriminant score of QDA minimum. The discriminant score of QDA is

$$d_k^Q(\mathbf{x}^*) = (\mathbf{x}^* - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}^* - \boldsymbol{\mu}_k) + \log |\Sigma_k| - 2 \log \pi_k,$$

and π_k is the prior probability that one individual comes from class k. Note that LDA is a special case of QDA when $\Sigma_k = \Sigma$ for all k.

To compute the discriminant scores of LDA and QDA, we need to estimate the covariance matrix. Like what we mentioned in Sections 1.1 and 1.2, the sample covariance matrix is a common estimator but it is not invertible for high-dimensional data. Therefore, new high-dimensional covariance matrix estimation is also necessary in discriminant analysis.

1.4 Overall Structure

The remainder of this thesis is organized as follows.

In Chapter 2, we propose the optimal shrinkage estimation of the covariance matrices, which is motivated by the quadratic discriminant analysis where many covariance matrices need to be estimated simultaneously. We compute the pooled sample covariance matrix which is the arithmetic mean of all sample covariance matrices, and shrink one sample covariance matrix towards the pooled sample covariance matrix

to obtain its respective estimate. We also investigate some properties and estimation methods of the optimal shrinkage parameter. Simulation studies and real data analysis are conducted to investigate the performance of our methods.

In Chapter 3, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. We also propose several different ways to derive the approximate null distribution under different scenarios of p and n for our proposed shrinkage-based test. Simulation studies show that the proposed method performs comparably to existing competitors when n is moderate or large, and it is better when n is small. In addition, we analyze four gene expression data sets and they demonstrate the advantage of our proposed shrinkage-based diagonal Hotelling's test.

In Chapter 4, we estimate the determinant of the covariance matrix using some recent proposals for estimating high-dimensional covariance matrix. Specifically, a total of nine covariance matrix estimation methods will be considered for comparison. Through extensive simulation studies, we explore and summarize some interesting comparison results among all compared methods. A few practical guidelines are also made on the sample size, the dimension, and the correlation of the data set for estimating the determinant of high-dimensional covariance matrix. Finally, from a perspective of the loss function, the comparison study in this chapter also serves as a proxy to assess the performance of the covariance matrix estimation.

In Chapter 5, we propose a negative binomial linear discriminant analysis for RNA-Seq data. By Bayes' rule, we construct the classifier by fitting a negative binomial model, and propose some plug-in rules to estimate the unknown parameters in the classifier. The relationship between the negative binomial classifier and the Poisson classifier is explored, with a numerical investigation of the impact of dispersion on the discriminant score. Simulation results show the superiority of our proposed method. We also analyze four real RNA-Seq data sets to demonstrate the advantage of our method in real-world applications.

In Chapter 6, we summarize the contents of this thesis and propose some future work.

Chapter 2

Optimal Shrinkage Estimation of the Covariance Matrices

2.1 Introduction

High-throughput techniques allow us to acquire thousands of or more gene expression values simultaneously, which introduces novel approaches to genetic research. One important goal of analyzing gene expression microarray data is to identify to which type of diseases a new patient belongs. Essentially, in statistics, this is a classification problem. Quadratic discriminant analysis (QDA) is a classical method which is used to solve such classification problem. Suppose we have G classes and the covariance matrices of G classes are $\Sigma_1, \ldots, \Sigma_G$, respectively. QDA assumes that Σ_g, $g = 1, \ldots, G$ are unequal and we need to compute the discriminant scores to decide to which class a new sample belongs to. In the discriminant scores, the covariance matrices Σ_g, $g = 1, \ldots, G$ are unknown and hence need to be estimated. In other words, to apply QDA to the classification problem, we need to estimate the many covariance matrices simultaneously.

To estimate one covariance matrix, many methods are developed in the literature. The sample covariance matrix is a common estimator but suffers from the singularity problem when the data set is high-dimensional. To overcome such singularity problem, the diagonal matrix of the sample variances is an available high-dimensional

covariance matrix estimator. This method has been used in high-dimensional classification (Dudoit et al.; 2002; Pang et al.; 2009). Another way is to shrink the sample covariance matrix towards a target matrix (Schäfer and Strimmer; 2005; Fisher and Sun; 2011). In recent years, a sparse structure of the covariance matrix is proposed. They assume most of covariates of the data are independent, and try to delete some weak correlations of the sample covariance matrix by thresholding (Bickel and Levina; 2008; Rothman et al.; 2009; Cai and Liu; 2011; Cai and Yuan; 2012). More recently, Fan et al. (2013) proposed a new method to estimate the covariance matrix using a factor model. They apply the spectral decomposition to the sample covariance matrix, keep the first K principle components, and at the same time apply thresholding to the principal orthogonal complement to obtain the final estimate.

To apply QDA, we can select one aforementioned method to estimate the covariance matrices Σ_g and then compute the respective discriminant scores. However, all these existing methods focus on the covariance matrix of one class and ignore the information from other classes. It is not uncommon that borrowing information can improve the variance estimation. Stein (1964) discovered that we can take advantage of the information from the sample mean to improve the variance estimation. Tong and Wang (2007) proposed an optimal shrinkage variance estimator which borrowing information from variances and demonstrated the improvement of the variance estimation. This motivates us to taking advantage of the information from other class to improve the covariance matrix estimation.

In this chapter, we propose new optimal shrinkage estimators to estimate many covariance matrices simultaneously. We shrink the sample covariance matrix towards the pooled sample covariance matrix which is the arithmetic mean of the sample covariance matrices of all classes. The shrinkage parameter in the estimators controls the shrinkage level. Two loss functions are considered in this chapter and we attempt to find the optimal shrinkage parameters and investigate their related properties under these two loss functions. We conduct simulation studies and real data analysis to investigate the performance of our methods.

2.2 Shrinkage Covariance Matrix Estimators

In this section, we introduce the optimal shrinkage covariance matrix estimation. Let $\mathbf{X}_{gi} = (X_{1gi}, \ldots, X_{pgi})^T$ be i.i.d. random vectors from the multivariate normal distribution $N_p(\boldsymbol{\mu}_g, \Sigma_g)$, where $i = 1, \ldots, n$, $\boldsymbol{\mu}_g$ is p-dimensional mean vector and Σ_g is the $p \times p$ covariance matrix. Traditionally, we use the sample covariance matrix $\mathbf{S}_{gn} = (1/n)S_g$ to estimate Σ_g, where

$$S_g = \sum_{i=1}^{n} (\mathbf{X}_{gi} - \bar{\mathbf{X}}_g)(\mathbf{X}_{gi} - \bar{\mathbf{X}}_g)^T, \tag{2.1}$$

and $\bar{\mathbf{X}}_g = (1/n) \sum_{i=1}^{n} \mathbf{X}_{gi}$. Note that $S_g \sim W_p(\Sigma_g, n)$, where $W_p(\Sigma_g, n)$ denotes the Wishart distribution with n degrees of freedom.

Let S_g be independent. To estimate Σ_g, we propose the following shrinkage estimators:

$$\hat{\Sigma}_g = \alpha \mathbf{S}_{gn} + (1 - \alpha)\hat{\Sigma}_{\text{pool}}, \quad 0 \leq \alpha \leq 1, \tag{2.2}$$

where

$$\hat{\Sigma}_{\text{pool}} = \frac{1}{G} \sum_{g=1}^{G} \mathbf{S}_{gn},$$

and α is the unknown shrinkage parameter which controls the shrinkage level. It is obvious that if $\alpha = 1$, the estimators are the sample covariance matrix, and if $\alpha = 0$, the estimates are shrunken to the pooled sample covariance matrix.

2.3 Optimal Shrinkage Parameters Estimation

In real application, the shrinkage parameter α is unknown and hence we need to find its estimation. To estimate the shrinkage parameter, we first consider to derive the average risk of the covariance matrix estimator under a loss function. Then we investigate some properties of the average risk and estimate the optimal shrinkage parameter based on the average risk. Note that the optimal shrinkage parameter in this chapter is not overall optimal. Our optimal shrinkage parameter is obtained in the domain specified by two loss functions which will be introduced in the next paragraph.

In this chapter, we consider the following two loss functions:

$$L_1(\hat{\Sigma}, \Sigma) = \mathrm{tr}(\hat{\Sigma}\Sigma^{-1}) - \log\det(\hat{\Sigma}\Sigma^{-1}) - p, \tag{2.3}$$

and

$$L_2(\hat{\Sigma}, \Sigma) = \mathrm{tr}(\hat{\Sigma}\Sigma^{-1} - I)^2, \tag{2.4}$$

where $\det(\cdot)$ and $\mathrm{tr}(\cdot)$ denote the determinant and trace of a covariance matrix, respectively. Note that both two loss functions have been used in the covariance matrix estimation (Haff; 1980, 1991; Yang and Berger; 1994). The first loss function is Stein Loss. For ease of interpretation, we take $p = 1$ for example. When $p = 1$, Stein Loss is

$$L(\hat{\sigma}^2, \sigma^2) = \frac{\hat{\sigma}^2}{\sigma^2} - \log\frac{\hat{\sigma}^2}{\sigma^2} - 1,$$

where $\hat{\sigma}^2$ is the estimator of σ^2. If $\hat{\sigma}^2 = \sigma^2$, the loss is 0. If $\hat{\sigma}^2$ goes to zero or infinity, the loss will go to infinity. This means that this loss function has the same heavy penalty for overestimation and underestimation. The second loss function is a commonly used function: Quadratic Loss.

2.3.1 Optimal Estimator Under the Loss Function L_1

Under L_1, the average risk is

$$\begin{aligned} R_1(\alpha, \hat{\Sigma}, \Sigma) &:= \frac{1}{G}\sum_{g=1}^{G} EL_1(\hat{\Sigma}_g, \Sigma_g) \\ &= \frac{1}{G}\sum_{g=1}^{G} E\mathrm{tr}(\hat{\Sigma}_g\Sigma_g^{-1}) - \frac{1}{G}\sum_{g=1}^{G} E\log\det(\hat{\Sigma}_g\Sigma_g^{-1}) - p \\ &:= B_1 + B_2 - p, \end{aligned} \tag{2.5}$$

where (by noting that $ES_g = n\Sigma_g$)

$$\begin{aligned} B_1 &= \frac{1}{G}\sum_{g=1}^{G} E\mathrm{tr}(\hat{\Sigma}_g\Sigma_g^{-1}) = \frac{1}{G}\sum_{g=1}^{G} \mathrm{tr}E(\hat{\Sigma}_g\Sigma_g^{-1}) \\ &= \alpha p + (1-\alpha)\mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g^{-1}\Big)\Big], \end{aligned}$$

and

$$B_2 = -\frac{1}{G}\sum_{g=1}^{G} E\log\det(\hat{\Sigma}_g\Sigma_g^{-1}) = -\frac{1}{G}\sum_{g=1}^{G} E\log\det\Big(\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat{\Sigma}_{\text{poo1}}\Big)\Sigma_g^{-1}\Big).$$

In the following discussion, the derivatives $R_k'(\alpha,\hat{\Sigma},\Sigma)$ and $R_k''(\alpha,\hat{\Sigma},\Sigma)$ are with respect to α, $k=1,2$.

Theorem 1. *$R_1(\alpha,\hat{\Sigma},\Sigma)$ is a strictly convex function of α on $[0,1]$ that satisfies*

(a). $R_1'(\alpha,\hat{\Sigma},\Sigma)|_{\alpha=0} \le 0$, the equality holds true if and only if $\Sigma_g = \Sigma_{g'}$ for all $g,g' \in \{1,2,\cdots,G\}$.

(b). $R_1'(\alpha,\hat{\Sigma},\Sigma)|_{\alpha=1} > 0$

By Theorem 1, there exists a unique $\alpha_1^* \in [0,1)$ satisfies

$$\alpha_1^* = \arg\min_{\alpha\in[0,1]} R_1(\alpha,\hat{\Sigma},\Sigma),$$

or, equivalently,

$$R_1'(\alpha,\hat{\Sigma},\Sigma)|_{\alpha=\alpha_1^*} = 0.$$

Theorem 2. *For any fixed G,p, as $n\to\infty$, we have*

(i) $\alpha_1^* \to 1$ when Σ_g are not all the same.

(ii) $R_1(\alpha,\hat{\Sigma},\Sigma)$ tends to a constant function of α when $\Sigma_g = \Sigma_{g'}$ for any $g = g'$.

Theorem 2 implies that it is unnecessary to borrow information across other classes if the sample size is large.

In real applications, the optimal shrinkage parameter α_1^* is unknown and need to be estimated. From the proof of Theorem 1, we know the optimal shrinkage parameter α_1^* is the unique solution to

$$\begin{aligned} R_1'(\alpha,\hat{\Sigma},\Sigma) &= p - \text{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g^{-1}\Big)\Big] \\ &\quad -\frac{1}{G}\sum_{g=1}^{G}\text{tr}E\Big(\Big(\frac{S_g}{n}-\hat{\Sigma}_{\text{poo1}}\Big)\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat{\Sigma}_{\text{poo1}}\Big)^{-1}\Big). \\ &= 0. \end{aligned} \tag{2.6}$$

We estimate Σ_g and Σ_g^{-1} by the unbiased estimators S_g/n and $(n-p-1)S_g^{-1}$, respectively and $R_1'(\alpha, \hat{\Sigma}, \Sigma)$ by

$$\begin{aligned}\hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma) &= p - \mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\frac{S_g}{n}\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}(n-p-1)S_g^{-1}\Big)\Big] \\ &\quad - \frac{1}{G}\sum_{g=1}^{G}\mathrm{tr}\Big(\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\Big(\alpha\frac{S_g}{n} + (1-\alpha)\hat{\Sigma}_{\mathrm{poo1}}\Big)^{-1}\Big). \quad (2.7)\end{aligned}$$

By using the same method as that in the proof of Theorem 1, for any $\alpha \in [0,1]$, we have

$$\hat{R}_1''(\alpha, \hat{\Sigma}, \Sigma) = -\frac{1}{G}\sum_{g=1}^{G}\Big(\log\det\Big(\alpha\frac{S_g}{n} + (1-\alpha)\hat{\Sigma}_{\mathrm{poo1}}\Big)\Big)''_{\alpha} > 0,$$

which implies that $\hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)$ is strictly increasing on $[0,1]$. Clearly,

$$\begin{aligned}\hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)|_{\alpha=0} &= p - \frac{n-p-1}{n}\mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}S_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}S_g^{-1}\Big)\Big], \\ \hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)|_{\alpha=1} &= \frac{p+1}{n}\mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}S_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}S_g^{-1}\Big)\Big] > 0.\end{aligned}$$

Here $\hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)|_{\alpha=0}$ is not guaranteed to be negative. If $\hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)|_{\alpha=0} \le 0$, then there exists a unique α satisfies $\hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma) = 0$ and we denote the solution as $\hat{\alpha}_1^*$. Otherwise, we define $\hat{\alpha}_1^* = 0$.

For any $p \times p$ matrix A, we define its Frobenius norm $\|A\|$ as

$$\|A\| := (\mathrm{tr}(AA^T))^{1/2} = \Big(\sum_{i,j=1}^{p}a_{ij}^2\Big)^{1/2}, \quad A \in \mathbb{H},$$

where a_{ij} are the coefficients of A. Then for any random matrix $A = (A_{ij})$, $E\|A\|^k < \infty$ is equivalent to $E|A_{ij}|^k$ for any $i, j \in \{1, 2, \cdots, p\}$.

In this chapter, for $p \times p$ matrix A, $A > 0$ means that A is a positive definite matrix. The following two theorems show some asymptotic results of the estimated optimal shrinkage parameters under the loss function L_1.

Theorem 3. *For any fixed G, p, as $n \to \infty$, we have $\hat{\alpha}_1^* \to 1$ when Σ_g are not all the same.*

Theorem 4. *For any fixed n and p with $n > p+1$, assume that $\Sigma_g \overset{i.i.d.}{\sim} \mu$, where μ is a probability measure supported on $\mathbb{H}^+ := \{A_{p\times p} : A > 0\}$ and $E\|S_1\|^7 < \infty$, $E\|S_1^{-1}\|^7 < \infty$. Then $R_1'(\alpha, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma) \to 0$ a.s. uniformly for $\alpha \in [0,1]$ as $G \to \infty$. In addition, $\hat{\alpha}_1^* - \alpha_1^* \to 0$ a.s.*

2.3.2 Optimal Estimator Under the Loss Function L_2

Under the loss function L_2, the average risk is

$$R_2(\alpha, \hat{\Sigma}, \Sigma) \;:=\; \frac{1}{G}\sum_{g=1}^{G} EL_2(\hat{\Sigma}_g, \Sigma_g) = \alpha^2 A_1 + 2\alpha A_2 + A_3,$$

where

$$A_1 = \frac{1}{G}\sum_{g=1}^{G} E\mathrm{tr}\Big(\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\Sigma_g^{-1}\Big)^2, \tag{2.8}$$

$$A_2 = \frac{1}{G}\sum_{g=1}^{G} E\mathrm{tr}\Big(\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\Sigma_g^{-1}\Big(\hat{\Sigma}_{\mathrm{poo1}}\Sigma_g^{-1} - I\Big)\Big), \tag{2.9}$$

$$A_3 = \frac{1}{G}\sum_{g=1}^{G} E\mathrm{tr}\Big(\hat{\Sigma}_{\mathrm{poo1}}\Sigma_g^{-1} - I\Big)^2. \tag{2.10}$$

Theorem 5. *$R_2(\alpha, \hat{\Sigma}, \Sigma)$ is a strictly convex function of α on $[0,1]$ with the unique minimum point at*

$$\alpha_2^* = -\frac{A_2}{A_1}.$$

Theorem 6. *For any fixed G, p, as $n \to \infty$, we have*

(i) $\alpha_2^* \to 1$ when Σ_g are not all the same.

(ii) $R_2(\alpha, \hat{\Sigma}, \Sigma)$ tends to a constant function of α when $\Sigma_g = \Sigma_{g'}$ for any $g = g'$.

To estimate the optimal shrinkage parameter α_2^*, we rewrite A_1 and A_2 as

$$\begin{aligned} A_1 \;&=\; \frac{1}{G}\sum_{g=1}^{G} \mathrm{tr}\Big(\Big(\Sigma_g - \bar{\Sigma}\Big)\Sigma_g^{-1}\Big)^2 + \frac{(G-1)^2(p^2+p)}{nG^2} \\ &\qquad + \frac{1}{nG^3}\sum_{g=1}^{G}\sum_{g'\neq g}\Big(\mathrm{tr}((\Sigma_{g'}\Sigma_g^{-1})^2) + (\mathrm{tr}(\Sigma_{g'}\Sigma_g^{-1}))^2\Big), \end{aligned} \tag{2.11}$$

$$A_2 \;=\; -A_1 + \frac{(G-1)(p^2+p)}{nG}, \tag{2.12}$$

where $\bar{\Sigma} = \frac{1}{G}\sum_{g=1}^{G}\Sigma_g$.

We go to estimate A_1. First, we must find a good estimator of

$$A := \frac{1}{G}\sum_{g=1}^{G}\mathrm{tr}\Big(\Big(\Sigma_g - \bar{\Sigma}\Big)\Sigma_g^{-1}\Big)^2. \tag{2.13}$$

As follows, assume that $n > p+3$. For any fixed constant matrix Σ_0, we have

$$E[\mathrm{tr}(I - r\Sigma_0 S_g^{-1})^2 - r(\mathrm{tr}(\Sigma_0 S_g^{-1})^2 + (\mathrm{tr}(\Sigma_0 S_g^{-1}))^2)] = \mathrm{tr}\Big((\Sigma_g - \Sigma_0)\Sigma_g^{-1}\Big)^2, \tag{2.14}$$

where $r = n - p - 1$. Thus we can use

$$\mathrm{tr}(I - r\Sigma_0 S_g^{-1})^2 - r(\mathrm{tr}(\Sigma_0 S_g^{-1})^2 + (\mathrm{tr}(\Sigma_0 S_g^{-1}))^2)$$

to estimate $\mathrm{tr}((\Sigma_g - \Sigma_0)\Sigma_g^{-1})^2$ for fixed Σ_0. By using $\hat{\Sigma}_{\mathrm{poo1}}$ to estimate $\bar{\Sigma}$ and noting that $A > 0$, we can define an estimator of A in (2.13) as

$$\hat{A} = \max\Big\{0,\ \frac{1}{G}\sum_{g=1}^{G}\Big(\mathrm{tr}\Big(I - r\hat{\Sigma}_{\mathrm{poo1}}S_g^{-1}\Big)^2 - r\Big(\mathrm{tr}(\hat{\Sigma}_{\mathrm{poo1}}S_g^{-1})^2 + (\mathrm{tr}(\hat{\Sigma}_{\mathrm{poo1}}S_g^{-1}))^2\Big)\Big)\Big\}.$$

Then we can estimate A_1, A_2 by

$$\begin{aligned}
\hat{A}_1 &= \hat{A} + \frac{(G-1)^2(p^2+p)}{nG^2} + \frac{1}{nG^3}\sum_{g=1}^{G}\sum_{g'\neq g}\Big(\mathrm{tr}((S_{g'}S_g^{-1})^2) + (\mathrm{tr}(S_{g'}S_g^{-1}))^2\Big),\\
\hat{A}_2 &= -\hat{A}_1 + \frac{(G-1)(p^2+p)}{nG}.
\end{aligned} \tag{2.15}$$

By using the similar method as that in the proof of Theorem 5, we have $0 \le -\hat{A}_2 \le \hat{A}_1$. Thus we can estimate the optimal shrinkage estimators by

$$\hat{\alpha}_2^* = -\frac{\hat{A}_2}{\hat{A}_1}.$$

Theorem 7. *For any fixed G, p, as $n \to \infty$, we have $\hat{\alpha}_2^* \to 1$ a.s. when Σ_g are not all the same.*

Theorem 8. *For any fixed n and p with $n > p+3$, assume that $\Sigma_g \overset{i.i.d.}{\sim} \mu$, where μ is a probability measure supported on $\mathbb{H}^+ := \{A_{p\times p} : A > 0\}$ and $E\|S_1\|^2 < \infty$, $E\|S_1^{-1}\|^2 < \infty$. Then $\hat{\alpha}_2^* - \alpha_2^* \to 0$ a.s. as $G \to \infty$.*

The above two theorems show the asymptotic results of the estimated optimal shrinkage parameters under the loss function L_2.

2.4 Simulation Studies

2.4.1 Simulation Design

In this section we conduct simulations to investigate the performance of the proposed optimal shrinkage covariance matrices estimators. The block diagonal covariance matrices are considered in our simulation studies. This type of structure has been considered in Guo et al. (2007), Pang et al. (2009), Tong, Chen and Zhao (2012) and so forth. Specifically, we consider the following block diagonal covariance matrices

$$\Sigma_g = D_g^{1/2} R D_g^{1/2},$$

where

$$R = \begin{pmatrix} \Sigma_\rho & 0 & \cdots & \cdots & 0 \\ 0 & \Sigma_{-\rho} & 0 & \ddots & \vdots \\ \vdots & 0 & \Sigma_\rho & 0 & \vdots \\ \vdots & \ddots & 0 & \Sigma_{-\rho} & \ddots \\ 0 & \cdots & \cdots & \ddots & \ddots \end{pmatrix}_{p\times p},$$

Σ_ρ is a $q \times q$ $(q \leq p)$ matrix, whose structure is $\Sigma_\rho = (\sigma_{ij})_{q\times q}$, where $\sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq q$, and $D_g = (\sigma_{1g}^2, \ldots, \sigma_{pg}^2)$ where $\sigma_{jg}^2 \sim (1/5)\chi_5^2$.

We generate the data from the multivariate normal distribution $N_p(\mathbf{0}, \Sigma_g)$. In our simulation studies, we compare our proposed approach with the sample covariance matrix under the setting of $p < n$. We investigate how the sample size and the number of classes affect the performance of our method, respectively. For the correlations in the block diagonal covariance matrices, we set $\rho = 0.2$, 0.4, 0.6 and 0.8, respectively. To compare different methods, we calculate the average matrix losses under L_1 and L_2 loss functions based on 1,000 simulations.

2.4.2 Simulation Results

First, we investigate the performance of our method under L_1 loss function. We set $(p, q) = (20, 10)$ for $n =$ 22, 30, 50 and 100. Figure 2.1 exhibit the average matrix

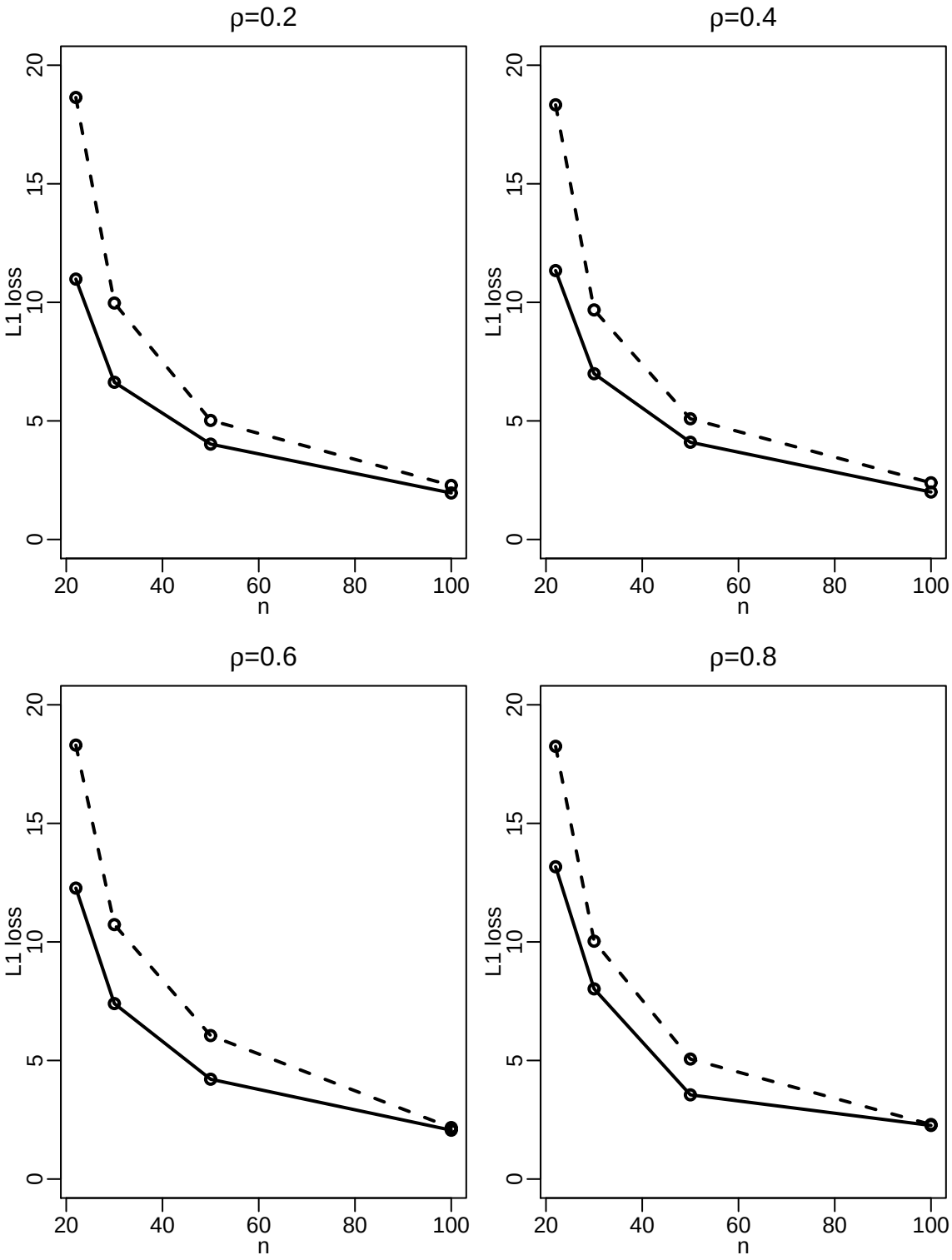

Figure 2.1: Average matrix losses of our method (solid lines) and $\mathbf{S}_n$ (dashed lines) against the sample size for L_1, $(p, q) = (20, 10)$ and $G = 2$.

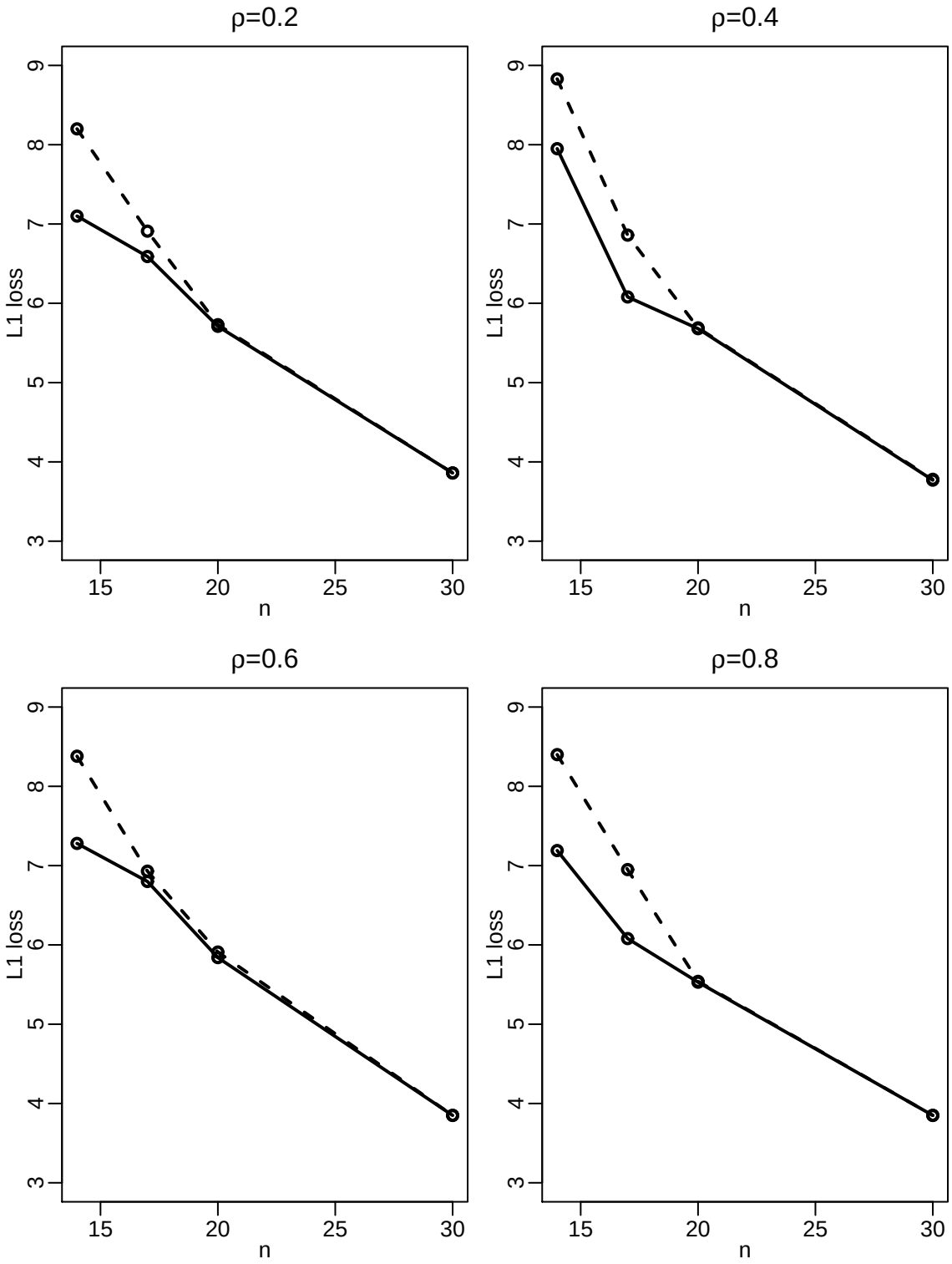

Figure 2.2: Average matrix losses of our method (solid lines) and $\mathbf{S}_n$ (dashed lines) against the sample size for L_2, $(p, q) = (10, 10)$ and $G = 4$.

losses of our method and the sample covariance matrix for $G = 2$. We observe that with an increasing number of observations, the average matrix losses for both methods are decreasing. When the sample size is large, both methods behave very similarly. However, when the sample size is close to the dimension, our proposed method shows its superiority over the sample covariance matrix.

Figure 2.2 investigate the performance of our method under L_2 loss function for $G = 4$. We set $(p, q) = (10, 10)$ for $n = 14$, 17, 20 and 30. From Figure (2.4.1) we also observe the decreasing average matrix losses with the increasing number of samples. Both approaches have similar performances under the case of large sample size and our approach performs better than the sample covariance matrix when the sample size is close to the dimension.

Besides, we consider the case of multiple class. Figure 2.3 shows the average matrix losses for an increasing number of classes. The left panel in Figure (2.4.2) exhibits the results of L_1 loss function and the right panel exhibits the results of L_2. Obviously, when the sample size is close to the dimension, an increasing number of classes will lead to a decreasing average matrix losses. However, the number of classes have little impact on the performance of our method under the case of large sample size. By simulations, we can also observe that once the sample size is much larger than the dimension, the estimated optimal shrinkage parameters are close to 1.

2.5 Real Data Analysis

In this section, we apply our proposed optimal shrinkage covariance matrices estimation to QDA. We use our method to estimate the covariance matrices in the discriminant scores and investigate the performance of classification. For real data set, we take advantage of Myeloma data (Zhan et al.; 2007). This data set includes two therapy groups who have multiple myeloma: Therapy 2 (TH2) with 351 samples and Therapy 3 (TH3) with 208 samples. The total number of genes is 54,675. This data set has been analyzed by Pang et al. (2009), and it can be downloaded from GEO Datasets using series number GSE2658.

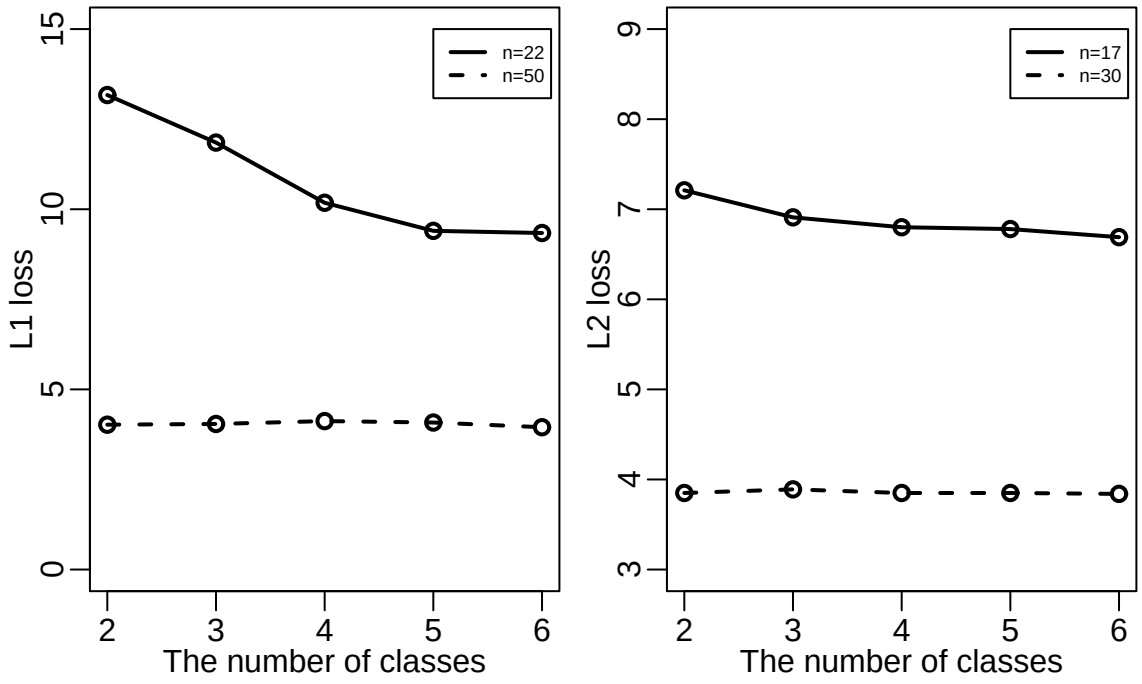

Figure 2.3: Average matrix losses of our method against the number of classes for $\rho = 0.8$

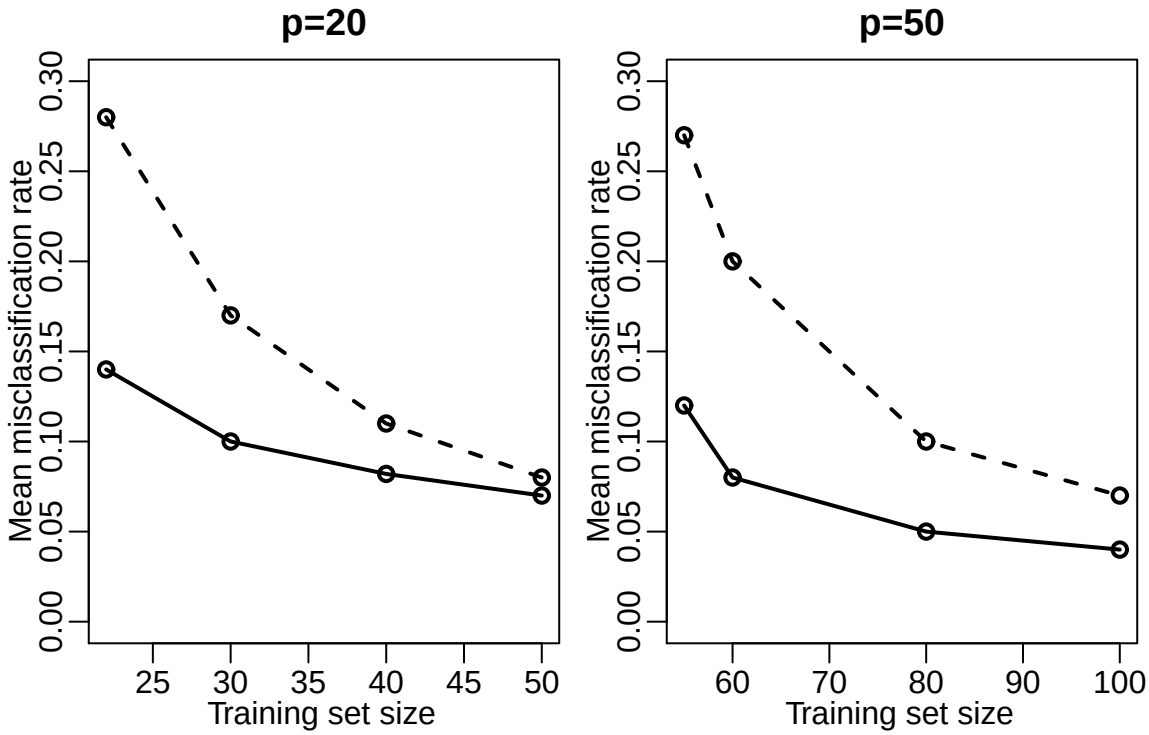

Figure 2.4: Mean misclassification rates of our method (solid lines) and $\mathbf{S}_n$ (dashed lines) against the training set size using Myeloma data.

In real data analysis, we compare our methods with the sample covariance matrix by computing the misclassification rate. First, we select top 20 and 50 genes using the gene screening method in Dudoit et al. (2002). They select top genes by ordering the ratio of between-class sums of squares to with-class. To compute the misclassification rate, we divide all samples into two sets randomly: the training set and the test set. For $p = 20$, we set the sample sizes in the training set are 22, 30, 40 and 50 for each groups, respectively. For $p = 50$, we set the sample sizes in the training set are 55, 60, 80 and 100 for each groups, respectively. All remaining samples are put into the test set. We compute the optimal shrinkage parameter under Stein Loss and repeat the whole procedure 1,000 times to compute the mean misclassification rates.

Figure 2.4 shows the mean misclassification rates and exhibits the superiority of our method in classification. From the figure, we can observe that the mean misclassification rates of our method are lower than the sample covariance matrix. When the sample size is large, our method performs similarly to the sample covariance matrix. However, when the sample size is close to the number of genes, the performance of our method is better.

2.6 Proofs

For any $p \times p$ matrix A, when the eigenvalues of A are all reals, we denote them as

$$\lambda_1(A) \geq \cdots \geq \lambda_p(A).$$

It is well known that if A is a symmetric matrix, then all its eigenvalues are reals, and further, we have $\lambda_p(A) > 0$ for any $A > 0$. By the Weilandt-Hoffman inequality (see, for instance, (1.67) in Tao (2012), page 55),

$$\sum_{i=1}^{p} |\lambda_i(A + B) - \lambda_i(A)|^2 \leq \|B\|^2 \tag{2.16}$$

holds for any symmetric matrices A and B.

Before the proof of theorems, we show some result on matrix calculation.

Lemma 1. *For any $p \times p$ matrices A and B, we have*

$$|tr(A)| \leq \sqrt{p}\,\|A\|, \quad \|AB\|^2 \leq \|A\|^2\|B\|^2.$$

Proof. By Cauchy-Schwarz inequality,

$$|\mathrm{tr}(A)| \leq \sum_{i=1}^{p} |a_{ii}| \leq \sqrt{p \sum_{i=1}^{p} a_{ii}^2} \leq \sqrt{p}\,\|A\|$$

and

$$\begin{aligned} \|AB\|^2 &= \sum_{i,j=1}^{p} (AB)_{ij}^2 = \sum_{i,j=1}^{p} \Big(\sum_{k=1}^{p} A_{ik} B_{kj} \Big)^2 \leq \sum_{i,j=1}^{p} \Big(\sum_{k=1}^{p} A_{ik}^2 \sum_{l=1}^{p} B_{lj}^2 \Big) \\ &= \sum_{i=1}^{p} \sum_{k=1}^{p} A_{ik}^2 \sum_{j=1}^{p} \sum_{l=1}^{p} B_{lj}^2 = \|A\|^2\|B\|^2. \end{aligned}$$

The proof of Lemma 1 is complete. □

Lemma 2. *Let $A > 0, B > 0$, then*

$$tr((A - B)(B^{-1} - A^{-1})) \geq 0, \tag{2.17}$$

where the equality holds true if and only if $A = B$.

Proof. By Exercise 12.28 in Abadir and Magnus (2005), page 338, we get (2.17), and the equality holds true if and only if $A^{1/2}B^{-1}A^{1/2} + A^{-1/2}BA^{-1/2} = 2I$.

Since $C := A^{1/2}B^{-1}A^{1/2} > 0$, there exists an orthogonal matrix P such that $C = P'\mathrm{diag}(\lambda_1(C), \cdots, \lambda_p(C))P$. If $C + C^{-1} = 2I$, then

$$2I = P'\mathrm{diag}(\lambda_1(C) + 1/\lambda_1(C), \cdots, \lambda_p(C) + 1/\lambda_p(C))P,$$

which implies $\lambda_1(C) = \cdots = \lambda_p(C) = 1$ and then $A = B$. Conversely, if $A = B$, then it is clear that $C + C^{-1} = 2I$.

Hence $A^{1/2}B^{-1}A^{1/2} + A^{-1/2}BA^{-1/2} = 2I$ is equivalent to $A = B$. □

Lemma 3. *Assume that $\alpha A + B > 0$ for any $\alpha \in [0, 1]$ and define*

$$f(\alpha) = \log\det(\alpha A + B),$$

then for any $\alpha \in [0, 1]$, we have $f''(\alpha) < 0$ and

$$f'(\alpha) = tr(A(\alpha A + B)^{-1}). \tag{2.18}$$

Proof. Let $\mu_1, \cdots, \mu_p$ be the eigenvalues of AB^{-1}. By noting that $AB^{-1} = AB^{-1/2}B^{-1/2}$ has the same eigenvalues as $B^{-1/2}AB^{-1/2}$, and $B^{-1/2}AB^{-1/2}$ is a symmetrical matrix, we know that $\mu_1, \cdots, \mu_p$ are positive reals and $I + B^{-1/2}AB^{-1/2}$ has eigenvalues $\{1+\mu_1, \cdots, 1+\mu_p\}$. Since $I + B^{-1/2}AB^{-1/2} > 0$, we have $1 + \mu_i > 0,\ i = 1, 2, \cdots, p$. Let

$$g(\alpha) := \log\det((\alpha A + B)B^{-1}) = \log\det(I + \alpha AB^{-1}) = \sum_{i=1}^{p} \log(1 + \alpha\mu_i),$$

then

$$f'(\alpha) = g'(\alpha) = \sum_{i=1}^{p} \frac{\mu_i}{1 + \alpha\mu_i}, \qquad f''(\alpha) = g''(\alpha) = -\sum_{i=1}^{p} \frac{\mu_i^2}{(1 + \alpha\mu_i)^2} < 0. \tag{2.19}$$

Next, we'll prove (2.18). For $\alpha = 0$, we have

$$f'(0) = \sum_{i=1}^{p} \mu_i = \mathrm{tr}(B^{-1/2}AB^{-1/2}) = \mathrm{tr}(AB^{-1}).$$

If $\alpha \neq 0$, then

$$\begin{aligned} f'(\alpha) &= \sum_{i=1}^{p} \frac{\mu_i}{1 + \alpha\mu_i} = \frac{p}{\alpha} - \sum_{i=1}^{p} \frac{1}{\alpha(1 + \alpha\mu_i)} \\ &= \frac{1}{\alpha}\mathrm{tr}(I - (I + \alpha B^{-1/2}AB^{-1/2})^{-1}) \\ &= \frac{1}{\alpha}\mathrm{tr}(I - B^{1/2}(B + \alpha A)^{-1}B^{1/2}) \\ &= \frac{1}{\alpha}\mathrm{tr}(I - B(\alpha A + B)^{-1}) = \mathrm{tr}(A(\alpha A + B)^{-1}). \end{aligned}$$

Hence (2.18) holds and proof of Lemma 7 is complete. □

Lemma 4. *For any $A, B, C > 0$ and $0 \leq \alpha \leq 1$, we have*

$$\begin{aligned} &(A - B)(\alpha A + (1 - \alpha)B)^{-1} - (A - C)(\alpha A + (1 - \alpha)C))^{-1} \\ &\quad = A(\alpha A + (1 - \alpha)B)^{-1}(C - B)(\alpha A + (1 - \alpha)C))^{-1}. \end{aligned} \tag{2.20}$$

Proof. For $\alpha = 1$, (2.20) is clear. When $\alpha \neq 1$, by noting that

$$\begin{aligned} &(A - B)(\alpha A + (1 - \alpha)B)^{-1} \\ &= \Big(\frac{1}{1 - \alpha}A - \frac{1}{1 - \alpha}\Big(\alpha A + (1 - \alpha)B\Big)\Big)(\alpha A + (1 - \alpha)B)^{-1} \\ &= \frac{1}{1 - \alpha}A(\alpha A + (1 - \alpha)B)^{-1} - \frac{1}{1 - \alpha}I, \end{aligned}$$

and, similarly,

$$(A-C)(\alpha A+(1-\alpha)C)^{-1}=\frac{1}{1-\alpha}A(\alpha A+(1-\alpha)C)^{-1}-\frac{1}{1-\alpha}I,$$

we have

$$\begin{aligned}
&(A-B)(\alpha A+(1-\alpha)B)^{-1}-(A-C)(\alpha A+(1-\alpha)C))^{-1}\\
&=\frac{1}{1-\alpha}A(\alpha A+(1-\alpha)B)^{-1}-\frac{1}{1-\alpha}A(\alpha A+(1-\alpha)C)^{-1}\\
&=\frac{1}{1-\alpha}A(\alpha A+(1-\alpha)B)^{-1}(\alpha A+(1-\alpha)C)(\alpha A+(1-\alpha)C)^{-1}\\
&\quad-\frac{1}{1-\alpha}A(\alpha A+(1-\alpha)B)^{-1}(\alpha A+(1-\alpha)B)(\alpha A+(1-\alpha)C)^{-1}\\
&=A(\alpha A+(1-\alpha)B)^{-1}(C-B)(\alpha A+(1-\alpha C))^{-1}.
\end{aligned}$$

The proof is complete. □

Lemma 5. *For any $A>0, B>0$, we have*

$$\sup_{0\le\alpha\le1}\|(\alpha A+(1-\alpha)B)^{-1}\|\le\sqrt{p}\,\|A^{-1}\|(1+\|A\|\,\|B^{-1}\|). \tag{2.21}$$

Proof. By the minimax formulae for eigenvalues (see, for instance, Theorem 1.3.2 in Tao (2012), page 49)

$$\begin{aligned}
\|(\alpha A+(1-\alpha)B)^{-1}\|^2&=\mathrm{tr}((\alpha A+(1-\alpha)B)^{-2})\\
&=\mathrm{tr}(A^{-1/2}WA^{-1}WA^{-1/2})\\
&\le p\max_{|v|=1}v^TA^{-1/2}WA^{-1}WA^{-1/2}v\\
&\le p\max_{|v|=1}v^TA^{-1}v\max_{|v|=1}\frac{v^TA^{-1/2}W^2A^{-1/2}v}{v^TA^{-1}v}\max_{|v|=1}\frac{v^TA^{-1/2}WA^{-1}WA^{-1/2}v}{v^TA^{-1/2}W^2A^{-1/2}v}\\
&\le p\max_{|v|=1}v^TA^{-1}v\max_{|v|=1}v^TW^2v\max_{|v|=1}v^TA^{-1}v\\
&=p(\lambda_1(A^{-1}))^2\lambda_1((\alpha I+(1-\alpha)A^{-1/2}BA^{-1/2})^{-2})\\
&=p(\lambda_1(A^{-1}))^2(\alpha+(1-\alpha)\lambda_p(A^{-1/2}BA^{-1/2}))^{-2},
\end{aligned}\tag{2.22}$$

where $W=W(\alpha,A,B):=(\alpha I+(1-\alpha)A^{-1/2}BA^{-1/2})^{-1}>0,\ v=(v_1,\cdots,v_n)\in\mathbb{R}^p, |v|^2=v_1^2+\cdots+v_n^2$ and v^T means the transposition of the column vector v. By using the similar method as that in (2.22), we can get that

$$\lambda_p(A^{-1/2}BA^{-1/2})=\min_{|v|=1}v^TA^{-1/2}BA^{-1/2}v\ge\min_{|v|=1}v^TBv\min_{|v|=1}v^TA^{-1}v=\lambda_p(B)\lambda_p(A^{-1}).$$

Hence

$$\begin{aligned}\sup_{0\le a\le 1}\|(\alpha A+(1-\alpha)B)^{-1}\| &\le \sqrt{p}\lambda_1(A^{-1})\sup_{0\le a\le 1}(\alpha+(1-\alpha)\lambda_p(B)\lambda_p(A^{-1}))^{-1}\\ &= \sqrt{p}\lambda_1(A^{-1})\max\Big(1,\frac{1}{\lambda_p(B)\lambda_p(A^{-1})}\Big)\\ &\le \sqrt{p}\lambda_1(A^{-1})(1+\lambda_1(A)\lambda_1(B^{-1}))\\ &\le \sqrt{p}\,\|A^{-1}\|(1+\|A\|\,\|B^{-1}\|),\end{aligned}$$

where we have used $\lambda_1(A)\le\|A\|$ for any $A>0$ since (2.16). The proof is complete.

□

2.6.1 Proof of Theorem 1

Proof. Now for any $g\in\{1,2,\cdots,G\}$, applying Lemma 7 with $A=\frac{S_g}{n}-\hat\Sigma_{\text{poo1}}, B=\hat\Sigma_{\text{poo1}}$, we have that for $\alpha\in[0,1]$,

$$\Big(\log\det\Big(\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat\Sigma_{\text{poo1}}\Big)\Sigma_g^{-1}\Big)\Big)''_\alpha=\Big(\log\det\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat\Sigma_{\text{poo1}}\Big)\Big)''_\alpha<0,$$

and

$$\begin{aligned}&\Big(\log\det\Big(\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat\Sigma_{\text{poo1}}\Big)\Sigma_g^{-1}\Big)\Big)'_\alpha\\ &\quad=\Big(\log\det\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat\Sigma_{\text{poo1}}\Big)\Big)'_\alpha\\ &\quad=\text{tr}\Big(\Big(\frac{S_g}{n}-\hat\Sigma_{\text{poo1}}\Big)\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat\Sigma_{\text{poo1}}\Big)^{-1}\Big).\end{aligned}$$

Hence

$$R_1''(\alpha,\hat\Sigma,\Sigma)=-\frac{1}{G}\sum_{g=1}^{G}E\Big(\log\det\Big(\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat\Sigma_{\text{poo1}}\Big)\Sigma_g^{-1}\Big)\Big)''_\alpha>0,$$

which implies that $R_1(\alpha,\hat\Sigma,\Sigma)$ is a strictly convex function on $[0,1]$, and

$$\begin{aligned}R_1'(\alpha,\hat\Sigma,\Sigma) &= p-\text{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g^{-1}\Big)\Big]\\ &\quad-\frac{1}{G}\sum_{g=1}^{G}\text{tr}E\Big(\Big(\frac{S_g}{n}-\hat\Sigma_{\text{poo1}}\Big)\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat\Sigma_{\text{poo1}}\Big)^{-1}\Big).\end{aligned}$$

Specially,

$$R_1'(\alpha,\hat{\Sigma},\Sigma)|_{\alpha=0} = p - \mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g^{-1}\Big)\Big], \tag{2.23}$$

and

$$\begin{aligned} R_1'(\alpha,\hat{\Sigma},\Sigma)|_{\alpha=1} &= \mathrm{tr}E\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}S_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}S_g^{-1}\Big)\Big] \\ &\quad -\mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g^{-1}\Big)\Big]. \end{aligned} \tag{2.24}$$

For any $g, g' \in \{1, 2, \cdots, G\}$, by Lemma 2, we get that

$$\mathrm{tr}(\Sigma_g - \Sigma_{g'})(\Sigma_{g'}^{-1} - \Sigma_g^{-1}) \geq 0,$$

which means

$$\mathrm{tr}(\Sigma_g\Sigma_{g'}^{-1} + \Sigma_{g'}\Sigma_g^{-1}) \geq 2p.$$

Thus

$$\mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g^{-1}\Big)\Big] \geq p. \tag{2.25}$$

This together with (2.23) implies $R_1'(\alpha,\hat{\Sigma},\Sigma)|_{\alpha=0} \leq 0$. And by Lemma 2, $R_1'(\alpha,\hat{\Sigma},\Sigma)|_{\alpha=0} = 0$ if and only if $\Sigma_g = \Sigma_{g'}$ for all $g, g' \in \{1, 2, \cdots, G\}$.

For any $g \in \{1, 2, \cdots, G\}$, since $S_g \sim W_p(\Sigma_g, n)$, we know that $ES_g = n\Sigma_g$, S_g^{-1} is an inverse Wishart random matrix with $S_g^{-1} \sim W_p^{-1}(\Sigma_g^{-1}, n)$ and $ES_g^{-1} = \frac{1}{n-p-1}\Sigma_g^{-1}$. Hence, for any $g, g' \in \{1, 2, \cdots, G\}$ with $g \neq g'$,

$$\begin{aligned} \mathrm{tr}(ES_{g'}S_g^{-1} - \Sigma_{g'}\Sigma_g^{-1}) &= \mathrm{tr}(ES_{g'}ES_g^{-1} - \Sigma_{g'}\Sigma_g^{-1}) = \frac{p+1}{n-p-1}\mathrm{tr}(\Sigma_{g'}\Sigma_g^{-1}) \\ &= \frac{p+1}{n-p-1}\mathrm{tr}(\Sigma_g^{-1/2}\Sigma_{g'}\Sigma_g^{-1/2}) > 0. \end{aligned}$$

This, together with (2.25) and (2.25), implies (b). □

2.6.2 Proof of Theorem 2

Proof. For any $g = 1, 2, \cdots, G$, we have that,

$$\frac{S_g}{n} \xrightarrow{L^1} \Sigma_g,$$

which implies that, for any $g \neq g'$,

$$\Big(\frac{S_g}{n} - \Sigma_g\Big)\Big(nS_{g'}^{-1} - \Sigma_{g'}^{-1}\Big) \xrightarrow{L^1} 0.$$

Thus

$$\begin{aligned} R_1'(\alpha, \hat{\Sigma}, \Sigma)|_{\alpha=1} &= \mathrm{tr}E\Big[\Big(\frac{1}{G}\sum_{g=1}^{G} S_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G} S_g^{-1}\Big)\Big] - \mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g^{-1}\Big)\Big] \\ &= \frac{1}{G^2}\sum_{g\neq g'}\mathrm{tr}E\Big[\Big(\frac{S_g}{n} - \Sigma_g\Big)\Big(nS_{g'}^{-1} - \Sigma_{g'}^{-1}\Big)\Big] \to 0. \end{aligned} \tag{2.26}$$

Now (ii) follows by the fact that $R_1(\alpha, \hat{\Sigma}, \Sigma)$ is a strictly convex function of α on $[0, 1]$ and

$$R_1'(\alpha, \hat{\Sigma}, \Sigma)|_{\alpha=0} = p - \mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g^{-1}\Big)\Big] = 0 \tag{2.27}$$

when $\Sigma_g = \Sigma_{g'}$ for any $g = g'$.

If Σ_g are not all the same, then by using SLLN and the dominated convergence theorem, we have that, for any $1 \leq \alpha \leq 1$,

$$\begin{aligned} R_1''(\alpha, \hat{\Sigma}, \Sigma) &= \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{p} E\frac{\Big(\lambda_i\Big(\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big)\Big)^2}{\Big(1 + \alpha\lambda_i\Big(\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big)\Big)^2} \\ &\geq \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{p} E\frac{\Big(\lambda_i\Big(\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big)\Big)^2}{2 + 2\Big(\lambda_i\Big(\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big)\Big)^2} \\ &\to \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{p}\frac{\Big(\lambda_i\Big(\bar{\Sigma}^{-1/2}(\Sigma_g - \bar{\Sigma})\bar{\Sigma}^{-1/2}\Big)\Big)^2}{2 + 2\Big(\lambda_i\Big(\bar{\Sigma}^{-1/2}(\Sigma_g - \bar{\Sigma})\bar{\Sigma}^{-1/2}\Big)\Big)^2} > 0. \end{aligned}$$

Now (i) follows by (2.26) and (2.28) □

2.6.3 Proof of Theorem 3

Proof. By SLLN, we have that, for any $g = 1, 2, \cdots, G$,

$$\frac{S_g}{n} \to \Sigma_g \quad a.s.$$

Thus

$$\mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G} S_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G} S_g^{-1}\Big)\Big] \to \mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G} \Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G} \Sigma_g^{-1}\Big)\Big] \quad a.s.,$$

and then

$$\hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)|_{\alpha=1} = \frac{p+1}{n}\mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G} S_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G} S_g^{-1}\Big)\Big] \to 0 \quad a.s.$$

By applying (2.19) and SLLN, and noting that $\lambda_i(\cdot)$ is a continuous function for any $i = 1, 2, \cdots, p$, we get that

$$\begin{aligned}
\hat{R}_1''(\alpha, \hat{\Sigma}, \Sigma) &= \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{p} \frac{\Big(\lambda_i\Big(\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big)\Big)^2}{\Big(1 + \alpha\lambda_i\Big(\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\hat{\Sigma}_{\mathrm{poo1}}^{-1/2}\Big)\Big)^2} \\
&\to \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{p} \frac{\Big(\lambda_i\Big(\bar{\Sigma}^{-1/2}(\Sigma_g - \bar{\Sigma})\bar{\Sigma}^{-1/2}\Big)\Big)^2}{\Big(1 + \alpha\lambda_i\Big(\bar{\Sigma}^{-1/2}(\Sigma_g - \bar{\Sigma})\bar{\Sigma}^{-1/2}\Big)\Big)^2} \geq 0,
\end{aligned}$$

where $\bar{\Sigma} = \frac{1}{G}\sum_{g=1}^{G} \Sigma_g$, and the equality holds if and only if $\Sigma_g = \Sigma_{g'}$ for any $g = g'$. Hence $\hat{\alpha}_1^* \to 1$ when Σ_g are not all the same. □

2.6.4 Proof of Theorem 4

Proof. By (2.6) and (2.7), we have

$$\sup_{0 \leq \alpha \leq 1} |R_1'(\alpha, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)| \leq \sum_{i=1}^{5} J_i,$$

where

$$J_1 = \Big|\mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\frac{S_g}{n}\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}(n-p-1)S_g^{-1}\Big)\Big] - \mathrm{tr}(E\Sigma_1 E\Sigma_1^{-1})\Big|$$

$$J_2 = \Big|\mathrm{tr}\Big[\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\Sigma_g^{-1}\Big)\Big] - \mathrm{tr}(E\Sigma_1 E\Sigma_1^{-1})\Big|$$

$$J_3 = \frac{1}{G}\sum_{g=1}^{G}\sup_{\alpha\in[0,1]}\Big|\mathrm{tr}\Big(\Big(\frac{S_g}{n}-\hat{\Sigma}_{\mathrm{poo1}}\Big)\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat{\Sigma}_{\mathrm{poo1}}\Big)^{-1}\Big) - \mathrm{tr}\Big(\Big(\frac{S_g}{n}-E\Sigma_1\Big)\Big(\alpha\frac{S_g}{n}+(1-\alpha)E\Sigma_1\Big)^{-1}\Big)\Big|$$

$$J_4 = \sup_{\alpha\in[0,1]}\Big|\frac{1}{G}\sum_{g=1}^{G}\mathrm{tr}\Big(\Big(\frac{S_g}{n}-E\Sigma_1\Big)\Big(\alpha\frac{S_g}{n}+(1-\alpha)E\Sigma_1\Big)^{-1}\Big) - E\mathrm{tr}\Big(\Big(\frac{S_1}{n}-E\Sigma_1\Big)\Big(\alpha\frac{S_1}{n}+(1-\alpha)E\Sigma_1\Big)^{-1}\Big)\Big|,$$

$$J_5 = \sup_{\alpha\in[0,1]}\Big|\mathrm{tr}E\Big(\Big(\frac{S_1}{n}-\hat{\Sigma}_{\mathrm{poo1}}\Big)\Big(\alpha\frac{S_1}{n}+(1-\alpha)\hat{\Sigma}_{\mathrm{poo1}}\Big)^{-1}\Big) - \mathrm{tr}E\Big(\Big(\frac{S_1}{n}-E\Sigma_1\Big)\Big(\alpha\frac{S_1}{n}+(1-\alpha)E\Sigma_1\Big)^{-1}\Big)\Big|.$$

In the following, we'll prove $J_i \to 0$ *a.s.* for $i = 1, \cdots, 5$.

Since

$$ES_g = E(E(S_g|\Sigma_g)) = nE\Sigma_1$$

and

$$(n-p+1)ES_g^{-1} = (n-p+1)E(E(S_g^{-1}|\Sigma_g)) = E\Sigma_1^{-1} > 0,$$

by the classical strong law of large number (SLLN), we have

$$\hat{\Sigma}_{\mathrm{pool}} = \frac{1}{G}\sum_{g=1}^{G}\frac{S_g}{n} \to E\Sigma_1 \ \ a.s., \qquad \frac{1}{G}\sum_{g=1}^{G}(n-p-1)S_g^{-1} \to E\Sigma_1^{-1} \ \ a.s.,$$

which, together with the fact that $\mathrm{tr}(\cdot)$ is a continuous function (by Lemma 1), implies that

$$J_1 \to 0 \ a.s.$$

Similarly,

$$J_2 \to 0 \ \ a.s.$$

By applying Lemma 4, we have

$$\Big(\frac{S_g}{n}-\hat{\Sigma}_{\text{poo1}}\Big)\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat{\Sigma}_{\text{poo1}}\Big)^{-1}-\Big(\frac{S_g}{n}-E\Sigma_1\Big)\Big(\alpha\frac{S_g}{n}+(1-\alpha)E\Sigma_1\Big)^{-1}$$
$$=\frac{S_g}{n}\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat{\Sigma}_{\text{poo1}}\Big)^{-1}\Big(E\Sigma_1-\hat{\Sigma}_{\text{poo1}}\Big)\Big(\alpha\frac{S_g}{n}+(1-\alpha)E\Sigma_1\Big)^{-1}. \quad (2.28)$$

Then, by Lemmas 1 and 5,

$$\begin{aligned} J_3 &\le \sqrt{p}\frac{1}{G}\sum_{g=1}^{G}\Big\|\frac{S_g}{n}\Big\|\sup_{a\in[0,1]}\Big\|\Big(\alpha\frac{S_g}{n}+(1-\alpha)\hat{\Sigma}_{\text{poo1}}\Big)^{-1}\Big\| \\ &\qquad\times\sup_{a\in[0,1]}\Big\|\Big(\alpha\frac{S_g}{n}+(1-\alpha)E\Sigma_1\Big)^{-1}\Big\|\Big\|E\Sigma_1-\hat{\Sigma}_{\text{poo1}}\Big\| \\ &\le (n\sqrt{p})^3\Big\|E\Sigma_1-\hat{\Sigma}_{\text{poo1}}\Big\|\times\frac{1}{G}\sum_{g=1}^{G}T(S_g,\hat{\Sigma}_{\text{poo1}}), \end{aligned} \quad (2.29)$$

where

$$\begin{aligned} & T(S_g,\hat{\Sigma}_{\text{poo1}}) \\ :=\ & \|S_g\|\,\|S_g^{-1}\|^2\Big(1+\|S_g\|\,\|(\hat{\Sigma}_{\text{poo1}})^{-1}\|\Big)\Big(1+\|S_g\|\,\|(E\Sigma_1)^{-1}\|\Big) \\ =\ & \|(\hat{\Sigma}_{\text{poo1}})^{-1}\|\Big(\|S_g\|^2\,\|S_g^{-1}\|^2\Big(1+\|S_g\|\,\|(E\Sigma_1)^{-1}\|\Big)\Big) \\ & \quad+\|S_g\|\,\|S_g^{-1}\|^2\Big(1+\|S_g\|\,\|(E\Sigma_1)^{-1}\|\Big). \end{aligned}$$

By noting that $\|(E\Sigma_1)^{-1}\|$ is a positive constant and recalling that $E\|S_1\|^7<\infty,\ E\|S_1^{-1}\|^7<\infty$, it is easy to get that

$$E\Big[\|S_1\|^2\,\|S_1^{-1}\|^2\Big(1+\|S_1\|\,\|(E\Sigma_1)^{-1}\|\Big)\Big]<\infty,$$
$$E\Big[\|S_1\|\,\|S_1^{-1}\|^2\Big(1+\|S_1\|\,\|(E\Sigma_1)^{-1}\|\Big)\Big]<\infty.$$

Hence by the SLLN, we have

$$\Big\|E\Sigma_1-\hat{\Sigma}_{\text{poo1}}\Big\|\to 0\ \ a.s.,\qquad \|(\hat{\Sigma}_{\text{poo1}})^{-1}\|\to\|(E\Sigma_1)^{-1}\|\ \ a.s.$$

and

$$\frac{1}{G}\sum_{g=1}^{G}T(S_g,\hat{\Sigma}_{\text{poo1}})\longrightarrow\Big(\|(E\Sigma_1)^{-1}\|\Big)\,E\Big[\|S_1\|^2\,\|S_1^{-1}\|^2\Big(1+\|S_1\|\,\|(E\Sigma_1)^{-1}\|\Big)\Big]$$
$$+E\Big[\|S_1\|\|S_1^{-1}\|^2\Big(1+\|S_1\|\|(E\Sigma_1)^{-1}\|\Big)\Big]<\infty\ \ a.s.$$

Thus

$$J_3 \to 0 \quad a.s.$$

By (2.28), we have

$$E\Big(\frac{S_1}{n} - \hat{\Sigma}_{\text{poo1}}\Big)\Big(\alpha\frac{S_1}{n} + (1-\alpha)\hat{\Sigma}_{\text{poo1}}\Big)^{-1} - E\Big(\frac{S_1}{n} - E\Sigma_1\Big)\Big(\alpha\frac{S_1}{n} + (1-\alpha)E\Sigma_1\Big)^{-1}$$
$$= E\Big[\frac{S_1}{n}\Big(\alpha\frac{S_1}{n} + (1-\alpha)\hat{\Sigma}_{\text{poo1}}\Big)^{-1}\Big(E\Sigma_1 - \hat{\Sigma}_{\text{poo1}}\Big)\Big(\alpha\frac{S_1}{n} + (1-\alpha)E\Sigma_1\Big)^{-1}\Big],$$

and, similarly to (2.29), we have

$$\begin{aligned} J_5 &\le (n\sqrt{p})^3 E\Big(T(S_1, \hat{\Sigma}_{\text{poo1}})\, \|E\Sigma_1 - \hat{\Sigma}_{\text{poo1}}\|\Big) \\ &\le (n\sqrt{p})^3\Big(1 + \|(E\Sigma_1)^{-1}\|\Big)^2 (J_{51} + J_{52}). \end{aligned}$$

where

$$J_{51} = E\Big[\|S_1\|\,\|S_1^{-1}\|^2\Big(1 + \|S_1\|\Big)^2\,\|E\Sigma_1 - \hat{\Sigma}_{\text{poo1}}\|\Big],$$
$$J_{52} = E\Big[\|S_1\|^2\,\|S_1^{-1}\|^2\Big(1 + \|S_1\|\Big)\|E\Sigma_1 - \hat{\Sigma}_{\text{poo1}}\|\,\|(\hat{\Sigma}_{\text{poo1}})^{-1}\|\Big].$$

By Hölder's inequality, we have

$$\begin{aligned} J_{52} \le\ & \Big(E\Big(\|S_1\|^2(1 + \|S_1\|)\Big)^{7/3}\Big)^{3/7} (E\|S_1^{-1}\|^7)^{2/7} \\ & \times (E\|E\Sigma_1 - \hat{\Sigma}_{\text{poo1}}\|^7)^{1/7}\, (E\|(\hat{\Sigma}_{\text{poo1}})^{-1}\|^7)^{1/7}. \end{aligned}$$

Clearly, by the assumption $E\|S_1\|^7 < \infty$, we have

$$E\Big(\|S_1\|^2(1 + \|S_1\|)\Big)^{7/3} < \infty,$$

and, by the L^p convergence theorem,

$$E\|E\Sigma_1 - \hat{\Sigma}_{\text{poo1}}\|^7 = E\Big\|\frac{1}{G}\sum_{g=1}^{G}\Big(\frac{S_g}{n} - \frac{ES_g}{n}\Big)\Big\|^7 \to 0.$$

Define a function $f : \mathbb{H}^+ \to \mathbb{R}$ with

$$f(A) = \|A^{-1}\|^2 = \text{tr}(A^{-2}) = \sum_{i=1}^{p}\frac{1}{(\lambda_i(A))^2}, \quad A > 0.$$

Since $x \mapsto 1/x^2$ is a convex function on $(0, \infty)$, by Klein's lemma (See, for instance, Guionnet (2009), page 78), f is a convex function on $\mathbb{H}^+$. Thus

$$\|(\hat{\Sigma}_{\text{poo1}})^{-1}\|^2 = \Big\|\Big(\frac{1}{G}\sum_{g=1}^{G}\frac{S_g}{n}\Big)^{-1}\Big\|^2 \leq \frac{1}{G}\sum_{g=1}^{G}\Big\|\Big(\frac{S_g}{n}\Big)^{-1}\Big\|^2.$$

Then, by Jensen's inequality,

$$E\|(\hat{\Sigma}_{\text{poo1}})^{-1}\|^7 \leq E\Big(\frac{1}{G}\sum_{g=1}^{G}\|nS_g^{-1}\|^2\Big)^{7/2} \leq E\Big(\frac{1}{G}\sum_{g=1}^{G}\|nS_g^{-1}\|^7\Big) = E\|nS_1^{-1}\|^7 < \infty.$$

Combing the above facts, we get that $J_{52} \to 0$. Similarly, $J_{51} \to 0$. Thus

$$J_5 \to 0.$$

Note that by Lemmas 1 and 5, we have that for any $0 \leq \alpha \leq 1$,

$$\begin{aligned}
&\Big|\text{tr}\Big(\Big(\frac{S_1}{n} - E\Sigma_1\Big)\Big(\alpha\frac{S_1}{n} + (1-\alpha)E\Sigma_1\Big)^{-1}\Big)\Big| \\
&\quad\leq \sqrt{p}\,\Big\|\frac{S_1}{n} - E\Sigma_1\Big\| \sup_{0\leq\alpha\leq 1}\Big\|\Big(\alpha\frac{S_1}{n} + (1-\alpha)E\Sigma_1\Big)^{-1}\Big\| \\
&\quad\leq np\,\Big(\|S_1\| + \|E\Sigma_1\|\Big)\,\|S_1^{-1}\|\Big(1 + \|S_1\|\,\|(E\Sigma_1)^{-1}\|\Big).
\end{aligned}$$

Since $E\Big[\Big(\|S_1\| + \|E\Sigma_1\|\Big)\,\|S_1^{-1}\|\Big(1 + \|S_1\|\,\|(E\Sigma_1)^{-1}\|\Big)\Big] < \infty$, by applying a uniform SLLN (see Theorem 16(a) in Ferguson (1996)), we have

$$J_4 \to 0 \quad a.s.$$

Now we get that

$$\sup_{0\leq\alpha\leq 1}|R_1'(\alpha, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)| \leq \sum_{i=1}^{5} J_i \to 0 \quad a.s.$$

As follows, we'll show that $\hat{\alpha}_1^* - \alpha_1^* \to 0$ $a.s.$ Since

$$R_1'(\hat{\alpha}_1^*, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\hat{\alpha}_1^*, \hat{\Sigma}, \Sigma) \geq R_1'(\alpha_1^*, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\hat{\alpha}_1^*, \hat{\Sigma}, \Sigma) \geq R_1'(\alpha_1^*, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\alpha_1^*, \hat{\Sigma}, \Sigma),$$

we have

$$|R_1'(\alpha^*, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\hat{\alpha}_1^*, \hat{\Sigma}, \Sigma)| \leq \sup_{0\leq\alpha\leq 1}|R_1'(\alpha, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)|.$$

Hence

$$
\begin{aligned}
&|\hat{R}_1'(\alpha_1^*, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\hat{\alpha}_1^*, \hat{\Sigma}, \Sigma)| \\
&\quad \le |R_1'(\alpha_1^*, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\alpha_1^*, \hat{\Sigma}, \Sigma)| + |R_1'(\alpha_1^*, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\hat{\alpha}_1^*, \hat{\Sigma}, \Sigma)| \\
&\quad \le 2 \sup_{0\le\alpha\le 1} |R_1'(\alpha, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\alpha, \hat{\Sigma}, \Sigma)| \to 0 \quad a.s..
\end{aligned}
$$

By the mean value theorem,

$$
|\alpha_1^* - \hat{\alpha}_1^*| \le |\hat{R}_1'(\alpha_1^*, \hat{\Sigma}, \Sigma) - \hat{R}_1'(\hat{\alpha}_1^*, \hat{\Sigma}, \Sigma)| / \inf_{0\le\alpha\le 1} \hat{R}_1''(\alpha, \hat{\Sigma}, \Sigma). \tag{2.30}
$$

Now, we'll estimate $\inf_{0\le\alpha\le 1} \hat{R}_1''(\alpha, \hat{\Sigma}, \Sigma)$. By applying (2.19) and using the similar method as that in (2.22), we get that

$$
\begin{aligned}
&\inf_{0\le\alpha\le 1} \hat{R}_1''(\alpha, \hat{\Sigma}, \Sigma) \\
= \;& \inf_{0\le\alpha\le 1} \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{p} \frac{\Big(\lambda_i\Big(\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big)\Big)^2}{\Big(1 + \alpha\lambda_i\Big(\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big)\Big)^2} \\
\ge \;& \frac{1}{G} \sum_{g=1}^{G} \sum_{i=1}^{p} \frac{\Big(\lambda_i\Big(\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big)\Big)^2}{2 + 2\Big(\lambda_i\Big(\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big)\Big)^2} \\
\ge \;& \frac{1}{4}\frac{1}{G} \sum_{g=1}^{G} \min\Big\{1, \lambda_p\Big(\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)\hat{\Sigma}_{\text{poo1}}^{-1}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)\hat{\Sigma}_{\text{poo1}}^{-1/2}\Big)\Big\} \\
\ge \;& \frac{1}{4}\frac{1}{G} \sum_{g=1}^{G} \min\Big\{1, \lambda_p\Big(\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)^2\Big)\Big(\lambda_p\Big(\hat{\Sigma}_{\text{poo1}}^{-1}\Big)\Big)^2\Big\} \\
\ge \;& \frac{1}{4} \min\Big\{1, \Big(\lambda_p\Big(\hat{\Sigma}_{\text{poo1}}^{-1}\Big)\Big)^2\Big\}\frac{1}{G} \sum_{g=1}^{G} \min\Big\{1, \Big(\lambda_p\Big(\hat{\Sigma}_{\text{poo1}}^{-1}\Big)\Big)^2\Big\}.
\end{aligned}
$$

By Lemma 1,

$$
\begin{aligned}
&\frac{1}{G} \sum_{g=1}^{G} \min\Big\{1, \lambda_p\Big(\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)^2\Big)\Big\} \\
\ge \;& \frac{1}{G} \sum_{g=1}^{G} \min\Big\{1, \lambda_p\Big(\Big(\frac{S_g}{n} - E\Sigma_1\Big)^2\Big)\Big\} - \frac{1}{G} \sum_{g=1}^{G} \Big\|\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)^2 - \Big(\frac{S_g}{n} - E\Sigma_1\Big)^2\Big\|.
\end{aligned}
$$

Note that

$$\begin{aligned}(A+B)^2-(A+C)^2 &= (A+B)^2-(A+B)(A+C)+(A+B)(A+C)-(A+C)^2\\ &= (A+B)(B-C)+(B-C)(A+C)\\ &= (A+C)(B-C)+(B-C)^2+(B-C)(A+C)\end{aligned}$$

holds for any $A, B, C > 0$. This, together with Lemma 1, SLLN and the fact that

$$E\Big\|\frac{S_1}{n}-E\Sigma_1\Big\| \le (1/n)E\|S_1\|+E\|E\Sigma_1\|<\infty,$$

yields that

$$\begin{aligned}&\frac{1}{G}\sum_{g=1}^{G}\Big\|\Big(\frac{S_g}{n}-\hat{\Sigma}_{\text{poo1}}\Big)^2-\Big(\frac{S_g}{n}-E\Sigma_1\Big)^2\Big\|\\ &\quad\le \Big\|E\Sigma_1-\hat{\Sigma}_{\text{poo1}}\Big\|^2+\frac{2}{G}\sum_{g=1}^{G}\Big\|E\Sigma_1-\hat{\Sigma}_{\text{poo1}}\Big\|\Big\|\frac{S_g}{n}-E\Sigma_1\Big\|\\ &\quad\to 0 \quad a.s.\end{aligned}$$

Thus, by noting that $\lambda_p(\cdot)$ is a continuous function (since (2.16)) and applying SLLN,

$$\begin{aligned}&\liminf_{G\to\infty}\inf_{0\le\alpha\le 1}\hat{R}_1''(\alpha,\hat{\Sigma},\Sigma)\\ \ge\ &\frac{1}{4}\min\{1,(\lambda_p((E\Sigma_1)^{-1}))^2\}E\min\Big\{1,\lambda_p\Big(\Big(\frac{S_1}{n}-E\Sigma_1\Big)^2\Big)\Big\}>0 \quad a.s.\end{aligned}$$

Now, by (2.30), we get $\hat{\alpha}_1^*-\alpha_1^*\to 0$ $a.s.$. □

2.6.5 Proof of Theorem 5

Proof. It is sufficient to show that $A_1 > 0$ and $-A_1 \le A_2 \le 0$. Define

$$V_g := \Sigma_g^{-1/2}\Big(\frac{S_g}{n}-\hat{\Sigma}_{\text{poo1}}\Big)\Sigma_g^{-1}\Big(\frac{S_g}{n}-\hat{\Sigma}_{\text{poo1}}\Big)\Sigma_g^{-1/2}.$$

Since $V_g \ge 0$, we have

$$E\text{tr}\Big(\Big(\frac{S_g}{n}-\hat{\Sigma}_{\text{poo1}}\Big)\Sigma_g^{-1}\Big)^2 = E\text{tr}(V_g)\ge 0,$$

and the equality holds if and only if $V_g = 0$ *a.s.* Note that $V_g = 0$ *a.s.* implies $S_g = \hat{\Sigma}_{\text{poo1}}$ *a.s.*, which is impossible. Hence $E\text{tr}\Big(\Big(\frac{S_g}{n} - \hat{\Sigma}_{\text{poo1}}\Big)\Sigma_g^{-1}\Big)^2 > 0$. Then $A_1 > 0$.

It follows from (2.12) that $-A_1 \leq A_2$. By Lemma 2, we know $\text{tr}(AB^{-1}+BA^{-1}) \geq 2p$ holds for any $A > 0, B > 0$. Thus, for any $g \neq g'$,

$$\begin{aligned}
&\text{tr}((\Sigma_{g'}\Sigma_g^{-1})^2) + \text{tr}((\Sigma_g\Sigma_{g'}^{-1})^2) = \text{tr}((\Sigma_{g'}\Sigma_g^{-1})^2) + \text{tr}(((\Sigma_{g'}\Sigma_g^{-1})^2)^{-1}) \geq 2p, \\
&(\text{tr}(\Sigma_{g'}\Sigma_g^{-1}))^2 + (\text{tr}(\Sigma_g\Sigma_{g'}^{-1}))^2 \geq \frac{1}{2}(\text{tr}(\Sigma_{g'}\Sigma_g^{-1}) + \text{tr}(\Sigma_g\Sigma_{g'}^{-1}))^2 \geq 2p^2.
\end{aligned}$$

Then, by (2.11) and (2.12),

$$\begin{aligned}
-A_2 \;&\geq\; \frac{(G-1)^2(p^2+p)}{nG^2} - \frac{(G-1)(p^2+p)}{nG} \\
&\qquad + \frac{1}{nG^3}\sum_{g=1}^{G}\sum_{g'\neq g}\Big(\text{tr}((\Sigma_{g'}\Sigma_g^{-1})^2) + (\text{tr}(\Sigma_{g'}\Sigma_g^{-1}))^2\Big) \\
&\geq\; \frac{(G-1)^2(p^2+p)}{nG^2} - \frac{(G-1)(p^2+p)}{nG} + \frac{G(G-1)(p^2+p)}{nG^3} = 0.
\end{aligned}$$

The proof of Theorem 5 is complete. □

2.6.6 Proof of Theorem 6

Proof. For any $g = 1, 2, \cdots, G$, we have that,

$$\frac{S_g}{n} \xrightarrow{L^2} \Sigma_g.$$

Then, as $n \to \infty$,

$$\begin{aligned}
A_1 \;&\to\; \frac{1}{G}\sum_{g=1}^{G}\text{tr}\Big(\Big(\Sigma_g - \bar{\Sigma}\Big)\Sigma_g^{-1}\Big(\Sigma_g - \bar{\Sigma}\Big)\Sigma_g^{-1}\Big) \\
&=\; \frac{1}{G}\sum_{g=1}^{G}\text{tr}\Big((I - \bar{\Sigma}\Sigma_g^{-1})^2\Big),
\end{aligned}$$

where $\bar{\Sigma} = \frac{1}{G}\sum_{g=1}^{G}\Sigma_g$. Similarly, as $n \to \infty$,

$$A_2 \to \frac{1}{G}\sum_{g=1}^{G}\text{tr}\Big((\Sigma_g - \bar{\Sigma})\Sigma_g^{-1}(\bar{\Sigma}\Sigma_g^{-1} - I)\Big) = -\frac{1}{G}\sum_{g=1}^{G}\text{tr}\Big((I - \bar{\Sigma}\Sigma_g^{-1})^2\Big),$$

and

$$A_3 \to \frac{1}{G}\sum_{g=1}^{G} \mathrm{tr}\Big(\bar{\Sigma}\Sigma_g^{-1} - I\Big)^2.$$

When Σ_g are not all the same, we have $\lim_{n\to\infty} A_1 > 0$ and then $\alpha_2^* = -A_2/A_1 \to 1$.

When $\Sigma_g = \Sigma_{g'}$ for any $g = g'$, we have

$$\lim_{n\to\infty} A_1 = -\lim_{n\to\infty} A_2 = \frac{1}{G}\sum_{g=1}^{G} \mathrm{tr}\Big((I - \bar{\Sigma}\Sigma_g^{-1})^2\Big) = 0.$$

Thus

$$\lim_{n\to\infty} R_2(\alpha, \hat{\Sigma}, \Sigma) = \lim_{n\to\infty} A_3 = \frac{1}{G}\sum_{g=1}^{G} \mathrm{tr}\Big(\bar{\Sigma}\Sigma_g^{-1} - I\Big)^2$$

is a constant function of α. □

2.6.7 Proofs of (2.11) and (2.12)

Proof. We have

$$\begin{aligned}
&E\mathrm{tr}\Big(\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\Sigma_g^{-1}\Big)^2 - \mathrm{tr}\Big(\Big(\Sigma_g - \bar{\Sigma}\Big)\Sigma_g^{-1}\Big)^2 \\
&= E\mathrm{tr}\Big(\Sigma_g^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}}\Big)\Sigma_g^{-1/2}\Big)^2 - \mathrm{tr}\Big(\Sigma_g^{-1/2}\Big(\Sigma_g - \bar{\Sigma}\Big)\Sigma_g^{-1/2}\Big)^2 \\
&= \sum_{i,j=1}^{p} E\Big(\Sigma_g^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}} - (\Sigma_g - \bar{\Sigma})\Big)\Sigma_g^{-1/2}\Big)_{ij}^2.
\end{aligned}$$

Let

$$S^{g,g'} = \Sigma_g^{-1/2} S_{g'} \Sigma_g^{-1/2}, \quad \Sigma^{g,g'} = \Sigma_g^{-1/2}\Sigma_{g'}\Sigma_g^{-1/2}, \qquad g' = 1, 2, \cdots,$$

then $S^{g,g'} \sim W_p(\Sigma^{g,g'}, n)$. By noting that $\mathrm{Var}(S_{ij}^{g,g'}) = n((\Sigma_{ij}^{g,g'})^2 + \Sigma_{ii}^{g,g'}\Sigma_{jj}^{g,g'})$, we have

$$\begin{aligned}
&E\Big(\Sigma_g^{-1/2}\Big(\frac{S_g}{n} - \hat{\Sigma}_{\mathrm{poo1}} - (\Sigma_g - \bar{\Sigma})\Big)\Sigma_g^{-1/2}\Big)_{ij}^2 \\
&= \frac{1}{n^2}\mathrm{Var}\Big(\frac{G-1}{G}S_{ij}^{g,g} - \frac{1}{G}\sum_{g'\neq g} S_{ij}^{g,g'}\Big) \\
&= \frac{(G-1)^2}{n^2G^2}\mathrm{Var}(S_{ij}^{g,g}) + \frac{1}{n^2G^2}\sum_{g'\neq g}\mathrm{Var}\Big(S_{ij}^{g,g'}\Big) \\
&= \frac{(G-1)^2}{nG^2}(\delta_{ij}+1) + \frac{1}{nG^2}\sum_{g'\neq g}((\Sigma_{ij}^{g,g'})^2 + \Sigma_{ii}^{g,g'}\Sigma_{jj}^{g,g'}).
\end{aligned}$$

Thus

$$E\mathrm{tr}\Big(\Big(\frac{S_g}{n}-\hat{\Sigma}_{\mathrm{poo1}}\Big)\Sigma_g^{-1}\Big)^2 \;=\; \mathrm{tr}\Big(\Big(\Sigma_g-\bar{\Sigma}\Big)\Sigma_g^{-1}\Big)^2+\frac{(G-1)^2(p^2+p)}{nG^2} + \frac{1}{nG^2}\sum_{g'\neq g}\Big(\mathrm{tr}((\Sigma_{g'}\Sigma_g^{-1})^2)+(\mathrm{tr}(\Sigma_{g'}\Sigma_g^{-1}))^2\Big).$$

Then we get (2.11) from (2.8).

For A_2, we have

$$A_2 \;=\; -A_1+\frac{1}{G}\sum_{g=1}^{G}E\mathrm{tr}\Big(\Big(\frac{S_g}{n}-\hat{\Sigma}_{\mathrm{poo1}}\Big)\Sigma_g^{-1}\Big(\Big(\frac{S_g}{n}\Big)\Sigma_g^{-1}-I\Big)\Big).$$

Since $ES_g=n\Sigma_g$ and $S_g, g=1,2,\cdots$ are independent, we have

$$\begin{aligned}
&E\mathrm{tr}\Big(\Big(\frac{S_g}{n}-\hat{\Sigma}_{\mathrm{poo1}}\Big)\Sigma_g^{-1}\Big(\Big(\frac{S_g}{n}\Big)\Sigma_g^{-1}-I\Big)\Big)\\
&=\frac{G-1}{nG}E\mathrm{tr}\Big(S_g\Sigma_g^{-1}\Big(\Big(\frac{S_g}{n}\Big)\Sigma_g^{-1}-I\Big)\Big)\\
&=\frac{G-1}{n^2G}E\mathrm{tr}(S^{g,g})^2-\frac{(G-1)p}{G}\\
&=\frac{(G-1)(p^2+p)}{nG},
\end{aligned}$$

where we have used

$$E\mathrm{tr}(S^{g,g})^2=\sum_{i,j=1}^{p}E(S_{i,j}^{g,g})^2=\sum_{i,j=1}^{p}(\mathrm{Var}(S_{i,j}^{g,g})+n^2\delta_{ij})=(n^2+n)p+np^2.$$

Hence

$$A_2=-A_1+\frac{(G-1)(p^2+p)}{nG},$$

and we get (2.12). □

2.6.8 Proof of (2.14)

Proof. Define

$$\tilde{S}:=(\Sigma_0)^{-1/2}S_g(\Sigma_0)^{-1/2},\quad \tilde{\Sigma}:=(\Sigma_0)^{-1/2}\Sigma_g(\Sigma_0)^{-1/2},$$

then $\tilde{S}\sim W_p(\tilde{\Sigma},n)$ and $\tilde{S}^{-1}$ has an inverse Wishart distribution, i.e. $\tilde{S}^{-1}\sim W_p^{-1}(\tilde{\Sigma}^{-1},n)$. Since (see, for instance, Letac and Massam (2004), page 308)

$$\begin{aligned}
E\tilde{S}^{-2}&=\frac{(n-p-1)\tilde{\Sigma}^{-2}+\tilde{\Sigma}^{-1}\mathrm{tr}(\tilde{\Sigma}^{-1})}{(n-p)(n-p-1)(n-p-3)},\\
E(\tilde{S}^{-1}\mathrm{tr}(\tilde{S}^{-1}))&=\frac{2\tilde{\Sigma}^{-2}+(n-p-2)\tilde{\Sigma}^{-1}\mathrm{tr}(\tilde{\Sigma}^{-1})}{(n-p)(n-p-1)(n-p-3)},
\end{aligned}$$

we have

$$(n-p-2)E\tilde{S}^{-2} - E(\tilde{S}^{-1}\mathrm{tr}(\tilde{S}^{-1})) = \frac{\tilde{\Sigma}^{-2}}{n-p-1}.$$

Then, by noting that $E(\tilde{S}^{-1}) = \frac{1}{n-p-1}\tilde{\Sigma}^{-1}$,

$$\begin{aligned}
&E[\mathrm{tr}(I-(n-p-1)\Sigma_0 S_g^{-1})^2 - (n-p-1)(\mathrm{tr}(\Sigma_0 S_g^{-1})^2 + (\mathrm{tr}(\Sigma_0 S_g^{-1}))^2)] \\
&= E[p - 2(n-p-1)\mathrm{tr}(\Sigma_0 S_g^{-1}) + (n-p-1)((n-p-2)\mathrm{tr}(\Sigma_0 S_g^{-1})^2 - (\mathrm{tr}(\Sigma_0 S_g^{-1}))^2)] \\
&= p - 2(n-p-1)\mathrm{tr}E(\tilde{S}^{-1}) + (n-p-1)((n-p-2)\mathrm{tr}E(\tilde{S}^{-2}) - (\mathrm{tr}E(\tilde{S}^{-1}))^2) \\
&= \mathrm{tr}\Big((\Sigma_g - \Sigma_0)\Sigma_g^{-1}\Big)^2.
\end{aligned}$$

where we have used that

$$\mathrm{tr}\Big((\Sigma_g - \Sigma_0)\Sigma_g^{-1}\Big)^2 = \mathrm{tr}(I - \tilde{\Sigma}^{-1})^2 = p - 2\mathrm{tr}(\tilde{\Sigma}^{-1}) + \mathrm{tr}(\tilde{\Sigma}^{-2}).$$

So we get (2.14). □

2.6.9 Proof of Theorem 7

Proof. For any $g = 1, 2, \cdots, G$, we have that,

$$\frac{S_g}{n} \to \Sigma_g \quad a.s.$$

When Σ_g are not all the same, we have that, as $n \to \infty$,

$$\begin{aligned}
\mathrm{tr}\Big(I - r\hat{\Sigma}_{\mathrm{poo1}} S_g^{-1}\Big)^2 &\to \mathrm{tr}\Big((I - \bar{\Sigma}\Sigma_g^{-1})^2\Big) > 0 \quad a.s., \\
r^2\Big(\mathrm{tr}(\hat{\Sigma}_{\mathrm{poo1}} S_g^{-1})^2 + (\mathrm{tr}(\hat{\Sigma}_{\mathrm{poo1}} S_g^{-1}))^2\Big) &\to \mathrm{tr}(\bar{\Sigma}\Sigma_g^{-1})^2 + (\mathrm{tr}(\bar{\Sigma}\Sigma_g^{-1})^2 \quad a.s., \\
\mathrm{tr}((S_{g'} S_g^{-1})^2) + (\mathrm{tr}(S_{g'} S_g^{-1}))^2 &\to \mathrm{tr}((\Sigma_{g'}\Sigma_g^{-1})^2) + (\mathrm{tr}(\Sigma_{g'}\Sigma_g^{-1}))^2 \quad a.s.
\end{aligned}$$

Thus, as $n \to \infty$,

$$\hat{A}_1 \to \frac{1}{G}\sum_{g=1}^{G} \mathrm{tr}\Big((I - \bar{\Sigma}\Sigma_g^{-1})^2\Big) > 0 \quad a.s.,$$

By (2.15), as $n \to \infty$,

$$\lim_{n\to\infty} \hat{A}_2 = -\lim_{n\to\infty} \hat{A}_1 \quad a.s.$$

Then $\hat{\alpha}_2^* \to 1$ *a.s.* □

2.6.10 Proof of Theorem 8

Proof. It is sufficient to prove that, as $G \to \infty$,

$$\hat{A} \to \mathrm{tr}E(I - (E\Sigma_1)\Sigma_1^{-1})^2 \quad a.s., \tag{2.31}$$

$$\frac{1}{G}\sum_{g=1}^{G} \mathrm{tr}\Big(\Big(\Sigma_g - \bar{\Sigma}\Big)\Sigma_g^{-1}\Big)^2 \to \mathrm{tr}E(I - (E\Sigma_1)\Sigma_1^{-1})^2 \quad a.s., \tag{2.32}$$

$$\frac{1}{G^3}\sum_{g=1}^{G}\sum_{g'\neq g}\Big(\mathrm{tr}((S_{g'}S_g^{-1})^2) + (\mathrm{tr}(S_{g'}S_g^{-1}))^2\Big) \to 0 \quad a.s., \tag{2.33}$$

$$\frac{1}{G^3}\sum_{g=1}^{G}\sum_{g'\neq g}\Big(\mathrm{tr}((\Sigma_{g'}\Sigma_g^{-1})^2) + (\mathrm{tr}(\Sigma_{g'}\Sigma_g^{-1}))^2\Big) \to 0 \quad a.s. \tag{2.34}$$

We only prove (2.31) and (2.33), the others' proofs are similar.

By SLLN and Lemma 1, as $G \to \infty$, we have

$$\begin{aligned}
&\frac{1}{G^3}\sum_{g=1}^{G}\sum_{g'\neq g}\Big(\mathrm{tr}((S_{g'}S_g^{-1})^2) + (\mathrm{tr}(S_{g'}S_g^{-1}))^2\Big) \\
&\quad\le \frac{2p}{G^3}\sum_{g=1}^{G}\sum_{g'\neq g}\|S_{g'}\|^2\,\|S_g^{-1}\|^2 \\
&\quad\le \frac{2p}{G}\Big(\frac{1}{G}\sum_{g=1}^{G}\|S_g\|^2\Big)\Big(\frac{1}{G}\sum_{g=1}^{G}\|S_g^{-1}\|^2\Big) \to 0 \quad a.s.
\end{aligned}$$

Hence we get (2.33).

By (2.14), we have

$$\begin{aligned}
E\Big(\mathrm{tr}\Big(I - r(E\Sigma_1)S_g^{-1}\Big)^2 - r\Big(\mathrm{tr}((E\Sigma_1)S_g^{-1})^2 + (\mathrm{tr}((E\Sigma_1)S_g^{-1}))^2\Big)\Big|\Sigma_g\Big) \\
= \mathrm{tr}(I - (E\Sigma_1)\Sigma_g^{-1})^2.
\end{aligned}$$

Thus, by SLLN,

$$\begin{aligned}
&\frac{1}{G}\sum_{g=1}^{G}\Big(\mathrm{tr}\Big(I - r(E\Sigma_1)S_g^{-1}\Big)^2 - r\Big(\mathrm{tr}((E\Sigma_1)S_g^{-1})^2 + (\mathrm{tr}((E\Sigma_1)S_g^{-1}))^2\Big)\Big) \\
&\quad\to \mathrm{tr}E(I - (E\Sigma_1)\Sigma_1^{-1})^2 \quad a.s.
\end{aligned}$$

In order to prove (2.31), it is sufficient to prove that

$$\frac{1}{G}\sum_{g=1}^{G}\Big(\mathrm{tr}\Big(I-r\hat{\Sigma}_{\mathrm{poo1}}S_g^{-1}\Big)^2-\mathrm{tr}\Big(I-r(E\Sigma_1)S_g^{-1}\Big)^2\Big)\to 0 \quad a.s., \tag{2.35}$$

$$\frac{1}{G}\sum_{g=1}^{G}\Big(\mathrm{tr}(\hat{\Sigma}_{\mathrm{poo1}}S_g^{-1})^2-\mathrm{tr}((E\Sigma_1)S_g^{-1})^2\Big)\to 0 \quad a.s., \tag{2.36}$$

$$\frac{1}{G}\sum_{g=1}^{G}\Big((\mathrm{tr}(\hat{\Sigma}_{\mathrm{poo1}}S_g^{-1}))^2-(\mathrm{tr}((E\Sigma_1)S_g^{-1}))^2\Big)\to 0 \quad a.s. \tag{2.37}$$

By noting that $\mathrm{tr}(A^2)-\mathrm{tr}(B^2)=\mathrm{tr}(A-B)(A+B)$ and applying SLLN and Lemma 1,

$$\begin{aligned}
&\Big|\frac{1}{G}\sum_{g=1}^{G}\Big(\mathrm{tr}\Big(I-r\hat{\Sigma}_{\mathrm{poo1}}S_g^{-1}\Big)^2-\mathrm{tr}\Big(I-r(E\Sigma_1)S_g^{-1}\Big)^2\Big)\Big| \\
&\quad\le \frac{r}{G}\sum_{g=1}^{G}\Big|\mathrm{tr}\Big(\Big((\hat{\Sigma}_{\mathrm{poo1}}-E\Sigma_1)S_g^{-1}\Big)\Big(2I-r(\hat{\Sigma}_{\mathrm{poo1}}+E\Sigma_1)S_g^{-1}\Big)\Big)\Big| \\
&\quad\le \frac{r\sqrt{p}}{G}\sum_{g=1}^{G}\Big\|\hat{\Sigma}_{\mathrm{poo1}}-E\Sigma_1\Big\|\,\Big\|S_g^{-1}\Big\|\,\Big(2\sqrt{p}+r\Big(\|\hat{\Sigma}_{\mathrm{poo1}}\|+\|E\Sigma_1\|\Big)\|S_g^{-1}\|\Big) \\
&\quad= \Big\|\hat{\Sigma}_{\mathrm{poo1}}-E\Sigma_1\Big\|\,\frac{2rp}{G}\sum_{g=1}^{G}\Big\|S_g^{-1}\Big\| \\
&\qquad\quad +\Big(\|\hat{\Sigma}_{\mathrm{poo1}}\|+\|E\Sigma_1\|\Big)\Big\|\hat{\Sigma}_{\mathrm{poo1}}-E\Sigma_1\Big\|\,\frac{r^2\sqrt{p}}{G}\sum_{g=1}^{G}\Big\|S_g^{-1}\Big\|^2\to 0 \quad a.s.
\end{aligned}$$

This proves (2.35). And we can similarly get (2.36) and (2.37).

Hence we get (2.31) and the proof of Theorem 8 is complete. □

Chapter 3

Shrinkage-Based Diagonal Hotelling's Tests for High-Dimensional Small Sample Size Data

3.1 Introduction

DNA microarrays allow us to acquire thousands or even tens of thousands of gene expression values simultaneously, which introduces novel approaches to genetic research. One important goal of analyzing gene expression microarray data is to detect differentially expressed genes. Recently, biologists and medical scientists have also recognized that testing the significance of gene sets or pathway analysis is an equally important problem (Efron and Tibshirani; 2007; Newton et al.; 2007; Chen and Qin; 2010; Maciejewski; 2014). Specifically, if we want to know whether a certain gene set, Z, is significantly differentially expressed in two different treatments, A and B, the testing hypothesis is $H_0 : \boldsymbol{\mu}_{\mathbf{Z}A} = \boldsymbol{\mu}_{\mathbf{Z}B}$, where $\boldsymbol{\mu}_{\mathbf{Z}A}$ and $\boldsymbol{\mu}_{\mathbf{Z}B}$ are the mean vectors of $\mathbf{Z}$ in A and B, respectively. In statistics, this is essentially a two-sample multivariate testing problem. One classical method used to solve such testing problems is Hotelling's T^2 test (Hotelling; 1931), which is a generalization of Student's t test.

This method works when the sample size, n, is larger than the data dimension, p. More generally, in a k-sample experiment, we are interested in whether or not there exist some differences among the k mean vectors of populations.

In this chapter, we focus on one-sample and two-sample multivariate testing problems for high-dimensional small sample size data, or equivalently, for "large p small n" data. In such settings, Hotelling's T^2 test suffers from a singularity problem in the covariance matrix estimation and therefore is not valid in this setting. To overcome the singularity problem, some remedies have been proposed in the literature; see, for example, the non-exact significance test and the randomization test in Dempster (1960). These approaches, however, are known to perform poorly in practice due to their complicated estimation of the degrees of freedom and some related issues (Bai and Saranadasa; 1996). In recent years, a number of approaches to improve Hotelling's T^2 test have emerged for testing high-dimensional data. In essence, these approaches can be classified into the following three categories, with the main difference among them how the covariance matrix is handled:

1) In the first category, the covariance matrix is removed from Hotelling's T^2 statistic to avoid the covariance matrix estimation. This idea was first considered by Bai and Saranadasa (1996). Specifically, they proposed to use $(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$ to replace $(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T\mathbf{S}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$ in Hotelling's T^2 statistic, where $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ are the sample mean vectors and $\mathbf{S}$ is the pooled sample covariance matrix. They demonstrated that the proposed test has better power than Hotelling's T^2 test under the requirement of p and n being of the same order. Recently, Zhang and Xu (2009) and Chen and Qin (2010) extended this method to "large p small n" data. We refer to the methods in this category as *the unscaled Hotelling's tests*.

2) In the second category, a regularization method is applied to the covariance matrix estimation to resolve the singularity problem. In this direction, Chen et al. (2011) have made a major contribution. They proposed a regularized Hotelling's T^2 test that estimates the covariance matrix by $\mathbf{S} + \lambda\mathbf{I}_p$, where $\mathbf{I}_p$

is the identity matrix and $\lambda > 0$ is a shrinkage parameter. This test works for both $p < n$ and $p \geq n$ cases. Note that a similar method was also proposed in Shen et al. (2011), where the form of $\lambda \mathbf{S} + (1 - \lambda)\mathbf{I}_p$ is used to estimate the covariance matrix with $0 \leq \lambda < 1$. In the special case of $\lambda = 0$, the test reduces to an unscaled Hotelling's test. We refer to the methods in this category as *the regularized Hotelling's tests*.

3) In the third category, the covariance matrix is assumed to be diagonal. Under this assumption, the singularity problem is circumvented since a diagonal matrix is always invertible for non-zero entries, whether or not p is larger than n. This idea was first considered by Wu et al. (2006) and then revisited by several other researchers; see, for example, Srivastava and Du (2008), Srivastava (2009), Park and Nag Ayyala (2013), and Srivastava et al. (2013). For more details, see Section 3.2.1 below. These methods are essentially all the same and we refer to them as *the diagonal Hotelling's tests*.

In our simulation studies, we note that the unscaled Hotelling's tests are often sensitive to the deviation of equal eigenvalues of the covariance matrix. If one eigenvalue is extremely large, then the performance of the test will be dominated by that individual component and thus a lower power will result. For more details, see the simulation studies in Section 3.4. In addition, even for the case of equal eigenvalues, Chen and Qin (2010) suggested $n = [20 \log(p)]$ to have a reasonably large power. For instance, n needs to be at least 46, 92 and 138 for $p = 10$, 100 and 1000, respectively. For high-dimensional data such as gene expression microarray data, however, it is not uncommon that n is very small, say for example less than 10 samples per group (Pomeroy et al.; 2002; Dong et al.; 2005). This has motivated researchers to consider more realistic testing methods for high-dimensional small sample size data, e.g., the regularized Hotelling's tests and the diagonal Hotelling's tests. Our additional simulation studies indicate that the existing regularized Hotelling's tests do not perform comparably to the diagonal Hotelling's tests when n is relatively small.

In view of the good performance of the diagonal Hotelling's tests, we also assume

that the covariance matrix is diagonal in this chapter. Before moving forward, we note that this diagonal covariance matrix assumption has been commonly used for high-dimensional small sample size data, e.g., Dudoit et al. (2002), Bickel and Levina (2004) and Tong and Wang (2007). In particular, Bickel and Levina (2004) pointed out that if the estimated correlations are all very noisy, then we are probably better off without estimating them. This, in essence, is the assumption of a diagonal covariance matrix when n is relatively small. In discriminant analysis, Lee et al. (2005a) have also observed that discriminant rules with an inverse generalized matrix may not perform as well as diagonal discriminant rules for microarray data. Although very promising, the performance of the diagonal Hotelling's tests themselves can be suboptimal due to the unreliable estimates of the sample variances from the limited number of observations. This suggests that some modifications to the diagonal Hotelling's tests are necessary to further improve their performance. We note that one such attempt has already been made by Dinu et al. (2007). They proposed a modified diagonal Hotelling's test, called "SAM-GS", by adding a small constant to each gene-specific variance estimate to stabilize the variance estimation, an idea originated in the SAM test of Tusher et al. (2001).

In this chapter, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. The test is structured by replacing the sample variances in the diagonal Hotelling's tests by the optimal shrinkage estimation of variances in Tong and Wang (2007). For the proposed shrinkage-based test, we then consider several different ways to derive the approximate null distribution under different scenarios of p and n. Simulation results show that the proposed method always performs comparably to existing competitors, especially when n is less than 10. In addition, to assess the performance of the proposed method using real data, we consider four gene expression data sets. A case study also demonstrates the advantage of the proposed shrinkage-based diagonal Hotelling's test.

The remainder of this chapter is organized as follows. The shrinkage-based diagonal Hotelling's tests are introduced in Section 3.2. In Section 3.3, we derive both a scaled chi-squared null distribution and a normal null distribution. Simulation stud-

ies and real data analysis are conducted in Sections 3.4 and 3.5, respectively. Section 3.6 provide some discussions and Section 3.7 give some proofs.

3.2 Improving the Diagonal Hotelling's Tests

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T,\ i = 1, \ldots, n$, be i.i.d. random vectors from a multivariate normal distribution, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ is the population mean vector and $\boldsymbol{\Sigma}$ is the population covariance matrix. Let also $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$ be the sample mean vector and $\mathbf{S} = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T/(n-1)$ be the sample covariance matrix. For the one-sample testing problem, the hypothesis is

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \tag{3.1}$$

where $\boldsymbol{\mu}_0$ is a fixed vector. To test hypothesis (3.1), the one-sample Hotelling's T^2 statistic is defined as

$$T_1^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0). \tag{3.2}$$

When $p \leq n-1$ so that $\mathbf{S}$ is invertible, under H_0, the scaled test statistic, $\{(n-p)/p(n-1)\}T_1^2$, follows an $F_{p,n-p}$ distribution with p and $n-p$ degrees of freedom.

For the two-sample testing problem, similarly, we assume that $\mathbf{X}_{ki} = (X_{ki1}, \ldots, X_{kip})^T$, $i = 1, \ldots, n_k$, are i.i.d. from a multivariate normal distribution, $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, for $k = 1$ and 2, respectively, where $\boldsymbol{\mu}_k$ are the population mean vectors and $\boldsymbol{\Sigma}$ is the common covariance matrix. Let $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ denote the class-specific sample means, and let $\mathbf{S}_{pool} = \{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2\}/(n_1+n_2-2)$ be the pooled sample covariance matrix. Then, to test the hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2, \tag{3.3}$$

the two-sample Hotelling's T^2 statistic is given by

$$T_2^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \mathbf{S}_{pool}^{-1} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2). \tag{3.4}$$

When $p \leq n_1 + n_2 - 2$ so that $\mathbf{S}_{pool}$ is invertible, under H_0, the scaled test statistic, $\{(n_1+n_2-p-1)/p(n_1+n_2-2)\}T_2^2$, follows an F_{p,n_1+n_2-p-1} distribution with p and n_1+n_2-p-1 degrees of freedom.

3.2.1 The Diagonal Hotelling's Tests

We note that Hotelling's T^2 statistics require $n-1 \geq p$ for the one-sample case and $n_1+n_2-2 \geq p$ for the two-sample case to ensure that the sample covariance matrix is nonsingular. Hence, these methods do not work under the "large p small n" paradigm. To avoid the singularity problem, Wu et al. (2006) proposed a pooled component test for the two-sample case, which essentially is a diagonal version of Hotelling's T^2 statistic (3.4). Specifically, their proposed test statistic is

$$\begin{aligned} T^2_{\mathrm{D2}} &= \frac{n_1 n_2}{n_1+n_2}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2)^T\{\mathrm{diag}(\mathbf{S}_{pool})\}^{-1}(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2) \\ &= \frac{n_1 n_2}{n_1+n_2}\sum_{j=1}^{p}(\bar{X}_{1j}-\bar{X}_{2j})^2/s^2_{j,pool}, \end{aligned} \tag{3.5}$$

where $\mathrm{diag}(\mathbf{S}_{pool}) = \mathrm{diag}(s^2_{1,pool},\ldots,s^2_{p,pool})$ with $s^2_{j,pool} = \{(n_1-1)s^2_{1j}+(n_2-1)s^2_{2j})\}/(n_1+n_2-2)$ for $j=1,\ldots,p$. Note that, for simplicity, the missing data problem is not considered in (3.5) unlike in Wu et al. (2006). Although only the two-sample case was considered in their paper, the diagonal idea can be readily extended to the one-sample case with T^2_{D1} taking the form

$$\begin{aligned} T^2_{\mathrm{D1}} &= n(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)^T\{\mathrm{diag}(\mathbf{S})\}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0) \\ &= n\sum_{j=1}^{p}(\bar{X}_j-\mu_{0j})^2/s^2_j. \end{aligned}$$

To define the rejection region of the diagonal Hotelling's tests, we reject H_0 if $T^2_{\mathrm{D1}} > C_1$ for the one-sample case and $T^2_{\mathrm{D2}} > C_2$ for the two-sample case, where C_1 and C_2 are two critical values.

Besides the aforementioned pooled components test, Srivastava and Du (2008) proposed scalar transformation invariant tests for both one-sample and two-sample cases, which essentially are the functions of the diagonal Hotelling's test. Srivastava (2009) constructed the test statistic using the diagonal Hotelling's test under non-normality. This test statistic is similar to Srivastava and Du (2008), and the only difference is that Srivastava (2009) deleted the adjustment coefficient appearing in Srivastava and Du (2008). Park and Nag Ayyala (2013) proposed new scalar transformation invariant tests for both one-sample and two-sample cases. Their tests

modified the test statistic of Srivastava (2009). This test statistic still assumes that the covariance matrix is diagonal. Srivastava et al. (2013) proposed a two-sample test under the condition of unequal covariance matrices, which essentially is also a function of the diagonal Hotelling's test.

3.2.2 Shrinkage-Based Diagonal Hotelling's Tests

To establish the diagonal Hotelling's tests, the aforementioned references used the diagonal matrix of sample variances to estimate the covariance matrix. However, when the number of observations is limited, such as when there are fewer than 10 observations, the sample variances are not reliable estimations any more, and the diagonal Hotelling's tests are thus unreliable. This point has been demonstrated by the simulation studies of Srivastava et al. (2013). Therefore, it is necessary to find an improved variance estimation. Dinu et al. (2007) made such an attempt. In this section, we use the optimal shrinkage estimation in Tong and Wang (2007) to improve the variance estimation.

Let diag($\mathbf{\Sigma}$)=diag($\sigma_1^2, \ldots, \sigma_p^2$), and $(\sigma_j^2)^t = \sigma_j^{2t}$, $j = 1, \ldots, p$. The shrinkage estimator of σ_j^{2t} is

$$\tilde{\sigma}_j^{2t}(\alpha) = \{h_{\nu,p}(t)\hat{\sigma}_{pool}^{2t}\}^{\alpha}\{h_{\nu,1}(t)\hat{\sigma}_j^{2t}\}^{1-\alpha}, \tag{3.6}$$

where $\hat{\sigma}_j^{2t}$ estimates σ_j^{2t}, $\hat{\sigma}_{pool}^{2t} = \prod_{j=1}^{p}(\hat{\sigma}_j^2)^{t/p}$, $\nu = n - 1$, $\Gamma(\cdot)$ is the gamma function, and

$$h_{\nu,p}(t) = \left(\frac{\nu}{2}\right)^t \left\{\frac{\Gamma(\nu/2)}{\Gamma(\nu/2 + t/p)}\right\}^p.$$

This shrinkage estimator includes a shrinkage parameter $\alpha \in [0,1]$. The estimator degenerates to the unbiased estimation of σ_j^{2t} if $\alpha = 0$, and it shrinks to the pooled variance estimation if $\alpha = 1$. Under the Stein loss function, $L(\sigma^2, \tilde{\sigma}^2) = \tilde{\sigma}^2/\sigma^2 - \ln(\tilde{\sigma}^2/\sigma^2) - 1$, Tong and Wang (2007) proved that there exists a unique optimal α in (0,1], denoted by α^*, to achieve the minimum average risk for any fixed p, ν and $t > -\nu/2$. A two-step procedure, proposed by Tong and Wang (2007), can be useful to estimate α^*.

Now, we move to the shrinkage-based diagonal Hotelling's tests. For ease of distinction, let σ_j^2 and $\sigma_{j,pool}^2$ denote the jth variance in diag($\mathbf{\Sigma}$) for one-sample and two-sample cases, respectively. Then, $\hat{\sigma}_j^2 = s_j^2$ and $\hat{\sigma}_{j,pool}^2 = s_{j,pool}^2$. Moreover, since the sample variances appear in the denominator of the diagonal Hotelling's tests, we consider estimating $\sigma_j^{-2} = 1/\sigma_j^2$ instead of σ_j^2, which is the case of $t = -1$. Pang et al. (2009) found that results for $t = 1$ and $t = -1$ were similar with the latter slightly better. Therefore, we focus on estimating σ_j^{-2} in the remaining of this chapter.

Let $\tilde{\alpha}^*$ denote the estimated optimal shrinkage parameter. For the one-sample test, we define the shrinkage-based diagonal Hotelling's test statistic as

$$\begin{aligned} T^2_{\text{SD1}}(\tilde{\alpha}^*) &= n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T \tilde{\mathbf{S}}^* (\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \\ &= n \sum_{j=1}^{p} (\bar{X}_j - \mu_{0j})^2 \tilde{\sigma}_j^{-2}(\tilde{\alpha}^*), \end{aligned} \tag{3.7}$$

where $\tilde{\mathbf{S}}^* = \text{diag}\{\tilde{\sigma}_1^{-2}(\tilde{\alpha}^*), \ldots, \tilde{\sigma}_p^{-2}(\tilde{\alpha}^*)\}$. Similarly, for the two-sample test, the shrinkage-based diagonal Hotelling's test statistic is

$$\begin{aligned} T^2_{\text{SD2}}(\tilde{\alpha}^*) &= \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \tilde{\mathbf{S}}^*_{pool} (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) \\ &= \frac{n_1 n_2}{n_1 + n_2} \sum_{j=1}^{p} (\bar{X}_{1j} - \bar{X}_{2j})^2 \tilde{\sigma}^{-2}_{j,pool}(\tilde{\alpha}^*), \end{aligned} \tag{3.8}$$

where $\tilde{\mathbf{S}}^*_{pool} = \text{diag}\{\tilde{\sigma}^{-2}_{1,pool}(\tilde{\alpha}^*), \ldots, \tilde{\sigma}^{-2}_{p,pool}(\tilde{\alpha}^*)\}$.

Tong and Wang (2007) showed that $\tilde{\alpha}^* \to 0$ for $n \to \infty$ and fixed p. This property demonstrates that, when the sample size is very large, on the one hand, our methods degenerate to the diagonal Hotelling's tests and thus it is unnecessary to borrow information from other genes. Our simulation studies indicate that our methods perform comparably to current approaches, and the diagonal Hotelling's tests are hence appropriate for testing the significance of gene sets. On the other hand, the approximate null distributions in the diagonal Hotelling's tests can also be used. However, when the sample size is small, the above approximations may not be accurate. Hence, it is of great importance to derive the approximate null distribution in the case of small sample sizes.

3.3 Null Distributions of Shrinkage-Based Diagonal Hotelling's Tests for Small Sample Size

When the sample size is large, two types of distributions are derived to be the approximate null distributions. One is the chi-squared distribution. Wu et al. (2006) considered a scaled chi-squared distribution as an approximation for both $p < n$ and $p \geq n$. If $(n, p) \to \infty$, the other possible choice, the normal distribution, is used as the asymptotic null distribution (Srivastava and Du; 2008; Srivastava; 2009; Park and Nag Ayyala; 2013; Srivastava et al.; 2013). This motivates us to derive approximate null distributions similarly when the sample size is very small. In this section, we follow Wu et al. (2006) and derive the scaled chi-squared null distribution, and the normal null distribution is derived as $p \to \infty$.

To obtain the null distributions of the shrinkage-based diagonal Hotelling's tests, we first derive the means and variances for $T^2_{\text{SD1}}(\alpha)$ and $T^2_{\text{SD2}}(\alpha)$ in Lemmas 1 and 2, respectively.

Lemma 6. *For any $\nu = n - 1 > 4$ and $\alpha \in (0, 1]$, the mean and variance of the test statistic, T^2_{SD1}, under H_0 are*

$$E\left\{T^2_{\text{SD1}}(\alpha)\right\} = C_1 \sigma_{pool}^{-2\alpha} \sum_{j=1}^{p} \sigma_j^{2\alpha},$$

and

$$Var\left\{T^2_{\text{SD1}}(\alpha)\right\} = (3C_2 - C_3)\sigma_{pool}^{-4\alpha} \sum_{j=1}^{p} \sigma_j^{4\alpha} + (C_3 - C_1^2)\sigma_{pool}^{-4\alpha} (\sum_{j=1}^{p} \sigma_j^{2\alpha})^2,$$

where

$$C_1 = \frac{h^{\alpha}_{\nu,p}(-1) h^{1-\alpha}_{\nu,1}(-1)}{h^{p-1}_{\nu,1}(-\alpha/p) h_{\nu,1}\{-\alpha/p - (1-\alpha)\}},$$

$$C_2 = \frac{h^{2\alpha}_{\nu,p}(-1) h^{2(1-\alpha)}_{\nu,1}(-1)}{h^{p-1}_{\nu,1}(-2\alpha/p) h_{\nu,1}\{-2\alpha/p - 2(1-\alpha)\}},$$

$$C_3 = \frac{h^{2\alpha}_{\nu,p}(-1) h^{2(1-\alpha)}_{\nu,1}(-1)}{h^{p-2}_{\nu,1}(-2\alpha/p) h^2_{\nu,1}\{-2\alpha/p - (1-\alpha)\}}.$$

Lemma 7. *For any $\nu = n_1 + n_2 - 2 > 4$ and $\alpha \in (0, 1]$, the mean and variance of the test statistic, T^2_{SD2}, under H_0 are*

$$E\left\{T^2_{\text{SD2}}(\alpha)\right\} = C_1 \sigma_{pool}^{-2\alpha} \sum_{j=1}^{p} \sigma_{j,pool}^{2\alpha},$$

and

$$Var\left\{T^2_{\text{SD2}}(\alpha)\right\} = (3C_2 - C_3)\sigma_{pool}^{-4\alpha} \sum_{j=1}^{p} \sigma_{j,pool}^{4\alpha} + (C_3 - C_1^2)\sigma_{pool}^{-4\alpha} (\sum_{j=1}^{p} \sigma_{j,pool}^{2\alpha})^2,$$

where

$$C_1 = \frac{h^{\alpha}_{\nu,p}(-1) h^{1-\alpha}_{\nu,1}(-1)}{h^{p-1}_{\nu,1}(-\alpha/p) h_{\nu,1}\{-\alpha/p - (1-\alpha)\}},$$

$$C_2 = \frac{h^{2\alpha}_{\nu,p}(-1) h^{2(1-\alpha)}_{\nu,1}(-1)}{h^{p-1}_{\nu,1}(-2\alpha/p) h_{\nu,1}\{-2\alpha/p - 2(1-\alpha)\}},$$

$$C_3 = \frac{h^{2\alpha}_{\nu,p}(-1) h^{2(1-\alpha)}_{\nu,1}(-1)}{h^{p-2}_{\nu,1}(-2\alpha/p) h^2_{\nu,1}\{-2\alpha/p - (1-\alpha)\}}.$$

The proof of Lemma 6 is given in Appendix 1. The proof of Lemma 7 is omitted since it is essentially the same as that for Lemma 6. Both lemmas are however necessary for determining the parameters of the approximate null distributions.

3.3.1 Chi-squared Approximation

For small p, the chi-squared distribution can be used as a good approximate null distribution. In this section, we approximate the null distributions of the proposed test statistics as a scaled chi-squared distribution, $c\chi^2_d$, as in Wu et al. (2006). To determine the scale parameter, c_1, and the degrees of freedom, d_1, for $T^2_{\text{SD1}}(\tilde{\alpha}^*)$, we equate the mean and variance of $c_1\chi^2_{d_1}$ with the mean and variance of $T^2_{\text{SD1}}(\tilde{\alpha}^*)$. Specifically, we have

$$E\{T^2_{\text{SD1}}(\tilde{\alpha}^*)\} = c_1 d_1 \qquad \text{and} \qquad \text{Var}\{T^2_{\text{SD1}}(\tilde{\alpha}^*)\} = 2c_1^2 d_1.$$

For $T^2_{\text{SD2}}(\tilde{\alpha}^*)$, we use the same approach to determine the corresponding scale parameter, c_2, and the degrees of freedom, d_2. The following theorems describe the approximate null distributions for our test statistics.

Theorem 9. *For any $n > 5$ and optimal shrinkage parameter estimation, $\tilde{\alpha}^*$, under the null hypothesis, we have*

$$T^2_{\mathrm{SD1}}(\tilde{\alpha}^*) \sim c_1 \chi^2_{d_1},$$

where

$$c_1 = \frac{(3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^p \sigma_j^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^p \sigma_j^{2\tilde{\alpha}^*})^2}{2C_1 \sigma_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^p \sigma_j^{2\tilde{\alpha}^*}},$$
$$d_1 = \frac{2C_1^2 \sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^p \sigma_j^{2\tilde{\alpha}^*})^2}{(3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^p \sigma_j^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^p \sigma_j^{2\tilde{\alpha}^*})^2}.$$

Theorem 10. *For any $n_1 + n_2 > 6$ and optimal shrinkage parameter estimation, $\tilde{\alpha}^*$, under the null hypothesis, we have*

$$T^2_{\mathrm{SD2}}(\tilde{\alpha}^*) \sim c_2 \chi^2_{d_2},$$

where

$$c_2 = \frac{(3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^p \sigma_{j,pool}^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^p \sigma_{j,pool}^{2\tilde{\alpha}^*})^2}{2C_1 \sigma_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^p \sigma_{j,pool}^{2\tilde{\alpha}^*}},$$
$$d_2 = \frac{2C_1^2 \sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^p \sigma_{j,pool}^{2\tilde{\alpha}^*})^2}{(3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^p \sigma_{j,pool}^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*} (\sum_{j=1}^p \sigma_{j,pool}^{2\tilde{\alpha}^*})^2}.$$

The proofs of Theorems 9 and 10 are simple and straightforward. They are thus omitted. Note that c_1, d_1, c_2 and d_2 involve some unknown quantities. Take c_1 and d_1 for example. Then, $b_1(\boldsymbol{\sigma}^2) = \sigma_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^p \sigma_j^{2\tilde{\alpha}^*}$ and $b_2(\boldsymbol{\sigma}^2) = \sigma_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^p \sigma_j^{4\tilde{\alpha}^*}$ are the unknown quantities. In practice, we suggest the following rules for estimating $b_1(\boldsymbol{\sigma}^2)$ and $b_2(\boldsymbol{\sigma}^2)$, according to the different scenarios:

(i) For any fixed p but large n, by noting that $\hat{\sigma}_j^2 \xrightarrow{a.s.} \sigma_j^2$ as $n \to \infty$, where $\xrightarrow{a.s.}$ denotes the almost sure convergence, we have the following consistent estimators:

$$\hat{b}_1(\boldsymbol{\sigma}^2) = \hat{\sigma}_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^p \hat{\sigma}_j^{2\tilde{\alpha}^*} \quad \text{and} \quad \hat{b}_2(\boldsymbol{\sigma}^2) = \hat{\sigma}_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^p \hat{\sigma}_j^{4\tilde{\alpha}^*};$$

(ii) For any fixed n but large p, by Lemma 2 of Tong and Wang (2007), we estimate

$$\breve{b}_1(\boldsymbol{\sigma}^2) = w(\tilde{\alpha}^*)\hat{\sigma}_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^p \hat{\sigma}_j^{2\tilde{\alpha}^*} \quad \text{and} \quad \breve{b}_2(\boldsymbol{\sigma}^2) = w(2\tilde{\alpha}^*)\hat{\sigma}_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^p \hat{\sigma}_j^{4\tilde{\alpha}^*},$$

where $\Psi(t) = \Gamma'(t)/\Gamma(t)$ and $w(\alpha) = (\nu/2)^{-\alpha} h_{\nu,1}(\alpha) \exp\{\alpha\Psi(\nu/2)\}$. More specifically, under some mild conditions, we have $\breve{b}_1(\boldsymbol{\sigma}^2) \xrightarrow{a.s.} b_1(\boldsymbol{\sigma}^2)$ and $\breve{b}_2(\boldsymbol{\sigma}^2) \xrightarrow{a.s.} b_2(\boldsymbol{\sigma}^2)$ as $p \to \infty$;

(iii) Otherwise, we estimate $b_1(\boldsymbol{\sigma}^2)$ and $b_2(\boldsymbol{\sigma}^2)$ by replacing σ_j^2 with the estimated optimal shrinkage estimates $\tilde{\sigma}_j^2(\tilde{\alpha}^*)$. Specifically, we estimate them by

$$\tilde{b}_1(\boldsymbol{\sigma}^2) = \tilde{\sigma}_{pool}^{-2\tilde{\alpha}^*} \sum_{j=1}^{p} \tilde{\sigma}_j^{2\tilde{\alpha}^*}(\tilde{\alpha}^*) \qquad \text{and} \qquad \tilde{b}_2(\boldsymbol{\sigma}^2) = \tilde{\sigma}_{pool}^{-4\tilde{\alpha}^*} \sum_{j=1}^{p} \tilde{\sigma}_j^{4\tilde{\alpha}^*}(\tilde{\alpha}^*).$$

3.3.2 Normal Approximation

For large p, the normal distribution can be a good approximation. The following content of this section can illustrate this point.

Take the one-sample shrinkage-based diagonal Hotelling's test for example. Consider the variances, σ_j^2, as random variables and assume that they are i.i.d. from a common distribution, F, with $E(\sigma_1^4) < \infty$ and $E\{\ln(\sigma_1^2)\} < \infty$. Let $U_j(\alpha) = n(\bar{X}_j - \mu_{0j})^2 \{h_{\nu,1}(-1)\hat{\sigma}_j^{-2}\}^{1-\alpha}$, where $j = 1, \ldots, p$ and $\alpha \in (0, 1]$. Then,

$$T_{\text{SD1}}^2(\alpha) = \{h_{\nu,p}(-1)\hat{\sigma}_{pool}^{-2}\}^{\alpha} \sum_{j=1}^{p} U_j(\alpha). \tag{3.9}$$

In Appendix 2, we show that

$$\hat{\sigma}_{pool}^{-2} \xrightarrow{a.s.} \exp\left[-E\left\{\ln(\sigma_1^2)\right\} + \ln\left(\frac{\nu}{2}\right) - \Psi\left(\frac{\nu}{2}\right)\right] \qquad \text{as } p \to \infty. \tag{3.10}$$

This implies that the first term in (3.9), $\{h_{\nu,p}(-1)\hat{\sigma}_{pool}^{-2}\}^{\alpha}$, converges to a constant when p is large. In addition, given that σ_j^2 are i.i.d. random variables, under H_0, it is easy to see that $U_j(\alpha)$ are also i.i.d. random variables. Thus, by the central limit theorem, for any $\nu > 4$ and $\alpha \in (0, 1]$, we have

$$\frac{\sum_{j=1}^{p} U_j(\alpha) - pE\{U_1(\alpha)\}}{\sqrt{p\text{Var}\{U_1(\alpha)\}}} \xrightarrow{\mathcal{D}} N(0, 1) \qquad \text{as } p \to \infty, \tag{3.11}$$

where $\xrightarrow{\mathcal{D}}$ denotes the convergence in distribution, $E\{U_1(\alpha)\} = E\left[E\left\{U_1(\alpha)|\sigma_1^2\right\}\right] =$

$h_{\nu,1}^{1-\alpha}(-1)E(\sigma_1^{2\alpha})/h_{\nu,1}\{-(1-\alpha)\}$ and

$$\begin{aligned}
\mathrm{Var}\{U_1(\alpha)\} &= E\left[\mathrm{Var}\left\{U_1(\alpha)|\sigma_1^2\right\}\right] + \mathrm{Var}\left[E\left\{U_1(\alpha)|\sigma_1^2\right\}\right] \\
&= \left[\frac{3h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}\{-2(1-\alpha)\}} - \frac{h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}^2\{-(1-\alpha)\}}\right] E(\sigma_1^{4\alpha}) + \frac{h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}^2\{-(1-\alpha)\}}\mathrm{Var}(\sigma_1^{2\alpha}) \\
&= \frac{3h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}\{-2(1-\alpha)\}}E(\sigma_1^{4\alpha}) - \frac{h_{\nu,1}^{2(1-\alpha)}(-1)}{h_{\nu,1}^2\{-(1-\alpha)\}}\left\{E(\sigma_1^{2\alpha})\right\}^2.
\end{aligned}$$

By (3.10) and (3.11), together with Slutsky's Theorem, we can claim that the test statistic $T_{\mathrm{SD1}}^2(\tilde{\alpha}^*)$ is approximately normally distributed when p is large. The same conclusion can also be obtained for $T_{\mathrm{SD2}}^2(\tilde{\alpha}^*)$.

Now as in Section 3.3.1, to have the normal approximation, we equate the mean and variance of $N(\xi_1, \tau_1)$ with the mean and variance of $T_{\mathrm{SD1}}^2(\tilde{\alpha}^*)$. Similarly, for the two-sample comparison, we use the same method to determine the mean, ξ_2, and the variance, τ_2, of $T_{\mathrm{SD2}}^2(\tilde{\alpha}^*)$. The results are summarized as the following theorems.

Theorem 11. *For any $n > 5$ and optimal shrinkage parameter estimation $\tilde{\alpha}^*$, under the null hypothesis, we have*

$$T_{\mathrm{SD1}}^2(\tilde{\alpha}^*) \sim N(\xi_1, \tau_1), \quad \text{as } p \to \infty,$$

where

$$\begin{aligned}
\xi_1 &= C_1\sigma_{pool}^{-2\tilde{\alpha}^*}\sum_{j=1}^{p}\sigma_j^{2\tilde{\alpha}^*}, \\
\tau_1 &= (3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*}\sum_{j=1}^{p}\sigma_j^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*}(\sum_{j=1}^{p}\sigma_j^{2\tilde{\alpha}^*})^2.
\end{aligned}$$

Theorem 12. *For any $n_1 + n_2 > 6$ and optimal shrinkage parameter estimation $\tilde{\alpha}^*$, under the null hypothesis, we have*

$$T_{\mathrm{SD2}}^2(\tilde{\alpha}^*) \sim N(\xi_2, \tau_2), \quad \text{as } p \to \infty,$$

where

$$\begin{aligned}
\xi_2 &= C_1\sigma_{pool}^{-2\tilde{\alpha}^*}\sum_{j=1}^{p}\sigma_{j,pool}^{2\tilde{\alpha}^*}, \\
\tau_2 &= (3C_2 - C_3)\sigma_{pool}^{-4\tilde{\alpha}^*}\sum_{j=1}^{p}\sigma_{j,pool}^{4\tilde{\alpha}^*} + (C_3 - C_1^2)\sigma_{pool}^{-4\tilde{\alpha}^*}(\sum_{j=1}^{p}\sigma_{j,pool}^{2\tilde{\alpha}^*})^2.
\end{aligned}$$

The practical rules for estimating the unknown quantities are the same as those in Section 3.3.1.

3.4 Monte Carlo Simulation Studies

In this section, we compare the shrinkage-based diagonal Hotelling's tests, including the chi-squared null distribution ($SDchi$) and the normal null distribution ($SDnor$), with some current methods in the aforementioned three categories:

- One unscaled Hotelling's test: Chen and Qin (2010) (CQ).
- One regularized Hotelling's test: Chen et al. (2011) (RHT).
- Two diagonal Hotelling's tests: Wu et al. (2006) (PCT) and Srivastava et al. (2013) (SR).

For RHT, we use the function "RHT.2samp" in the R package "RHT" provided by Chen et al. (2011). In our simulation studies, we simulate both the type I error rate and the power to assess the performances of all approaches. Moreover, we compare all methods by plotting the receiver operating characteristic (ROC) curves. The ROC curve describes the performance of the true positive rate (TPR) as the false positive rate (FPR) varies. The area under the curve (AUC) values are also provided. We mainly focus on small sample sizes in this section; but moderate to large sample sizes are also considered. Note that most existing methods were proposed for the two-sample case. For ease of comparison, we consider only the two-sample test in the simulation studies.

3.4.1 Simulation Design

In our simulation, we generate data from the multivariate normal distribution with a common covariance matrix $\mathbf{\Sigma}$. To assess the type I error rate, the data are generated for both groups from $N_p(\mathbf{0}, \mathbf{\Sigma})$. To assess the power, one group of data is generated from $N_p(\mathbf{0}, \mathbf{\Sigma})$ and the other one from $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\mu_j = c\sigma_j^2$ for $j = 1, \ldots, p$

with c being the effect size, and σ_j^2 is randomly drawn from the scaled chi-squared distribution $(1/5)\chi_5^2$.

The structure of the common covariance matrix is $\mathbf{\Sigma} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$, where $\mathbf{D} = \text{diag}(\mathbf{\Sigma}) = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and $\mathbf{R}$ is the correlation matrix. We use the following block-diagonal matrix as the correlation matrix:

$$\mathbf{R} = \begin{pmatrix} \mathbf{\Sigma}_\rho & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{-\rho} & \mathbf{0} & \ddots & \vdots \\ \vdots & \mathbf{0} & \mathbf{\Sigma}_\rho & \mathbf{0} & \vdots \\ \vdots & \ddots & \mathbf{0} & \mathbf{\Sigma}_{-\rho} & \ddots \\ \mathbf{0} & \cdots & \cdots & \ddots & \ddots \end{pmatrix}_{p \times p},$$

where $\mathbf{\Sigma}_\rho$ is a $q \times q$ matrix and $q \leq p$. We consider the following two settings for $\mathbf{\Sigma}_\rho$:

- $\mathbf{\Sigma}_\rho = (\sigma_{ij})_{q \times q}$, where $\sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq q$. Let $\mathbf{\Sigma}_{AR}$ denote this type of common covariance matrix.

- $\mathbf{\Sigma}_\rho = (\sigma_{ij})_{q \times q}$, where $\sigma_{ij} = 1$ for $i = j$ and $\sigma_{ij} = \rho$ for $1 \leq i \neq j \leq q$. Let $\mathbf{\Sigma}_{CS}$ denote this type of common covariance matrix.

For $\mathbf{\Sigma}_{AR}$, the correlation matrix is autoregressive of the order-1 structure (Guo et al.; 2007; Tong, Chen and Zhao; 2012). For $\mathbf{\Sigma}_{CS}$, the correlation matrix takes the compound symmetry structure.

In our simulation, we set n_1=n_2=n from 5 to 10, and the effect sizes are c=0.55, 0.52, 0.50, 0.47, 0.43, 0.39 and 0.35. For different correlations, we set ρ=0, 0.2 and 0.4. Note that $\mathbf{\Sigma}_{CS}$ is not positive if both $\pm$0.4 are included in the block-diagonal correlation matrix. For the case of $\rho = 0.4$, we set all correlations in $\mathbf{\Sigma}_{CS}$ to be positive. The type I error rate and power are obtained by running 1000 simulations under the settings of $p = 50$, $q = 5$ and $\alpha = 0.05$, where α is the significance level.

3.4.2 Simulation Results

We first focus on the performances of all approaches for small sample sizes. The type I error rate and power are reported in Tables 3.1 and 3.2, respectively. Two different

structures of the correlation are considered, and we find that the results are similar for both $\mathbf{\Sigma}_{AR}$ and $\mathbf{\Sigma}_{CS}$. From the results in Tables 3.1 and 3.2, we observe that the shrinkage-based diagonal Hotelling's tests outperform the other methods for different ρ. When the correlation is weak, our methods control the type I error rate well and at the same time maintain high power. As the correlation increases, the type I error rate of our methods becomes higher but it is still better than those of the other methods. PCT and SR have high type I error rates when the sample size is smaller than 10. CQ selects the null hypothesis too often; RHT tends to be conservative; and both of their powers are low. For higher correlations, these four approaches perform similarly.

The superiority of the shrinkage-based diagonal Hotelling's tests is also demonstrated in Figure 3.1 and Table 3.3. Figure 3.1 shows the plots of the ROC curves and their respective AUC values are shown in Table 3.3. As in Si and Liu (2013), we plot ROC curves with a range of FPR values from 0 to 0.1. AUC values are also calculated in the same range as the FPR values. Without loss of generality, we plot the ROC curves for n=6 in Figure 3.1 to assess the overall performances of all approaches for small sample sizes. The figure shows that our methods, $SDchi$ and $SDnor$, have the largest AUC values and highest ROC curves for all three correlations. Additionally, we observe a very interesting and important result from Figure 3.1 and Table 3.3. The six curves appearing in Figure 3.1 can be divided into three groups, and these three groups clearly represent the aforementioned three categories. We can see that diagonal Hotelling's tests, including our shrinkage-based methods, have the highest ROC curves. The unscaled Hotelling's test has the second highest ROC curves and the regularized Hotelling's test has the lowest ROC curves. This demonstrates that with limited numbers of observations, the diagonal Hotelling's tests are the best options.

Finally, we keep an eye on the case of the large sample size; for example, $n = 50$. The type I error rates and powers of all approaches are also reported in Tables 3.1 and 3.2. We find that the the shrinkage-based diagonal Hotelling's tests perform similarly to when the small sample size is small. However, the other approaches obtain satisfactory results that different greatly from the small sample size case. This demonstrates that for large sample sizes, it is unnecessary to borrow information

Table 3.1: Type I error rates for p=50 under the null case

ρ	$\mathbf{\Sigma}$	n	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	$\mathbf{\Sigma}_{AR} = \mathbf{\Sigma}_{CS}$	5	0.049	0.060	0.228	0.356	0.000	0.042
		6	0.044	0.059	0.132	0.250	0.000	0.039
		7	0.045	0.056	0.108	0.178	0.000	0.038
		8	0.052	0.056	0.072	0.148	0.000	0.040
		9	0.050	0.051	0.056	0.128	0.000	0.032
		10	0.046	0.053	0.047	0.120	0.000	0.023
		50	0.052	0.054	0.045	0.059	0.067	0.020
0.2	$\mathbf{\Sigma}_{AR}$	5	0.054	0.086	0.181	0.217	0.000	0.042
		6	0.051	0.075	0.135	0.253	0.000	0.026
		7	0.052	0.072	0.147	0.202	0.000	0.028
		8	0.050	0.076	0.105	0.154	0.000	0.034
		9	0.054	0.070	0.072	0.149	0.000	0.034
		10	0.055	0.068	0.049	0.140	0.000	0.016
		50	0.054	0.068	0.043	0.055	0.065	0.012
	$\mathbf{\Sigma}_{CS}$	5	0.055	0.090	0.273	0.338	0.000	0.044
		6	0.052	0.089	0.136	0.257	0.000	0.040
		7	0.052	0.094	0.110	0.204	0.000	0.038
		8	0.051	0.085	0.094	0.160	0.000	0.031
		9	0.056	0.085	0.060	0.156	0.000	0.023
		10	0.054	0.075	0.046	0.125	0.000	0.019
		50	0.059	0.076	0.044	0.064	0.070	0.008
0.4	$\mathbf{\Sigma}_{AR}$	5	0.069	0.106	0.280	0.257	0.000	0.063
		6	0.067	0.092	0.136	0.223	0.000	0.053
		7	0.073	0.087	0.110	0.175	0.000	0.033
		8	0.070	0.088	0.102	0.151	0.000	0.032
		9	0.069	0.083	0.070	0.116	0.000	0.020
		10	0.067	0.081	0.064	0.113	0.000	0.019
		50	0.068	0.083	0.046	0.058	0.064	0.010
	$\mathbf{\Sigma}_{CS}$	5	0.081	0.119	0.270	0.252	0.000	0.038
		6	0.085	0.119	0.136	0.214	0.000	0.036
		7	0.091	0.116	0.110	0.176	0.000	0.034
		8	0.085	0.114	0.106	0.143	0.000	0.034
		9	0.091	0.105	0.096	0.127	0.000	0.027
		10	0.092	0.106	0.063	0.113	0.000	0.015
		50	0.089	0.103	0.033	0.061	0.056	0.007

Table 3.2: Powers for p=50 under the alternative case

ρ	$\mathbf{\Sigma}$	n	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	$\mathbf{\Sigma}_{AR} = \mathbf{\Sigma}_{CS}$	5	0.859	0.904	0.910	0.978	0.000	0.070
		6	0.912	0.927	0.925	0.969	0.000	0.059
		7	0.887	0.953	0.830	0.952	0.000	0.060
		8	0.818	0.950	0.842	0.928	0.000	0.076
		9	0.855	0.873	0.756	0.915	0.002	0.062
		10	0.788	0.810	0.713	0.869	0.003	0.045
		50	0.771	0.798	0.739	0.776	0.804	0.406
0.2	$\mathbf{\Sigma}_{AR}$	5	0.866	0.848	0.878	0.884	0.000	0.056
		6	0.823	0.853	0.824	0.941	0.000	0.038
		7	0.839	0.950	0.896	0.936	0.000	0.040
		8	0.842	0.888	0.752	0.914	0.000	0.036
		9	0.867	0.859	0.724	0.902	0.002	0.040
		10	0.789	0.847	0.668	0.823	0.002	0.033
		50	0.761	0.772	0.696	0.734	0.744	0.572
	$\mathbf{\Sigma}_{CS}$	5	0.843	0.864	0.963	0.965	0.000	0.061
		6	0.847	0.908	0.950	0.961	0.000	0.050
		7	0.897	0.874	0.796	0.924	0.000	0.053
		8	0.849	0.893	0.736	0.913	0.000	0.046
		9	0.876	0.855	0.800	0.892	0.002	0.042
		10	0.747	0.825	0.750	0.818	0.003	0.033
		50	0.750	0.838	0.733	0.728	0.750	0.690
0.4	$\mathbf{\Sigma}_{AR}$	5	0.793	0.868	0.834	0.824	0.000	0.051
		6	0.836	0.902	0.910	0.866	0.000	0.034
		7	0.873	0.940	0.713	0.889	0.000	0.040
		8	0.784	0.880	0.729	0.846	0.000	0.041
		9	0.859	0.861	0.730	0.823	0.003	0.040
		10	0.728	0.782	0.714	0.725	0.003	0.036
		50	0.713	0.697	0.589	0.588	0.571	0.668
	$\mathbf{\Sigma}_{CS}$	5	0.792	0.789	0.847	0.884	0.000	0.038
		6	0.882	0.837	0.732	0.841	0.000	0.038
		7	0.875	0.886	0.715	0.861	0.000	0.041
		8	0.770	0.783	0.690	0.782	0.000	0.033
		9	0.778	0.796	0.700	0.710	0.000	0.031
		10	0.726	0.685	0.570	0.694	0.002	0.037
		50	0.664	0.689	0.522	0.636	0.556	0.470

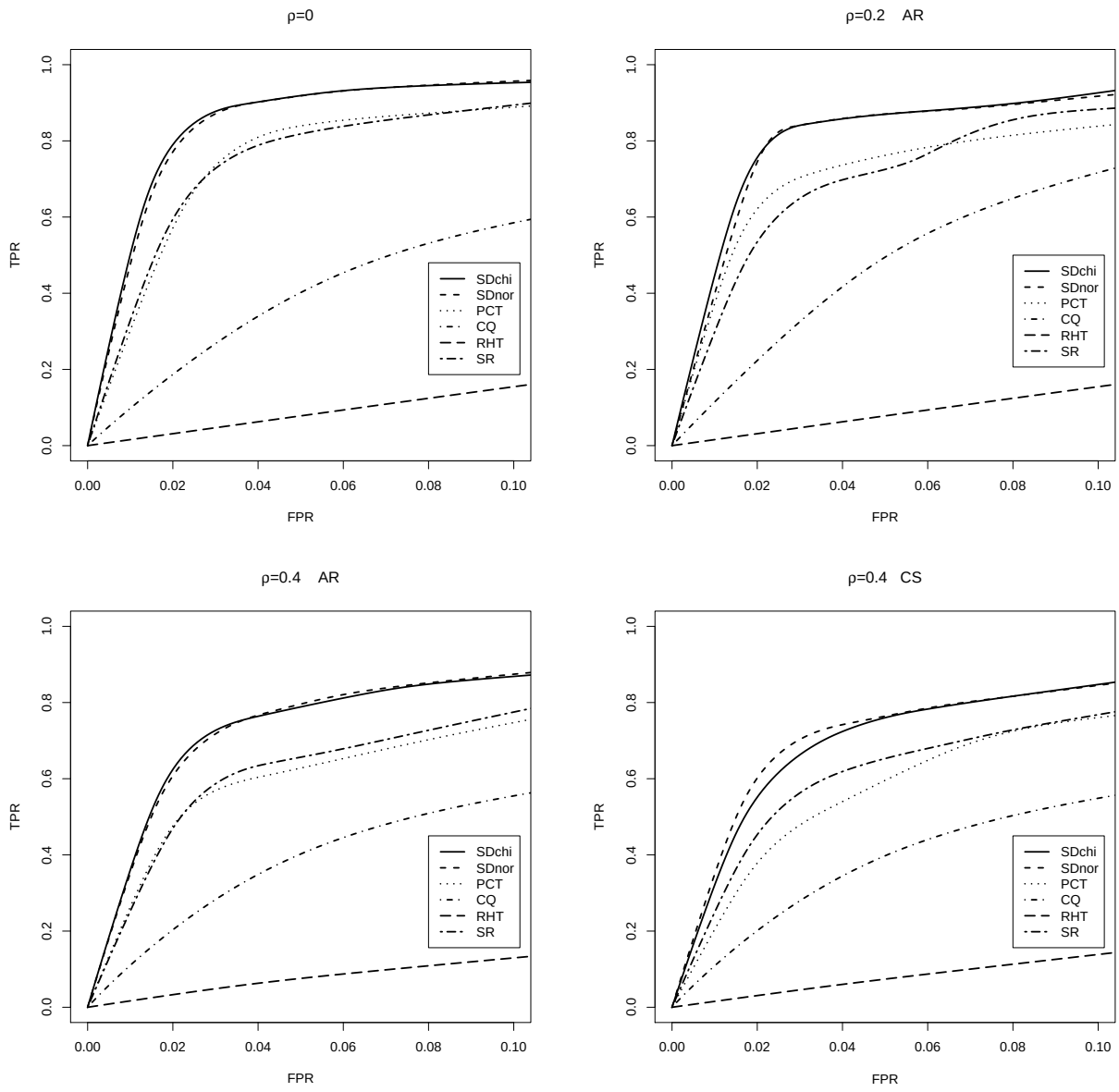

Figure 3.1: ROC curves for n=6 and p=50. "AR" represents $\mathbf{\Sigma}_{AR}$ and "CS" represents $\mathbf{\Sigma}_{CS}$.

Table 3.3: AUC values for n=6 and p=50

ρ	$\mathbf{\Sigma}$	$SDchi$	$SDnor$	PCT	SR	CQ	RHT
0	$\mathbf{\Sigma}_{AR}$	0.0823	0.0822	0.0711	0.0707	0.0351	0.0076
0.2	$\mathbf{\Sigma}_{AR}$	0.0772	0.0771	0.0680	0.0656	0.0429	0.0074
0.4	$\mathbf{\Sigma}_{AR}$	0.0698	0.0706	0.0558	0.0581	0.0353	0.0070
0.4	$\mathbf{\Sigma}_{CS}$	0.0668	0.0683	0.0528	0.0579	0.0350	0.0071

across all variables.

3.5 Case Studies

A real gene expression data set, whose sample size is smaller than 10, is not uncommon. In addition to the references in Section 3.1, the following data sets also demonstrate this point:

- Kuster et al. (2011). The data set includes two groups: the sham control group ($n_1 = 8$) and the myocardial infarction group ($n_2 = 8$). The total number of genes is 24,123.

- Bchetnia et al. (2012). The data set includes two groups: the control group ($n_1 = 6$) and the epidermolysis bullosa simplex group ($n_2 = 3$). The total number of genes is 32,321.

- Kaur et al. (2012). The data set includes two groups: the control group ($n_1 = 3$) and the polycystic ovary syndrome group ($n_2 = 7$). The total number of genes is 54,675.

- Mokry et al. (2012). The data set includes two groups: the Ls174T-L8 group ($n_1 = 6$) and the Ls174T-pTER-β-catenin group ($n_2 = 8$). The total number of genes is 54,675.

- Searcy et al. (2012). The data set includes two groups: the control group ($n_1 = 8$) and the Pioglitazone group ($n_2 = 8$). The total number of genes is 45,101.

In this section, we assess the performances of the shrinkage-based diagonall Hotelling's tests when they are applied to real gene expression data sets. Besides the Myeloma data which we have analyzed in Chapter 2, we also use the following three gene expression data sets in our case studies.

1. *Nakayama data*(Nakayama et al.; 2007).

   In this data set, there are ten types of soft tissue tumors based on 105 samples and 22,283 probe sets. Without loss of generality, we use the first two types of soft tissue tumors: synovial sarcoma (SS) and myxoid/round cell liposarcoma (MRCL), and the sample sizes are 16 and 19, respectively. All samples are log-transformed as described by Nakayama et al. (2007). This data set has been analyzed by Witten and Tibshirani (2011), who discussed the classification problem, and it can be downloaded from Gene Expression Omnibus (GEO) Datasets using access number GDS2736.

2. *Glioma data* (Sun et al.; 2006).

   There are four classes of data in this data set: one non-tumor class and three glioma classes. Totally, the data include 54,613 genes and 180 samples. We use the non-tumor (NON) class and the astrocytomas (AS) class, and the sample sizes are 23 and 26, respectively. The data set has also been analyzed by Witten and Tibshirani (2011), and it can be downloaded from GEO Datasets with access number GDS1962.

3. *Leukemia data* (Golub et al.; 1999).

   There are two different groups in this data set: the acute lymphoblastic leukemia (ALL) patients group and the acute myeloid leukemia (AML) patients group. The data contain 7,129 genes and 72 samples. We follow the method of Dudoit

et al. (2002) to threshold, filter, logarithmically (base 10) transform and standardize the data. Finally, we obtain leukemia data with 3,571 genes, 47 ALL patients and 25 AML patients, which are used in our analysis. The data set is available from the package "golubEsets" in Bioconductor.

To plot the ROC curves and calculate the AUC values, we first randomly select p genes from each data set for further analysis. Throughout this section, we consider $p = 50$ and $n = 5$. We then choose one class from each data set to calculate FPR. Specifically, they are SS, TH2, NON and ALL in our analysis. Now we use the first data set (SS and MRCL) for illustration to describe how the FPR and TPR are calculated. For the FPR, we randomly sample two distinct groups (each with size n) from SS and then use them to assess the type I errors. Instead, for the TPR, we sample one group (with size n) from SS and the other group (with size n) from MRCL and use them to assess the power. The FPR and TPR are calculated based on 1000 simulations.

Table 3.4: AUC values for all data sets when p=50 and n=5

DataSet	*SDchi*	*SDnor*	*PCT*	*SR*	*CQ*	*RHT*
Nakayama	0.0942	0.0941	0.0899	0.0880	0.0255	0.0668
Myloma	0.0174	0.0165	0.0148	0.0148	0.0088	0.0780
Glioma	0.0868	0.0856	0.0831	0.0781	0.0276	0.0399
Leukemia	0.0484	0.0479	0.0435	0.0382	0.0434	0.0291

Figure 3.2 shows the ROC curves for all four data sets, and AUC values are provided in Table 3.4. Similar to Figure 3.1, the ROC curves in Figure 3.2 are also generated with a range of FPR values from 0 to 0.1, and the same for AUC values. In Figure 3.2 and Table 3.4, our proposed approaches have the highest ROC curves and largest AUC values. This illustrates the advantage of the shrinkage-based diagonal Hotelling's tests. Additionally, the same result as in Section 3.4 can also be obtained; that is, the diagonal Hotelling's tests perform better than the unscaled Hotelling's tests and the regularized Hotelling's tests when the sample size is small.

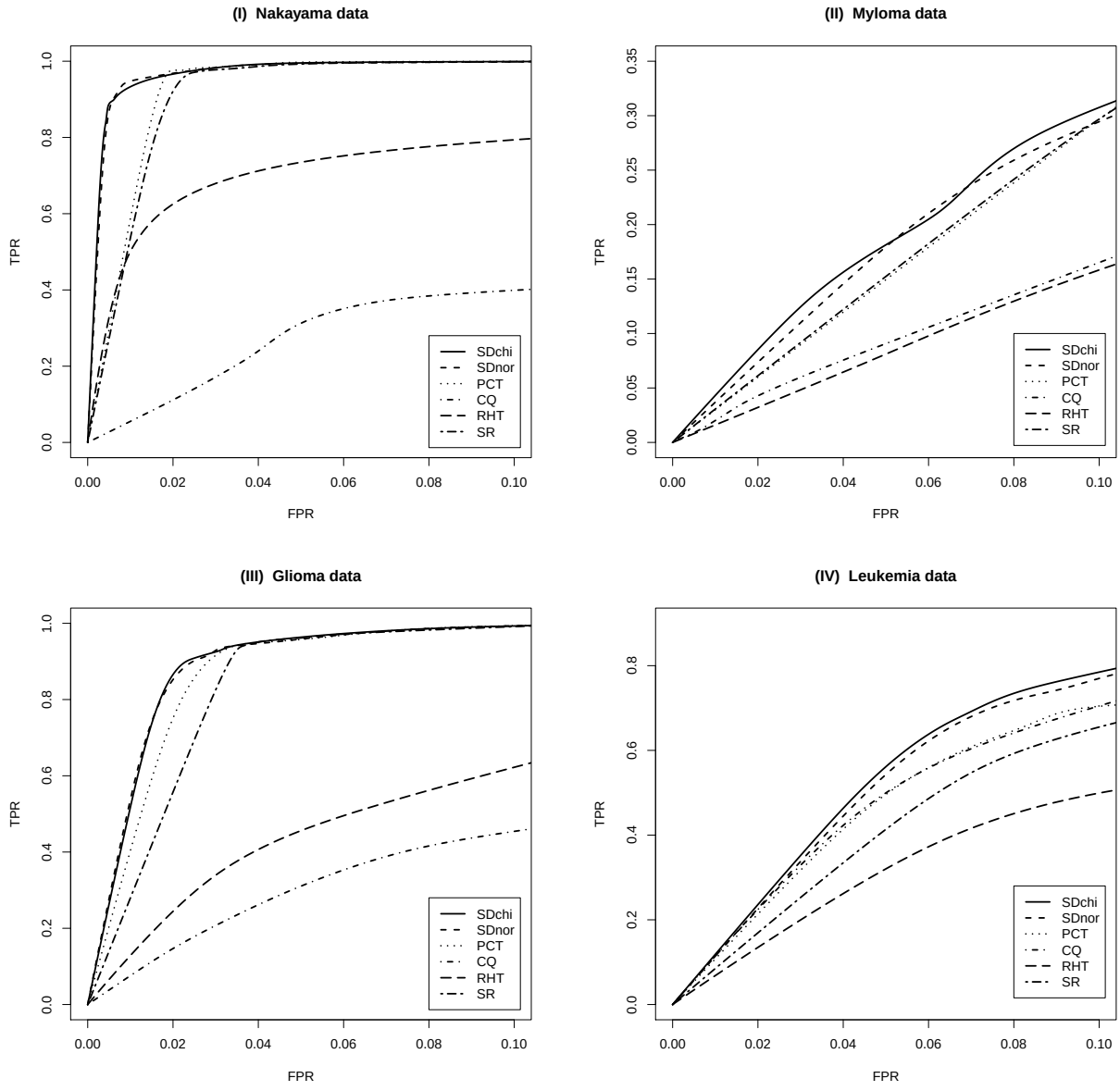

Figure 3.2: ROC curves for all data sets when p=50 and n=5

3.6 Discussion

The Hotelling's T^2 test is an important and useful tool for testing multivariate differences in means. However, its requirement that the sample size must be larger than the number of variables is violated in gene expression data analysis. Testing the significance of gene sets would be impossible with the Hotelling's T^2 test due to the difficulty of estimating $\mathbf{\Sigma^{-1}}$. It is therefore necessary to develop new methods to tackle "large p small n" multivariate testing problems. Currently, many statisticians have devoted themselves to solving this problem and some available approaches have been discovered, such as the unscaled Hotelling's tests, the regularized Hotelling's tests and the diagonal Hotelling's tests. However, because of cost or rarity of samples, a small sample size is a very common case. Current available approaches encounter difficulties while testing high-dimensional small sample size data. Our Monte Carlo simulation studies have demonstrated these issues.

In this chapter, we proposed a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. For high-dimensional small sample size data, the diagonal Hotelling's tests are better than the unscaled Hotelling's tests and the regularized Hotelling's tests. However, sample variance is an unreliable variance estimator for limited observations. Therefore, we use optimal shrinkage variance estimations to improve the performance of the diagonal Hotelling's test. The improvements are shown in our simulation studies. Consequently, we suggest using shrinkage-based diagonal Hotelling's tests to test the significance of gene sets with small sample sizes. Furthermore, if the number of genes in the gene sets is not large, the scaled chi-squared null distribution is recommended.

Nevertheless, from our simulation studies, we find that when the correlation becomes high, our methods have higher type I error rates, although higher ROC curves and larger AUC values than those of other methods can be obtained. This phenomenon is likely because the approximate null distributions in this chapter are not accurate enough. In addition, real data might not come from a multivariate normal distribution. Some heavy-tailed distributions or even discrete distributions are possi-

ble in real data. For example, RNA-seq data, obtained by next-generation sequencing technologies, have better coverage than microarray data have and such data have already been applied in medical science. RNA-seq data from typical high-dimensional small sample size discrete data sets and thus the shrinkage-based diagonal Hotelling's tests, based on multivariate normal distributions, are not suitable for testing the significance of RNA-seq gene sets.

3.7 Proofs

3.7.1 Proof of Lemma 6

For any non-zero $t > -\nu/2$, by Lemma 1 of Tong and Wang (2007), we have

$$E(\hat{\sigma}_j^{2t}) = \sigma_j^{2t}/h_{\nu,1}(t), \qquad j = 1, \ldots, p.$$

This leads to

$$\begin{aligned}
E\{\tilde{\sigma}_j^{-2}(\alpha)\} &= E\left[\{h_{\nu,p}(-1)\hat{\sigma}_{pool}^{-2}\}^{\alpha}\{h_{\nu,1}(-1)\hat{\sigma}_j^{-2}\}^{1-\alpha}\right] \\
&= h_{\nu,p}^{\alpha}(-1)h_{\nu,1}^{1-\alpha}(-1)E\left(\hat{\sigma}_1^{-2\alpha/p}\cdots\hat{\sigma}_j^{-2\alpha/p-2(1-\alpha)}\cdots\hat{\sigma}_p^{-2\alpha/p}\right) \\
&= h_{\nu,p}^{\alpha}(-1)h_{\nu,1}^{1-\alpha}(-1)E(\hat{\sigma}_1^{-2\alpha/p})\cdots E(\hat{\sigma}_j^{-2\alpha/p-2(1-\alpha)})\cdots E(\hat{\sigma}_p^{-2\alpha/p}) \\
&= h_{\nu,p}^{\alpha}(-1)h_{\nu,1}^{1-\alpha}(-1)\frac{\sigma_1^{-2\alpha/p}}{h_{\nu,1}(-\alpha/p)}\cdots\frac{\sigma_j^{-2\alpha/p-2(1-\alpha)}}{h_{\nu,1}\{-\alpha/p-(1-\alpha)\}}\cdots\frac{\sigma_p^{-2\alpha/p}}{h_{\nu,1}(-\alpha/p)} \\
&= C_1\sigma_{pool}^{-2\alpha}\sigma_j^{-2(1-\alpha)}.
\end{aligned}$$

Further, noting that $\bar{X}_j$ and σ_j^2 are independent of each other, we have

$$\begin{aligned}
E\left\{T_{SD1}^2(\alpha)\right\} &= nE\left\{\sum_{j=1}^{p}(\bar{X}_j-\mu_{0j})^2\tilde{\sigma}_j^{-2}(\alpha)\right\} \\
&= n\sum_{j=1}^{p}E(\bar{X}_j-\mu_{0j})^2E\{\tilde{\sigma}_j^{-2}(\alpha)\} \\
&= n\sum_{j=1}^{p}\frac{\sigma_j^2}{n}C_1\sigma_{pool}^{-2\alpha}\sigma_j^{-2(1-\alpha)} \\
&= C_1\sigma_{pool}^{-2\alpha}\sum_{j=1}^{p}\sigma_j^{2\alpha}.
\end{aligned}$$

To find the variance of $T^2_{SD1}(\alpha)$, it suffices to compute the second moment of $T^2_{SD1}(\alpha)$. For any $j \neq k$, by similar algebra as above, we have

$$E\{\tilde{\sigma}_j^{-2}(\alpha)\tilde{\sigma}_k^{-2}(\alpha)\} = C_3\sigma_{pool}^{-4\alpha}\sigma_j^{-2(1-\alpha)}\sigma_k^{-2(1-\alpha)}.$$

In addition, by the fact that $E(\bar{X}_j - \mu_{oj})^4 = 3\sigma_j^4/n^2$, it gives

$$E(\tilde{\sigma}_j^{-4}(\alpha)) = 3C_2\sigma_{pool}^{-4\alpha}\sigma_j^{-4(1-\alpha)}.$$

Therefore,

$$\begin{aligned} E\left\{T^2_{SD1}(\alpha)\right\}^2 &= n^2 E\left\{\sum_{j=1}^{p}\sum_{k=1}^{p}(\bar{X}_j - \mu_{0j})^2(\bar{X}_k - \mu_{0k})^2\tilde{\sigma}_j^{-2}(\alpha)\tilde{\sigma}_k^{-2}(\alpha)\right\} \\ &= n^2\sum_{j=1}^{p}E(\bar{X}_j - \mu_{0j})^4\tilde{\sigma}_j^{-4}(\alpha) \\ &\quad + n^2\sum_{j\neq k}E(\bar{X}_j - \mu_{0j})^2E(\bar{X}_k - \mu_{0k})^2E\{\tilde{\sigma}_j^{-2}(\alpha)\tilde{\sigma}_k^{-2}(\alpha)\} \\ &= 3C_2\sigma_{pool}^{-4\alpha}\sum_{j=1}^{p}\sigma_j^{4\alpha} + C_3\sigma_{pool}^{-4\alpha}\sum_{j\neq k}\sigma_j^{2\alpha}\sigma_k^{2\alpha}. \end{aligned}$$

Finally, we have

$$\begin{aligned} \mathrm{Var}\{T^2_{SD1}(\alpha)\} &= E\left\{T^2_{SD1}(\alpha)\right\}^2 - \left[E\{T^2_{SD1}(\alpha)\}\right]^2 \\ &= (3C_2 - C_3)\sigma_{pool}^{-4\alpha}\sum_{j=1}^{p}\sigma_j^{4\alpha} + (C_3 - C_1^2)\sigma_{pool}^{-4\alpha}(\sum_{j=1}^{p}\sigma_j^{2\alpha})^2. \end{aligned}$$

3.7.2 Derivation of formula (3.10)

Note that

$$\ln(\hat{\sigma}_{pool}^{-2}) = -\frac{1}{p}\sum_{j=1}^{p}\ln(\sigma_j^2) - \frac{1}{p}\sum_{j=1}^{p}\ln\left(\frac{\nu\hat{\sigma}_j^2}{\sigma_j^2}\right) + \ln(\nu).$$

Given that σ_j^2 are i.i.d. random variables with $E\left\{\ln(\sigma_1^2)\right\} < \infty$, by the strong law of large numbers,

$$\frac{1}{p}\sum_{j=1}^{p}\ln(\sigma_j^2) \xrightarrow{a.s.} E\left\{\ln(\sigma_1^2)\right\} \quad \text{as } p \to \infty.$$

In addition, noting that $\nu\hat{\sigma}_j^2/\sigma_j^2$ are i.i.d. chi-squared distributed with ν degrees of freedom, we have

$$\frac{1}{p}\sum_{j=1}^{p}\ln\left(\frac{\nu\hat{\sigma}_j^2}{\sigma_j^2}\right)\xrightarrow{a.s.}E\left\{\ln(\frac{\nu\hat{\sigma}_1^2}{\sigma_1^2})\right\}=\Psi\left(\frac{\nu}{2}\right)+\ln(2)\ \text{ as } p\to\infty.$$

Then, by Slutsky's theorem,

$$\ln(\hat{\sigma}_{pool}^{-2})\xrightarrow{a.s.}-E\left\{\ln(\sigma_1^2)\right\}+\ln\left(\frac{\nu}{2}\right)-\Psi\left(\frac{\nu}{2}\right)\ \text{ as } p\to\infty,$$

which leads to (3.10).

Chapter 4

A Comparison of Methods for Estimating the Determinant of High-Dimensional Covariance Matrix

4.1 Introduction

With the advancement of technology, high-dimensional data are becoming more and more common in scientific research including gene expression study, financial engineering and signal processing. One significant feature of such data is that the dimension is usually larger or much larger than the sample size. This results in the so-called "large p small n" data, where p is the dimension of the data and n is the sample size. One such example is the gene expression microarray data, in which one often measures thousands of gene expression values simultaneously for each individual. In contrary, due to the cost or other reasons such as the limited availability of patients, the number of samples in microarray experiments is usually small compared to the number of genes. It is not even uncommon to see microarray data with less than 10 samples (Kuster et al.; 2011; Mokry et al.; 2012; Kaur et al.; 2012; Searcy et al.; 2012). As seen in the literature, there are many statistical and computational

challenges in analyzing the "large p small n" data.

Let $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})^T$, $i = 1, \ldots, n$, be i.i.d. random vectors from the multivariate normal distribution $N_p(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ is a p-dimensional mean vector and Σ is a symmetric, nonnegative-definite, covariance matrix of size $p \times p$. When p is larger than n, it is known that the sample covariance matrix $\mathbf{S}_n$ is a singular matrix and so it is no longer a valid estimate of Σ. To overcome the singularity problem, various methods for estimating Σ have been proposed in the recent literature, e.g., the ridge-type estimators in Ledoit and Wolf (2003) and Fisher and Sun (2011), and the sparse estimators in Bickel and Levina (2008), Cai and Yuan (2012) and Rothman (2012). For more details, see also Tong et al. (2014) and the references therein.

Apart from the covariance matrix estimation, there are situations where one needs an estimate of the determinant (or the log-determinant) of the covariance matrix for high-dimensional data. To illustrate it, we write the log-likelihood function of the data as

$$\log(L) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{X}_i - \boldsymbol{\mu}),$$

where $|\Sigma|$ denotes the determinant of the covariance matrix Σ. The term $|\Sigma|$ plays an important role in statistical inference. It has many real applications including classification, multivariate multiple regression, and information theory. To cater for this demand, we present here several examples.

- Quadratic discriminant analysis (QDA) is an important method of classification. Assuming that the data in class k follow $N_p(\boldsymbol{\mu}_k, \Sigma_k)$, the quadratic discriminant scores are given by

  $$d_k(\mathbf{Y}) = (\mathbf{Y} - \boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{Y} - \boldsymbol{\mu}_k) + \log|\Sigma_k| - 2\log\pi_k, \qquad k = 1, \ldots, K,$$

  where $\mathbf{Y}$ is the new sample, K is the total number of classes, and π_k is the prior probability of observing a sample from class k. The classification rule is to assign $\mathbf{Y}$ to class k that minimizes $d_k(\mathbf{Y})$ among all classes. To implement QDA, it is obvious that we need an estimate of $|\Sigma_k|$ or $\log|\Sigma_k|$.

- To estimate the high-dimensional precision matrix $\Omega = \Sigma^{-1}$, Yuan and Lin (2007) and Friedman et al. (2008) proposed to solve the following optimization problem:

$$\hat{\Omega} = \arg\min_{\Omega>0}\{\text{tr}(\mathbf{S}_n\Omega) - \log|\Omega| + \lambda\|\Omega\|_1\},$$

where $\text{tr}(\cdot)$ is the trace, $\|\cdot\|_1$ is the ℓ_1 norm, and λ is a tuning parameter. The purpose of the term, $\log|\Omega| = -\log|\Sigma|$, is to ensure that the optimization problem has a unique global positive definite minimizer (Rothman; 2012). Other proposals in this direction include Banerjee et al. (2008), Witten and Tibshirani (2009), Ravikumar et al. (2011), Yin and Li (2013), and among others.

- In probability theory and information theory, the differential entropy is commonly used by extending the concept of entropy to the continuous probability distribution. For a random vector from $N_p(\boldsymbol{\mu}, \Sigma)$, the differential entropy is

$$h(\Sigma) = \frac{p}{2} + \frac{p\log(2\pi)}{2} + \frac{\log|\Sigma|}{2}.$$

Another related quantity is the Kullback-Leibler divergence that measures the difference between two probability distributions. Specifically, the Kullback-Leibler divergence between $N_p(\boldsymbol{\mu}, \Sigma)$ and $N_p(\boldsymbol{\mu}_1, \Sigma_1)$ is given by

$$D_{\text{KL}} = \frac{1}{2}\left\{\text{tr}(\Sigma_1^{-1}\Sigma) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu})^T\Sigma_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}) - p - \log|\Sigma| + \log|\Sigma_1|\right\}.$$

Needless to say, an estimate of the determinant $|\Sigma|$, or equivalently an estimate of the log-determinant $\log|\Sigma|$, is increasingly needed in high-dimensional data analysis. For ease of notation, we let

$$\theta = \log|\Sigma|$$

throughout this chapter. In contrast to the covariance matrix estimation, the importance of estimating θ is relatively overlooked in the literature. In practice, one often estimates the covariance matrix first and then use it to compute the log-determinant, rather than a direct estimate of θ. Recently, Cai et al. (2013) investigated the estimation of θ under various settings. Under some "moderate" setting with $p \leq n$, they proposed to estimate θ by the determinant of the sample covariance matrix, i.e.,

$\log|\mathbf{S}_n|$. A central limit theorem was also established for $\log|\mathbf{S}_n|$ in the setting where p can grow with n. For the "large p small n" data, however, they showed that it is impossible to estimate θ consistently, unless some structural assumption such as sparsity on the parameter can be imposed.

Note that a good estimate of Σ may not necessarily lead to a satisfactory estimate of θ. This motivates us to conduct a comprehensive simulation study that evaluates the performance of the existing methods for estimating Σ. In Section 4.2, we consider a total of nine methods for estimating θ. A brief review on each of the methods is also given. In Section 4.3, we conduct simulation studies to evaluate and compare their performance under various settings. In particular, we will consider different types of correlation structures including a non-positive definite covariance matrix that is often ignored in the existing literature. We then explore and summarize some useful findings, and provide some practical guidelines for scientists in Section 4.4. Finally, we note that the comparison study in this chapter also provides a way for assessing the performance of existing covariance matrix estimation methods. Technical details are provided in Section 4.5.

4.2 Methods for Estimating θ

Needless to say, estimating the high-dimensional covariance matrix has attracted a lot of attention. Various approaches have been proposed in the literature. In this section, we review eight representative methods for estimating the covariance matrix. We then estimate the log-determinant θ using the eight estimates of Σ, respectively. In addition, we also propose a new method for estimating θ under the assumption of a diagonal covariance matrix. For ease of presentation, we divide the nine methods into four different categories: the diagonal estimation, the shrinkage estimation, the sparse estimation, and the factor model estimation.

4.2.1 Diagonal Estimation

Method 1: Diagonal Estimator (DE)

Under the "large p small n" setting, one naive approach is to estimate Σ by the diagonal sample covariance matrix, i.e. $\text{diag}(\mathbf{S}_n)$. This estimator was first considered in Dudoit et al. (2002) where the authors proposed a diagonal linear discriminant analysis. It was also considered in Bickel and Levina (2004) where the authors demonstrated that a diagonal covariance matrix estimation can be sometimes reasonable when p is much larger than n. Let $\text{diag}(\Sigma) = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ where σ_j^2 are the covariate-specific variances for $j = 1, \ldots, p$, and $\text{diag}(\mathbf{S}_n) = \text{diag}(s_1^2, \ldots, s_p^2)$ where s_j^2 are the sample variances of σ_j^2, respectively. By letting $\hat{\Sigma} = \text{diag}(\mathbf{S}_n)$, we define the first estimator of θ as

$$\hat{\theta}_{(1)} = \log \left|\text{diag}(\mathbf{S}_n)\right| = \sum_{j=1}^{p} \log s_j^2. \tag{4.1}$$

We refer to $\hat{\theta}_{(1)}$ as the diagonal estimator (DE). To be specific, DE is proposed to estimate $\log |\text{diag}(\Sigma)|$ rather than $\log |\Sigma|$.

Method 2: Improved Diagonal Estimator (IDE)

It is noteworthy that DE may not perform well as an estimate of $\log |\text{diag}(\Sigma)|$ when the sample size is small, mainly due to the unreliable estimates of the sample variances. Various approaches have been proposed to improving the variance estimation in the literature. See, for example, Baldi and Long (2001), Wright and Simon (2003), Cui et al. (2005), Tong and Wang (2007), and Tong, Jang and Wang (2012).

To improve DE, we consider the optimal shrinkage estimator in Tong and Wang (2007),

$$\hat{\sigma}_j^2 = \{h_p(1) s_{pool}^2\}^{\alpha} \{h_1(1) s_j^2\}^{1-\alpha},$$

where $s_{pool}^2 = \prod_{i=1}^{p} (s_j^2)^{1/p}$, $h_p(1) = (\nu/2) \{\Gamma(\nu/2)/\Gamma(\nu/2 + 1/p)\}^p$ with $\nu = n - 1$, $\Gamma(\cdot)$ is the Gamma function, and $\alpha \in [0, 1]$ is the shrinkage parameter. Replacing s_j^2 in DE by $\hat{\sigma}_j^2$, we have

$$\hat{\theta} = \sum_{j=1}^{p} \log \hat{\sigma}_j^2 = \hat{\theta}_{(1)} + C, \tag{4.2}$$

where $C = \log \{h_p^{\alpha p}(1) h_1^{(1-\alpha)p}(1)\}$ is a constant.

The estimation structure in (4.2) shows that the DE estimator, $\hat{\theta}_{(1)}$, can be further improved. Specifically, if we have C_0 such that $E(\hat{\theta}_{(1)} + C_0) = \log|\text{diag}(\Sigma)|$, then C_0 defines as the optimal C value so that the estimator $\hat{\theta}_{(1)} + C_0$ minimizes the mean squared error in the family of estimators $\{\hat{\theta}_{(1)} + C : \ C \in (-\infty, \infty)\}$.

Theorem 13. *Let $s_j^2 = \sigma_j^2 \chi_{\nu,j}^2/\nu$, where $\chi_{\nu,j}^2$ are i.i.d random variables with a common chi-squared distribution with ν degrees of freedom, and $C_0 = -p\{\log(2/\nu) + \psi(\nu/2)\}$, where $\psi(\cdot) = \Gamma'(\cdot)/\Gamma(\cdot)$ is the digamma function. Then for any fixed $\nu > 0$, we have*

(1) $\hat{\theta}_{(1)} + C_0$ is an unbiased estimator of $\log|\text{diag}(\Sigma)|$.

(2) Assume also that σ_j^2 are i.i.d random variables from a common distribution F and $E(\log\sigma_1^2) < \infty$. Then

$$\frac{1}{p}\left(\hat{\theta}_{(1)} + C_0 - \log|\text{diag}(\Sigma)|\right) \xrightarrow{a.s.} 0 \qquad \text{as} \quad p \to \infty,$$

where $\xrightarrow{a.s.}$ denotes almost sure convergence.

The proof of Theorem 13 is given in Section 4.5. By (4.2) and Theorem 13, we define the second estimator of θ as

$$\hat{\theta}_{(2)} = \sum_{j=1}^{p} \log s_j^2 - p\{\log(2/\nu) + \psi(\nu/2)\}.$$

We refer to $\hat{\theta}_{(2)}$ as the improved diagonal estimator (IDE).

4.2.2 Shrinkage Estimation

Recall that the sample covariance matrix $\mathbf{S}_n$ is singular when the dimension is larger than the sample size. To overcome the singularity problem, other than the diagonal methods in Section 4.2.1, one may also estimate the covariance matrix by the following convex combination:

$$\mathbf{S}^* = \delta\mathbf{T} + (1-\delta)\mathbf{S}_n,$$

where $\mathbf{T}$ is the target matrix and $\delta \in [0,1]$ is the shrinkage parameter. Both the target matrix and the shrinkage parameter play an important role in the shrinkage

estimation. For instance, if we let $\mathbf{T} = \text{diag}(\mathbf{S}_n)$ and $\delta = 1$, then $\mathbf{S}^*$ reduces to the DE estimator.

The appropriate choice of the target matrix has been extensively studies in the literature. See, for example, Ledoit and Wolf (2003), Schäfer and Strimmer (2005), Warton (2008), Warton (2011), and Fisher and Sun (2011) and the references therein. Note that $\mathbf{T}$ is often chosen to be positive definite and well-conditioned, and consequently, the final estimate $\mathbf{S}^*$ is also guaranteed positive definite and well-conditioned for any dimensionality. As suggested in Schäfer and Strimmer (2005) and Fisher and Sun (2011), we consider two popular target matrices: (1) the "diagonal, unit variance" matrix, i.e., the identity matrix $\mathbf{I}$, and (2) the "diagonal, unequal variance" matrix, i.e., the diagonal sample covariance matrix $\text{diag}(\mathbf{S}_n)$.

Note also that, given the target matrix, the estimation of the shrinkage parameter δ is also crucial to the final estimation. The available estimation methods for the shrinkage parameter are mainly: (1) the unbiased estimation, and (2) the consistent estimation. The unbiased estimation is obtained by minimizing a risk function to compute the optimal value, replacing unknown terms in the optimal value by their unbiased estimation, and truncating the estimated optimal value. The consistent estimation is obtained by deriving the (n, p)-consistent estimators for the unknown terms in the optimal shrinkage parameter derived in Ledoit and Wolf (2004). Taken together, we consider a total of four methods for estimating the covariance matrix and consequently for estimating θ.

Method 3: Unbiased Shrinkage Estimator with $\mathbf{T} = \mathbf{I}$ *(USIE)*

Letting the target matrix be $\mathbf{T} = \mathbf{I}$, Schäfer and Strimmer (2005) proposed an unbiased estimator for the shrinkage parameter, denoted by $\hat{\delta}_1^*$. This leads to $\mathbf{S}^* = \hat{\delta}_1^*\mathbf{I} + (1 - \hat{\delta}_1^*)\mathbf{S}_n$. We then define the third estimator of θ as

$$\hat{\theta}_{(3)} = \log|\hat{\delta}_1^*\mathbf{I} + (1 - \hat{\delta}_1^*)\mathbf{S}_n|. \tag{4.3}$$

Method 4: Consistent Shrinkage Estimator with $\mathbf{T} = \mathbf{I}$ *(CSIE)*

Letting the target matrix be $\mathbf{T} = \mathbf{I}$, Fisher and Sun (2011) proposed a consistent

estimator for the shrinkage parameter, denoted by $\hat{\delta}_2^*$. This leads to $\mathbf{S}^* = \hat{\delta}_2^*\mathbf{I} + (1 - \hat{\delta}_2^*)\mathbf{S}_n$. We then define the fourth estimator of θ as

$$\hat{\theta}_{(4)} = \log|\hat{\delta}_2^*\mathbf{I} + (1 - \hat{\delta}_2^*)\mathbf{S}_n|. \tag{4.4}$$

Method 5: Unbiased Shrinkage Estimator with $\mathbf{T} = \text{diag}(\mathbf{S}_n)$ *(USDE)*

Letting $\mathbf{T} = \text{diag}(\mathbf{S}_n)$, Schäfer and Strimmer (2005) also proposed an unbiased estimator for the shrinkage parameter, denoted by $\hat{\delta}_3^*$. This leads to $\mathbf{S}^* = \hat{\delta}_3^*\text{diag}(\mathbf{S}_n) + (1 - \hat{\delta}_3^*)\mathbf{S}_n$. We then define the fifth estimator of θ as

$$\hat{\theta}_{(5)} = \log|\hat{\delta}_3^*\text{diag}(\mathbf{S}_n) + (1 - \hat{\delta}_3^*)\mathbf{S}_n|. \tag{4.5}$$

Method 6: Consistent Shrinkage Estimator with $\mathbf{T} = \text{diag}(\mathbf{S}_n)$ *(CSDE)*

Letting $\mathbf{T} = \text{diag}(\mathbf{S}_n)$, Fisher and Sun (2011) also proposed a consistent estimator for the shrinkage parameter, denoted by $\hat{\delta}_4^*$. This leads to $\mathbf{S}^* = \hat{\delta}_4^*\text{diag}(\mathbf{S}_n) + (1 - \hat{\delta}_4^*)\mathbf{S}_n$. We then define the sixth estimator of θ as

$$\hat{\theta}_{(6)} = \log|\hat{\delta}_4^*\text{diag}(\mathbf{S}_n) + (1 - \hat{\delta}_4^*)\mathbf{S}_n|. \tag{4.6}$$

4.2.3 Sparse Estimation

When p is much larger than n, the shrinkage methods in Section 4.2.2 may not achieve a significant improvement over $\mathbf{S}_n$. In such settings, to have a good estimate of Σ, one may have to impose some structural assumptions such as sparsity in the parameters. A typical sparsity is to assume that most of the off-diagonal elements in the covariance matrix are zero. To estimate the covariance matrix under a sparsity condition, various thresholding-based methods have been proposed in the literature that aim to locate some "large" off-diagonal elements. See, for example, Bickel and Levina (2008), Karoui (2008), Rothman et al. (2009), Lam and Fan (2009), Cai and Liu (2011), Cai and Yuan (2012), and Cai and Zhou (2012). In this section, we consider two representative sparsity methods and then use them to estimate θ, i.e., the log-determinant of the covariance matrix.

Method 7: Hard Thresholding Estimator (HTE)

Under the sparsity assumption, Bickel and Levina (2008) introduces a universal thresholding method for estimating the sparse covariance matrix. Specifically, given $\mathbf{S}_n = (s_{ij})_{p\times p}$, their proposed estimator is

$$\hat{\Sigma}_\gamma = (\tilde{\sigma}_{ij})_{p\times p} \quad \text{with} \quad \tilde{\sigma}_{ij} = s_{ij}\mathbf{1}\{\text{abs}(s_{ij}) \geq \gamma\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, abs$(\cdot)$ is the absolute operation, and γ is a threshold value. Now by $\hat{\Sigma}_\gamma$, we define the seventh estimator of θ as

$$\hat{\theta}_{(7)} = \log|\hat{\Sigma}_\gamma|. \tag{4.7}$$

We refer to $\hat{\theta}_{(7)}$ as the hard thresholding estimator (HTE).

Method 8: Adaptive Thresholding Estimator (ATE)

Note that HTE is a universal thresholding method where all entries in the sample covariance matrix are thresholded by a common value γ. To make HTE well behaved, we require that the variances σ_j^2 are uniformly bounded by a constant K, and consequently, the variances of the entries of the sample covariance matrix are uniformly bounded.

Cai and Liu (2011) showed that HTE is suboptimal over a certain class of sparse covariance matrices. To improve it, they proposed an adaptive thresholding estimator for the covariance matrix:

$$\hat{\Sigma}^* = (\tilde{\sigma}^*_{ij})_{p\times p} \quad \text{with} \quad \tilde{\sigma}^*_{ij} = s_{\gamma_{ij}}(s_{ij}),$$

where γ_{ij} are individual threshold values of all entries in the sample covariance matrix, and $s_{\gamma_{ij}}(\cdot)$ is a generalized thresholding operator which is defined similarly as those in Rothman et al. (2009). They further showed that, with the proper choices of γ_{ij} and $s_{\gamma_{ij}}(\cdot)$, the estimator $\hat{\Sigma}^*$ adaptively achieves the optimal rate of convergence over a large class of sparse covariance matrices under the spectral norm. Now by $\hat{\Sigma}^*$, we define the eighth estimator of θ as

$$\hat{\theta}_{(8)} = \log|\hat{\Sigma}^*|. \tag{4.8}$$

We refer to $\hat{\theta}_{(8)}$ as the adaptive thresholding estimator (ATE).

4.2.4 Factor Model Estimation

The sparsity condition on the covariance matrix assumes that most of covariates are uncorrelated to each other. Note that, however, this assumption may not be realistic in practice. Recently, under the assumption of conditional sparsity, Fan et al. (2013) introduced a principle orthogonal complement thresholding method using the factor model. In this section, we briefly review their method and then apply it to estimate the log-determinant of the covariance matrix.

Method 9: Principal Orthogonal Complement Thresholding Estimator (POCTE)

Fan et al. (2013) considered the approximate factor model:

$$\mathbf{y}_g = \mathbf{B}\mathbf{f}_g + \mathbf{u}_g, \quad g = 1, \ldots, G,$$

where $\mathbf{y}_g = (y_{1g}, \ldots, y_{pg})^T$ is the observed response, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)^T$ is the loading matrix, $\mathbf{f}_g$ is a $Q \times 1$ vector of common factors, and $\mathbf{u}_g = (u_{1g}, \ldots, u_{pg})^T$ is the error vector. In this model, we can only observe $\mathbf{y}_g$. Let

$$\Sigma = \mathbf{B}\text{cov}(\mathbf{f}_g)\mathbf{B}^T + \Sigma_u, \quad g = 1, \ldots, G,$$

where Σ_u is the covariance matrix of $\mathbf{u}_g$. To estimate Σ, Fan et al. (2013) applied the spectral decomposition on the sample covariance matrix:

$$\mathbf{S}_n = \sum_{j=1}^{Q} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \hat{\mathbf{R}}_Q,$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ are eigenvalues of $\mathbf{S}_n$, $\hat{\xi}_j$, $j = 1, \ldots, p$ are the corresponding eigenvectors, and $\hat{\mathbf{R}}_Q = \sum_{j=Q+1}^{p} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T$ is the principle orthogonal complement. For this decomposition, the first Q principle components were kept and the thresholding was applied on $\hat{\mathbf{R}}_Q$. Here, the generalized thresholding operator can be used. In addition, Fan et al. (2013) also introduced a method to obtain an estimation of Q, denoted by $\hat{Q}$. Their final estimator of Σ is

$$\hat{\Sigma}_{\hat{Q}} = \sum_{j=1}^{\hat{Q}} \hat{\lambda}_j \hat{\xi}_j \hat{\xi}_j^T + \hat{\mathbf{R}}_{\hat{Q}}^{\mathcal{T}}, \tag{4.9}$$

where $\hat{\mathbf{R}}_{\hat{Q}}^{\mathcal{T}}$ is the thresholding result of $\hat{\mathbf{R}}_Q$. Now by (4.9), we define the ninth estimator of θ as

$$\hat{\theta}_{(9)} = \log|\hat{\Sigma}_{\hat{Q}}|. \tag{4.10}$$

We refer to $\hat{\theta}_{(9)}$ as the principal orthogonal complement thresholding estimator (POCTE).

4.3 Simulation Studies

In this section, we evaluate and compare the performance of all nine methods under various simulation settings. To compare these methods, we compute the mean squared error (MSE) as below:

$$\text{MSE}(\theta, \hat{\theta}) = \frac{1}{Mp}\sum_{m=1}^{M}(\hat{\theta}_m - \theta)^2,$$

where M is the repeated time. Throughout the simulations, we take $M = 1,000$. The data are generated from the multivariate normal distribution $N_p(\mathbf{0}, \Sigma)$. In what follows, three different setups of Σ will be considered, respectively.

4.3.1 Setup I

In this setup, we consider to generate a realistic covariance matrix from the Myeloma data (Zhan et al.; 2007), which is a real microarray data set including a total of 54,675 genes, 351 samples in the first group and 208 samples in the second group. To generate the covariance matrix, we first select p genes (50 or 300) randomly from the first group and then compute the sample covariance matrix using the selected genes, denote by Σ_r. Next, to evaluate the performance of the estimators under different levels of dependence, we follow Tong et al. (2013) and define the true covariance matrix as

$$\Sigma_1 = (1-\rho)\text{diag}(\Sigma_r) + \rho\Sigma_r,$$

where ρ controls the level of dependence. We set $\rho = 0$, $1/3$, $2/3$, and 1. Note that $\rho = 0$ corresponds to a diagonal covariance matrix, and $\rho = 1$ treats the generated covariance matrix as the true covariance matrix.

Figures 4.1 and 4.2 show the log(MSE) of all nine methods for different correlations, numbers of genes and sample sizes. Among the nine methods, the shrinkage estimation performs the best when the sample size is very small. If the sample size becomes large, the performances of the nine methods are different under different correlations. IDE outperforms the other eight methods if the covariates are independent. USDE and CSDE perform better if the correlation among covariates becomes strong. In addition, we note that under the case of strong correlations, POCTE tends to outperform others if we have larger number of genes.

4.3.2 Setup II

In this setup, we consider a block diagonal structure for the covariance matrix. This structure is widely adopted in the literature, e.g., in Guo et al. (2007) and Pang et al. (2009). Specifically, we let

$$\Sigma_2 = D^{1/2} R D^{1/2},$$

where $D = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ with σ_j^2 being i.i.d. from the distribution $(1/5)\chi_5^2$, and R follows the following block diagonal structure:

$$R = \begin{pmatrix} \Sigma_\rho & 0 & \cdots & \cdots & 0 \\ 0 & \Sigma_{-\rho} & 0 & \ddots & \vdots \\ \vdots & 0 & \Sigma_\rho & 0 & \vdots \\ \vdots & \ddots & 0 & \Sigma_{-\rho} & \ddots \\ 0 & \cdots & \cdots & \ddots & \ddots \end{pmatrix}_{p \times p}.$$

In our simulations, we consider $\Sigma_\rho = (\sigma_{ij}(\rho))_{q \times q}$ with $\sigma_{ij}(\rho) = \rho^{|i-j|}$ for $1 \leq i, j \leq q$. In addition, we set $\rho = 0$, 0.3, 0.6, and 0.9, to represent different levels of dependence, and $(p, q) = (50, 5)$ and $(300, 15)$, respectively.

Figures 4.3 and 4.4 display the log(MSE) of all nine methods for different correlations, numbers of genes and sample sizes. Both figures exhibit similar results as those of Setup I in the case of small sample sizes and independent covariates. For larger sample sizes, however, we can see the improved performances of sparse estimation and factor model estimation. In particular, POCTE gives the best performance for

strong correlations. Figure 4.5 displays the performances of the nine methods for different levels of dependence. It is also shown that POCTE outperforms the other eight methods for high level of dependence when the sample size is large.

4.3.3 Setup III

Recall that most existing methods require that Σ is positive definite. In this setup, we consider a non-positive definite covariance matrix and investigate the performance of the nine methods under the violation of the positive definite assumption. Note that this new setting is often overlooked in the literature. To construct a non-positive definite covariance matrix, we define the affine transformation $\mathbf{C}$ as

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & 1 & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}_{p \times p}.$$

We then apply the affine transformation to the covariance matrix in Setup II and form

$$\Sigma_3 = \mathbf{C}\Sigma_2\mathbf{C}^T.$$

It is obvious that $|\Sigma_3| = 0$ since $|\mathbf{C}| = 0$. We set $(p, q) = (50, 5)$, and $\rho = 0$, 0.3, 0.6, and 0.9. Note that the log-determinant of Σ_3 is negative infinity. Hence, for this degenerate setting, the MSE is defined on the determinant rather than on the log-determinant. Specifically, it is

$$\text{MSE}(e^{\theta}, e^{\hat{\theta}}) = \frac{1}{Mp}\sum_{m=1}^{M}\left(e^{\hat{\theta}_m} - e^{\theta}\right)^2.$$

Figure 4.6 shows the log(MSE) of all nine methods for different correlations and sample sizes. We can see that the simulation results are different from those of Setups

I and II. It is evident that POCTE outperforms the other eight methods when the positive definite assumption on Σ is violated.

4.4 Conclusion and Discussion

In this chapter, we have compared a total of nine methods for estimating the log-determinant, or equivalently the determinant, of high-dimensional covariance matrix. Three types of covariance structures are considered in our simulation studies. In what follows, we summarize some useful findings of the comparison results and also provide some practical guidelines for scientists.

The diagonal estimator, DE, is the simplest method for estimating the determinant of high-dimensional covariance matrix. However, it assumes that all covariates are independent. IDE, which is an unbiased estimator of $\log|\text{diag}(\Sigma)|$, has the best performance if the number of samples is moderate and the assumption of independence is satisfied. Once the correlation becomes strong, there are many other methods that have better performances. Therefore, if we have prior information that most of the covariates are independent or weakly correlated, IDE can be recommended for estimating the determinant of high-dimensional covariance matrix.

In general, the shrinkage estimators perform well for small sample sizes. For the shrinkage estimation, different choices of the target matrix and shrinkage parameter result in different performances for the estimation. First, we observed that USIE and CSIE tend to perform better for weakly correlated covariates, and USDE and CSDE tend to perform better for highly correlated covariates. Therefore, we suggest to adopt the identity matrix as the target matrix when estimating the determinant of weakly correlated covariance matrix, and the diagonal sample covariance matrix as the target matrix when the covariates are highly correlated. Second, we observed that for a common target matrix, USIE and USDE tend to have better performances than CSIE and CSDE, respectively. This indicates that when the sample size is very small, the unbiased estimators are better choices for estimating θ.

For the sparse estimators, we observed that they perform well when the sample size

is moderate and the correlations are strong. We also noted that the performances of HTE and ATE are always similar to each other. In practice, if we have prior information that most of the covariates are weakly correlated and the sample size is not small, then the sparse estimators can be recommended for estimating the determinant of high-dimensional covariance matrix.

As a factor model estimation, POCTE is very attractive for strong correlated data sets. Fan et al. (2013) assumed that the data are weakly correlated after extracting the common factors which can result in high levels of dependence among the covariates. Hence POCTE has better performance if the data are strong correlated. Note also that POCTE can select $Q = 0$ automatically if the true covariance matrix is sparse. Then consequently, their method will degenerate to the sparse estimation in such cases. In addition, the superiority of POCTE can be seen in our simulation studies when the true covariance matrix is non-positive definite.

Overall, there is no single method that dominates other methods under all considered circumstances. In general, the sample size and the correlation of the data sets have a great impact on the accuracy of estimation. Consequently, we may select appropriate estimation methods according to the sample size and the prior information on the correlation structure of the covariates. In the situations, however, when such prior information is not available, we recommend to use POCTE to estimate the determinant of high-dimensional covariance matrix.

Finally, it is noteworthy that the comparison study in this chapter also serves as a proxy to assess the performance of the covariance matrix estimation. Specifically, from a perspective of the loss function, if we define the loss function as

$$L(\hat{\Sigma}, \Sigma) = (\log|\hat{\Sigma}| - \log|\Sigma|)^2 \quad \text{or} \quad L(\hat{\Sigma}, \Sigma) = (|\hat{\Sigma}| - |\Sigma|)^2,$$

then the conducted simulations in Section 4.3 provide exactly a comparison for the nine methods for estimating Σ rather than θ. Of course, we do not intend to claim that the above loss function should be consistently recommended. In contrast, for evaluating the covariance matrix estimation, other commonly used loss functions are also available. For instance, we may also consider the following loss functions:

- Loss function in Karoui (2008), Rothman et al. (2009) and Fan et al. (2011): $L(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_2 = \sqrt{\lambda_{\max}\{(\hat{\Sigma} - \Sigma)^T(\hat{\Sigma} - \Sigma)\}}$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue.

- Loss function in Cai and Liu (2011) and Fan et al. (2013): $L(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_F = \sqrt{\sum_{i,j}(\hat{\sigma}_{ij} - \sigma_{ij})^2}$, where $\Sigma = (\sigma_{ij})_{p\times p}$, $\hat{\Sigma} = (\hat{\sigma}_{ij})_{p\times p}$.

- Loss function in Fan et al. (2013): $L(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma} - \Sigma\|_{\max} = \max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}|$.

Further research is needed to investigate which loss function provides to be the best candidate for evaluating the estimation methods for the covariance matrix.

4.5 Proofs

4.5.1 Proof of Theorem 13

(1) From $s_j^2 = \sigma_j^2 \chi_{\nu,j}^2/\nu$, we have $\log s_j^2 = \log \sigma_j^2 + \log(\chi_{\nu,j}^2/\nu)$. Then,

$$\sum_{j=1}^{p} \log s_j^2 = \sum_{j=1}^{p} \log \sigma_j^2 + p \log \chi_{\nu,j}^2 - p \log \nu.$$

Hence

$$\begin{aligned} E\left(\sum_{j=1}^{p} \log s_j^2\right) &= \sum_{j=1}^{p} \log \sigma_j^2 + pE(\log \chi_{\nu,j}^2) - p \log \nu \\ &= \sum_{j=1}^{p} \log \sigma_j^2 + p\{\log 2 + \psi(\nu/2)\} - p \log \nu. \end{aligned}$$

This leads to

$$\begin{aligned} E\left\{\hat{\theta}_{(1)} + C_0\right\} &= E\left(\sum_{j=1}^{p} \log s_j^2\right) - p\{\log 2 + \psi(\nu/2)\} + p \log \nu \\ &= \sum_{j=1}^{p} \log \sigma_j^2 \\ &= \log |\mathrm{diag}(\Sigma)|. \end{aligned}$$

This proves that $\hat{\theta}_{(1)} + C_0$ is an unbiased estimator of $\log |\mathrm{diag}(\Sigma)|$.

(2) For $E(\log \sigma_1^2) < \infty$, we have

$$\frac{1}{p}\sum_{j=1}^{p} \log \sigma_j^2 \xrightarrow{a.s.} E(\log \sigma_1^2) \quad as\ p \to \infty.$$

Since $E(\log s_1^2) = E\{E(\log s_1^2 | \sigma_1^2)\} = E(\log \sigma_1^2) + \log(2/\nu) + \psi(\nu/2)$, we have

$$\frac{1}{p}\sum_{j=1}^{p} \log s_j^2 - \log(2/\nu) - \psi(\nu/2) \xrightarrow{a.s.} E(\log \sigma_1^2) \quad as\ p \to \infty.$$

Hence

$$\frac{1}{p}\sum_{j=1}^{p} \log s_j^2 - \log(2/\nu) - \psi(\nu/2) - \frac{1}{p}\sum_{j=1}^{p} \log \sigma_j^2 \xrightarrow{a.s.} 0 \quad as\ p \to \infty.$$

Finally, we have

$$\begin{aligned} \frac{1}{p}\left\{\hat{\theta}_{(1)} + C_0 - \log|\mathrm{diag}(\Sigma)|\right\} &= \frac{1}{p}\sum_{j=1}^{p} \log s_j^2 - \log(2/\nu) - \psi(\nu/2) - \frac{1}{p}\sum_{j=1}^{p} \log \sigma_j^2 \\ &\xrightarrow{a.s.} 0 \quad as\ p \to \infty. \quad \square \end{aligned}$$

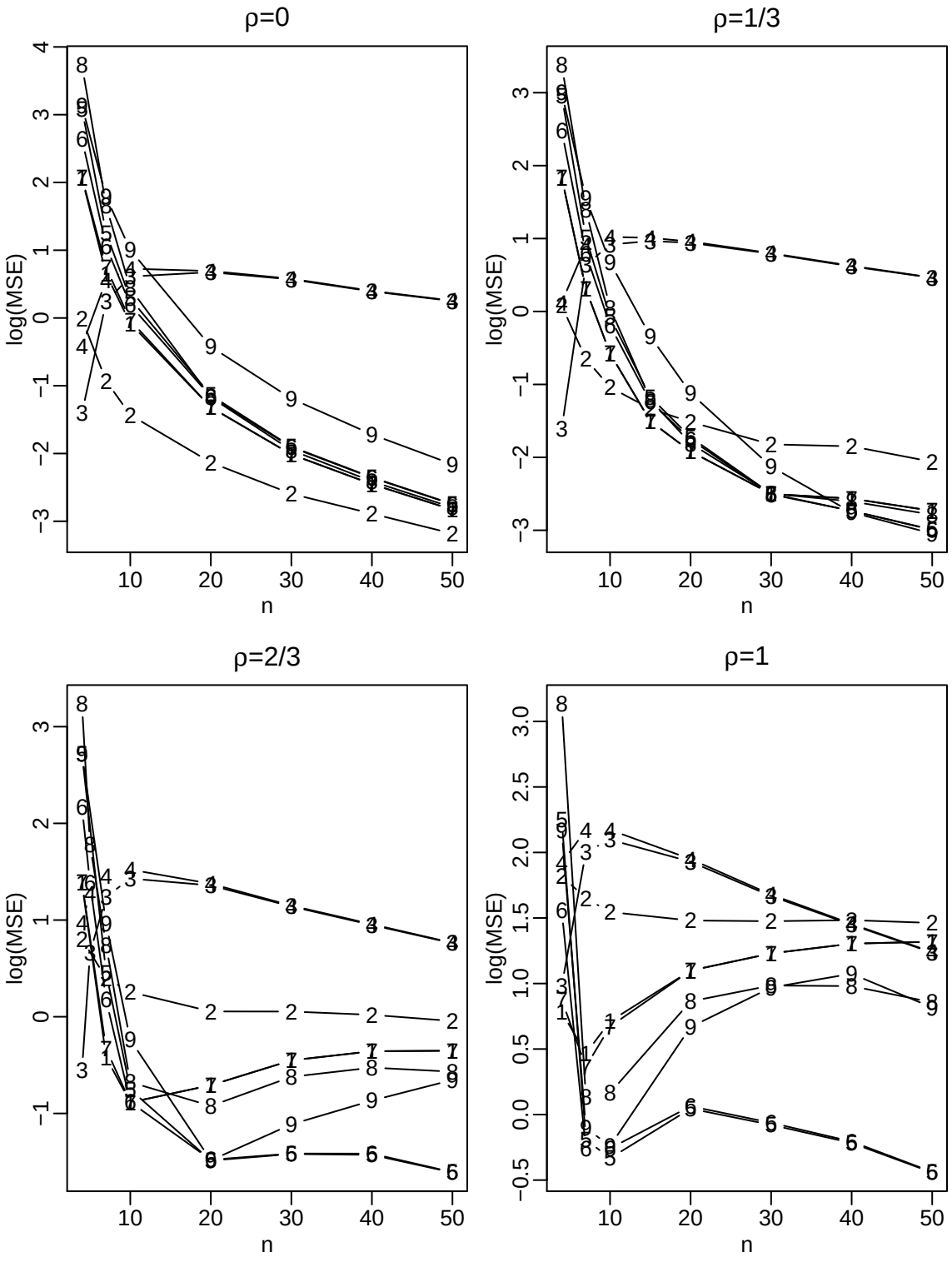

Figure 4.1: MSEs for Setup I and p=50. The sample size ranges from 4 to 50. In all figures, "1" to "9" represent Methods 1 to 9, respectively.

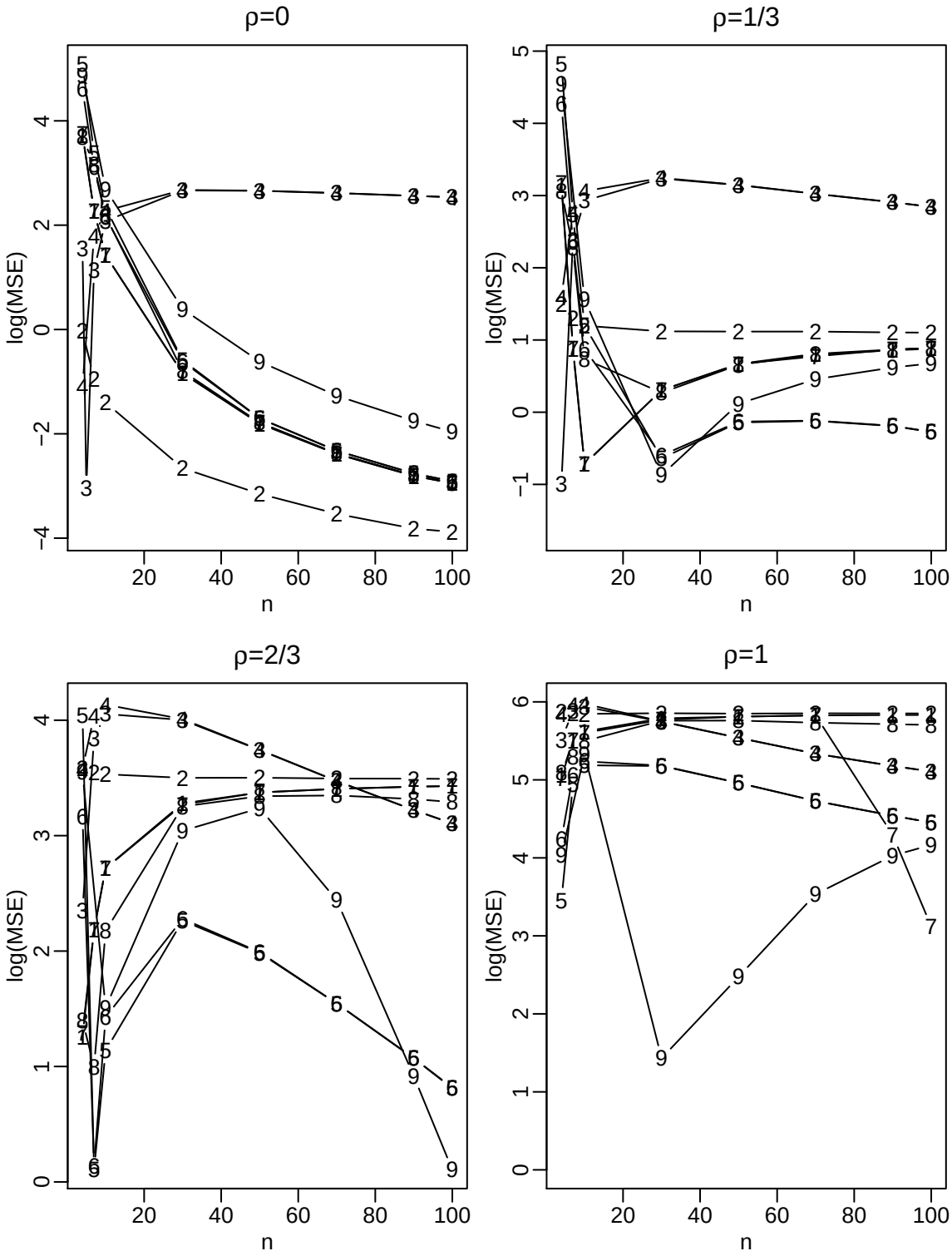

Figure 4.2: MSEs for Setup I and p=300. The sample size ranges from 4 to 100.

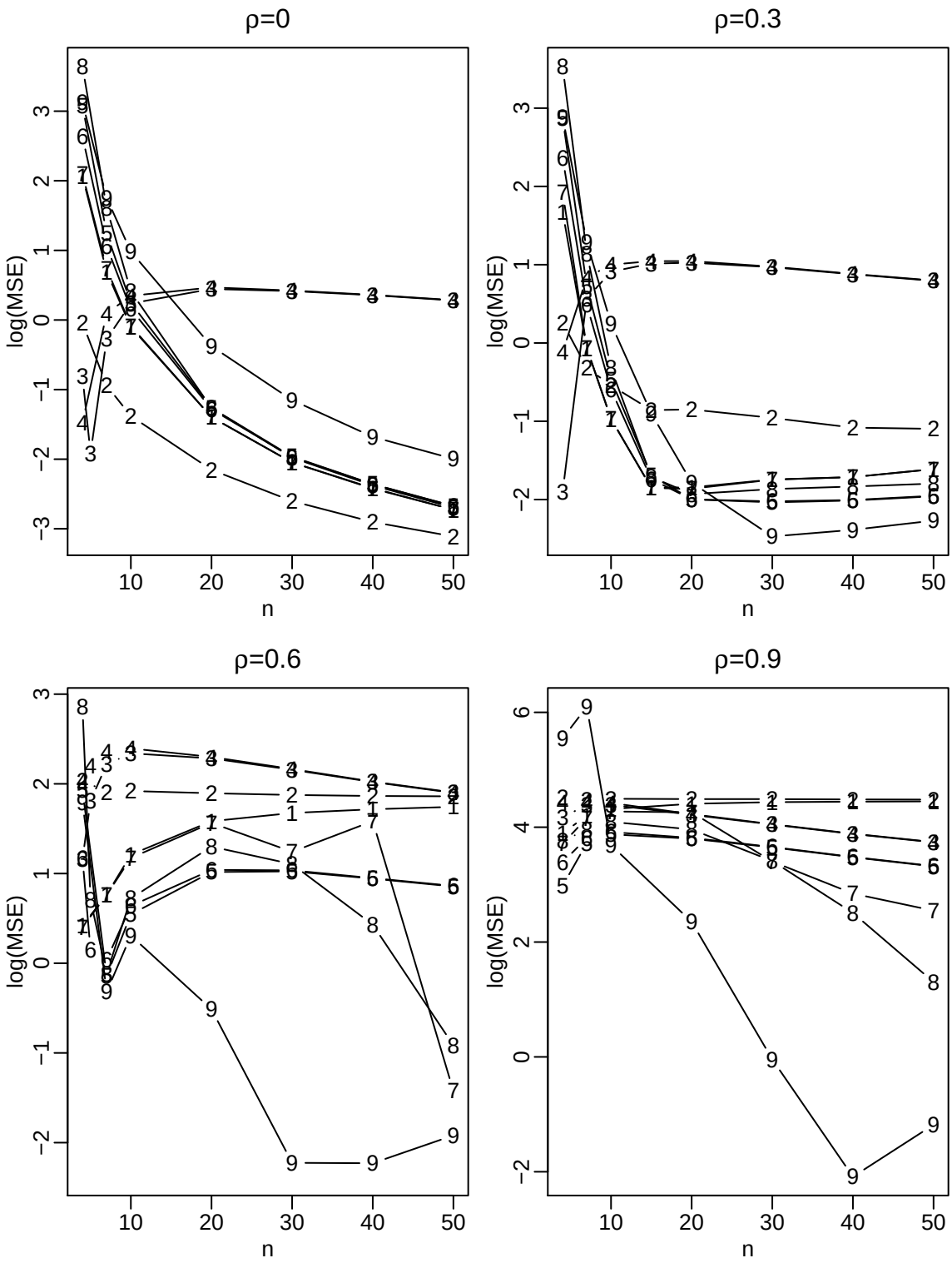

Figure 4.3: MSEs for Setup II and p=50. The sample size ranges from 4 to 50.

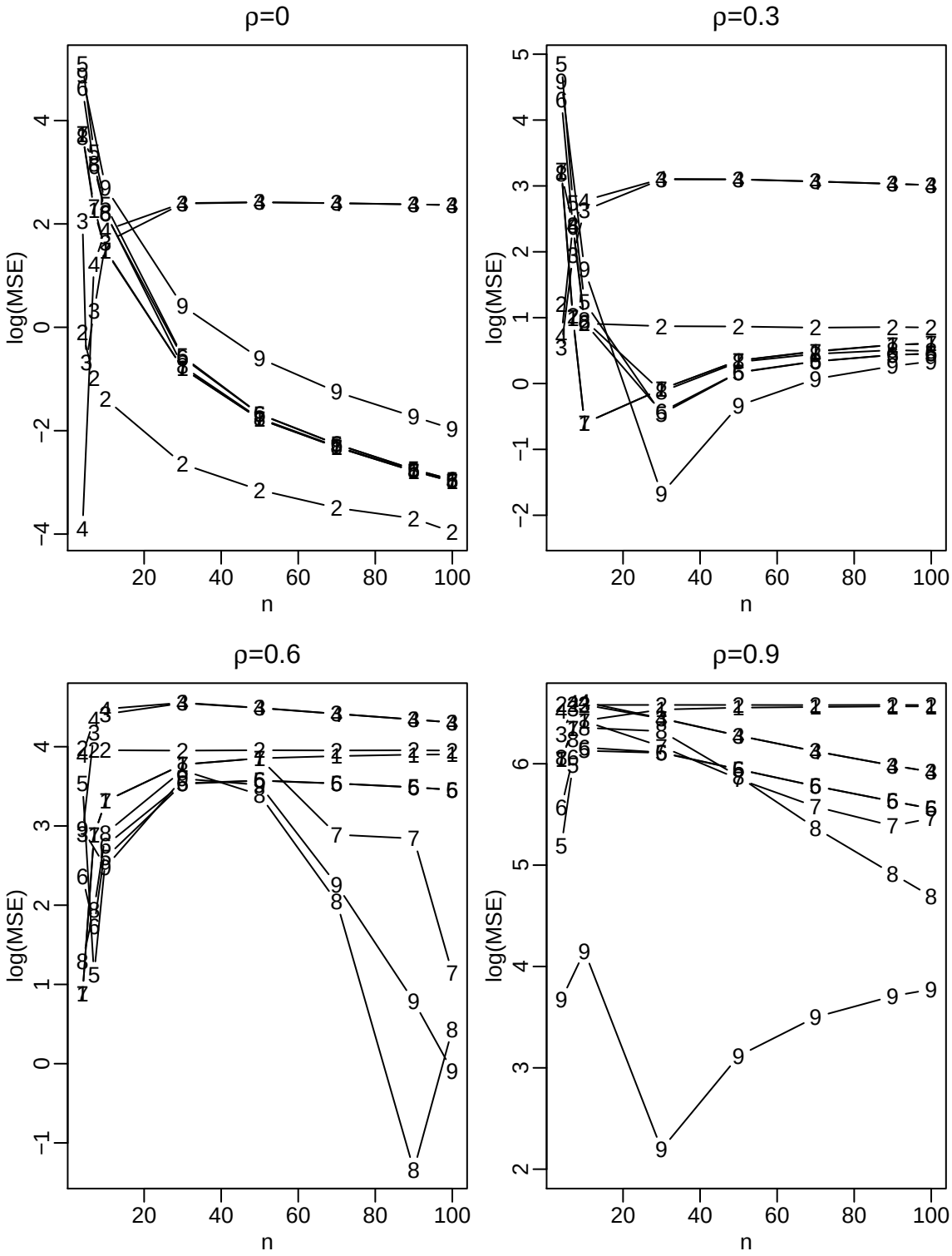

Figure 4.4: MSEs for Setup II and p=300. The sample size ranges from 4 to 100.

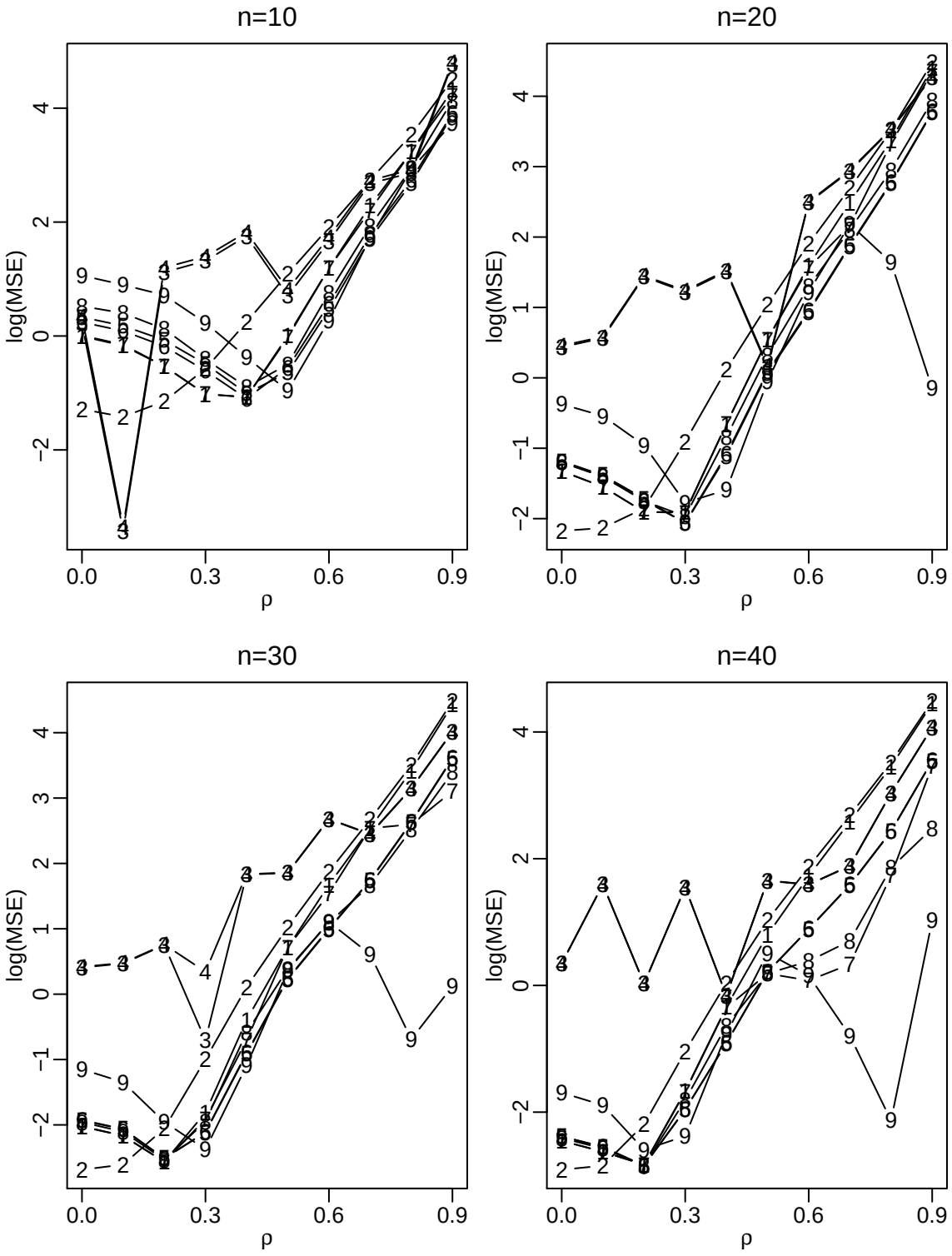

Figure 4.5: MSEs for Setup II and p=50 with different ρ ranging from 0.1 to 0.9.

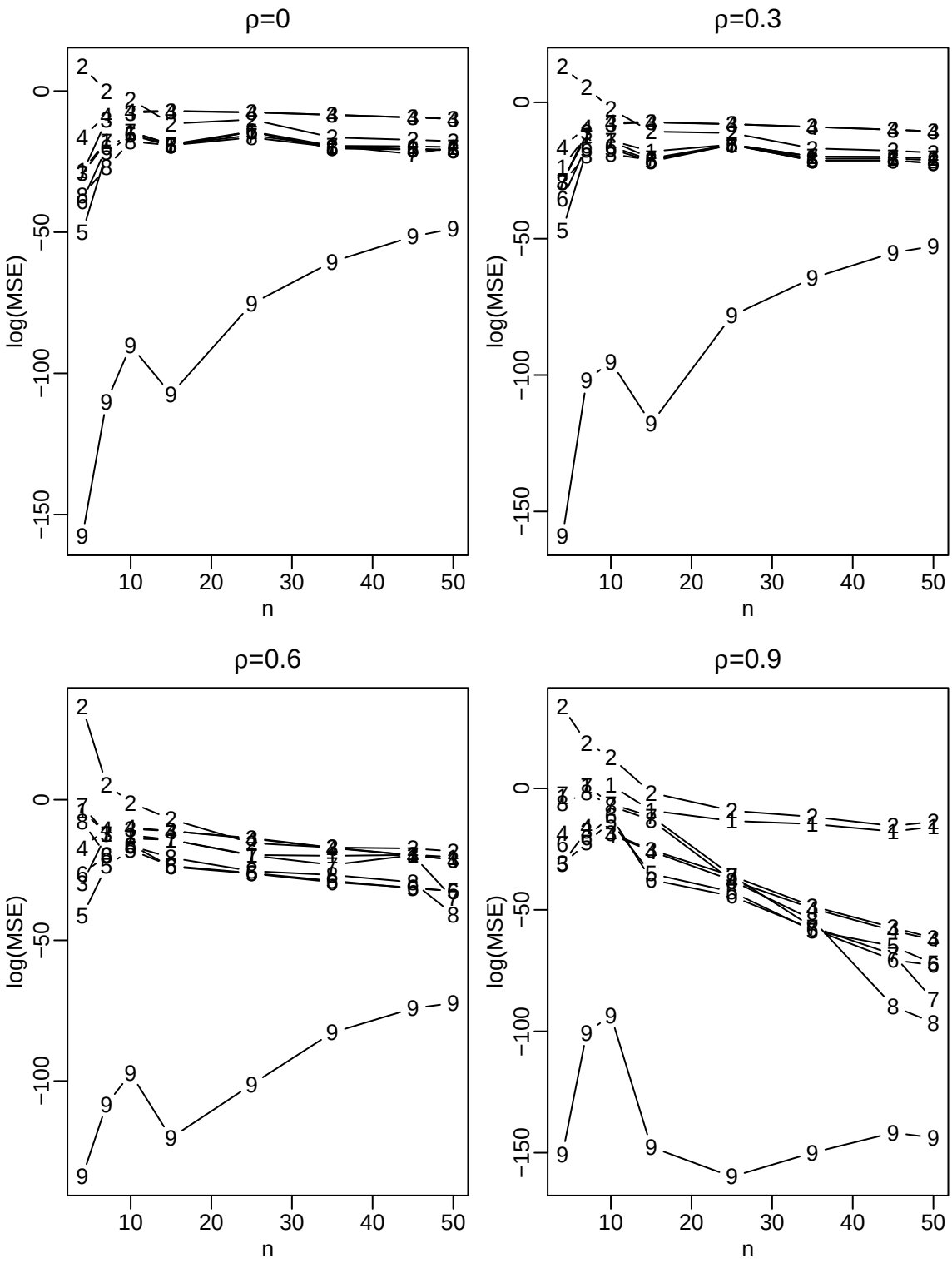

Figure 4.6: MSEs for Setup III and p=50. The sample size ranges from 4 to 50.

Chapter 5

NBLDA: Negative Binomial Linear Discriminant Analysis for RNA-Seq Data

5.1 Introduction

RNA-sequencing (RNA-Seq) is a revolutionary technology that uses the capabilities of next-generation sequencing to quantify gene expression levels (Mardis; 2008; Wang et al.; 2009; Morozova et al.; 2009). Compared to microarray technology, RNA-Seq has many advantages including the detection of novel transcripts, low background signal, and the increased specificity and sensitivity. Due to reduced sequencing cost, RNA-Seq has been widely used in biomedical research in recent years (Lorenz et al.; 2014). In real-world applications, the gene expression profile of biopsy or serum sample from an individual can be used to test whether this individual has a disease and/or a specific type of disease, which is essentially a classification problem. Different from the microarray technology that measures the level of gene expression on a continuous scale, RNA-Seq counts the number of reads that are mapped to one gene and measures the level of gene expression with nonnegative integers. As a result, popular tools that assume a Gaussian distribution in microrray data analysis, such as linear discriminant analysis, may not perform as well as those methods that adopt

appropriate discrete distributions for RNA-Seq data.

For RNA-Seq data, the Poisson distribution and negative binomial distribution are two common distributions considered in the expression detection and classification. Many methods have been proposed to detect differentially expressed genes, including edgeR (Robinson and Smyth; 2008; Robinson et al.; 2010), DESeq (Anders and Huber; 2010), baySeq (Hardcastle and Kelly; 2010), BBSeq (Zhou et al.; 2011), SAMseq (Li and Tibshirani; 2013), DSS (Wu et al.; 2013), AMAP (Si and Liu; 2013), sSeq (Yu et al.; 2013), and LFCseq (Lin et al.; 2014). However, there is less progress on the classification using RNA-Seq data until recently. Witten (2011) proposed a Poisson linear discriminant analysis (PLDA) which assumes that RNA-Seq data follow the Poisson distribution. Tan et al. (2014) further discussed many methods, such as logistic regression and partial least squares, and showed that PLDA is a comparable method. The Poisson distribution is suitable for modeling RNA-Seq data when biological replicates are not available. However, if biological replicates are available, the Poisson distribution may not be a proper choice owing to the overdispersion issue, where the variances of such data are likely to exceed their means (Anders and Huber; 2010; Si and Liu; 2013). The overdispersion issue can have a significant effect on classification accuracies. In real-world applications, biological replicates can provide more convincing results than technical replicates. Therefore, it is necessary to look for some solutions to take the overdispersion issue into consideration.

We note that Witten (2011) has considered this problem and pointed out that the classification accuracy can be further improved for overdispersed data by extending the Poisson model to the negative binomial model. However, to construct an appropriate negative binomial classifier for practical use, two major issues remain to be solved. The first issue is that the probability density function (pdf) of the negative binomial distribution is more complicated than that of the Poisson distribution, which gives rise to a more complicated classifier. The second issue is that the negative binomial distribution contains a dispersion parameter, which controls how much its variance exceeds its mean. To construct the classifier using the negative binomial model, we need to estimate the dispersion parameter. To avoid fitting the compli-

cated negative binomial model, Witten (2011) proposed a transformation method for the overdispersed data and found that this method works well if the overdispersion is mild.

In light of the importance of the dispersion in modelling RNA-Seq data with the negative binomial distribution, some dispersion estimation methods have been proposed recently in the literature. For example, Wu et al. (2013) proposed a dispersion estimator using empirical Bayes method and applied it to find differentially expressed genes. Yu et al. (2013) proposed a shrinkage estimator of dispersion which shrinks the estimates obtained by the method of moments towards a target value, and also applied it to detect differentially expressed genes. These new methods on estimating the dispersion parameter make it possible to construct a negative binomial classifier to achieve better classification accuracy on RNA-Seq data.

In this chapter, we propose a negative binomial linear discriminant analysis (NBLDA) for RNA-Seq data. The main contributions of this chapter are in, but not limited to, the following two aspects:

1. We extend Witten (2011) to build a new classifier based on the negative binomial model. Under the assumption of independent genes, we define the discriminant score by Bayes' rule and propose some plug-in rules for estimating the unknown parameters in the classifier.

2. We further explore the relationship between NBLDA and PLDA. A numerical comparison is conducted to explore how the dispersion changes the discriminant score. The comparison results will provide some guidelines for scientists to decide which method should be used in the discriminant analysis of RNA-Seq data.

To demonstrate the performance of our proposed method, we conduct several simulation studies under different numbers of genes, sample sizes, and proportions of differentially expressed genes. Simulation results show that the proposed NBLDA outperforms existing methods in many settings. Four real RNA-Seq data sets are also analyzed to demonstrate the advantage of NBLDA. Specifically, we propose the

negative binomial classifier study, the relationship between NBLDA and PLDA, and present the parameter estimation in Section 5.2. Simulation studies and real data analysis are conducted in Sections 5.3 and 5.4. We conclude this chapter with some discussions in Section 5.5.

5.2 Negative Binomial Linear Discriminant Analysis

Let X_{ig} denote the numbers of reads mapped to gene g in sample i, $i = 1, \ldots, n$ and $g = 1, \ldots, G$. Our goal is to identify which class a new observation belongs to. Witten (2011) proposed a PLDA for classifying RNA-Seq data. When biological replicates are available, however, overdispersion occurs for RNA-Seq data and hence the Poisson distribution may no longer be appropriate. In this section, we propose a new discriminant analysis for RNA-Seq data by assuming that the data follow the negative binomial distribution.

5.2.1 Methodology

Consider the following negative binomial distribution for RNA-Seq data:

$$X_{ig} \sim \text{NB}(\mu_{ig}, \phi_g), \qquad \mu_{ig} = s_i \lambda_g, \tag{5.1}$$

where s_i is the size factor which is used to scale gene counts for the ith sample due to different sequencing depth, λ_g is the total number of reads per gene, and $\phi_g \geq 0$ is the dispersion parameter. We have $E(X_{ig}) = \mu_{ig}$ and $\text{Var}(X_{ig}) = \mu_{ig} + \mu_{ig}^2 \phi_g$. Note that the variance is larger than the mean for the negative binomial distribution.

Let K be the total number of classes and $C_k \in \{1, \ldots, n\}$ the indices of samples in class k for $k = 1, \ldots, K$. Then the class-specific model for RNA-Seq data is given by

$$(X_{ig} | y_i = k) \sim \text{NB}(\mu_{ig} d_{kg}, \phi_g), \tag{5.2}$$

where d_{kg} represents the differences among K classes, and $y_i = k \in \{1, \ldots, K\}$ represents the label of sample i. We also follow the independence assumption in Witten (2011) that all genes are independent of each other. Note that the independence assumption is frequently assumed in microarray data analysis.

Let $\mathbf{x}^* = (X_1^*, \ldots, X_G^*)^T$ be a test sample with s^* the size factor and y^* the class label. By Bayes' rule, we have

$$P(y^* = k|\mathbf{x}^*) \ \propto \ f_k(\mathbf{x}^*)\pi_k, \tag{5.3}$$

where f_k is the pdf of the sample in class k, and π_k is the prior probability that one sample comes from class k. The pdf of $X_{ig} = x_{ig}$ in model (5.2) is

$$\begin{aligned} P(X_{ig} = x_{ig}|y_i = k) \ &= \ \frac{\Gamma(x_{ig} + \phi_g^{-1})}{x_{ig}!\Gamma(\phi_g^{-1})} \left(\frac{s_i\lambda_g d_{kg}\phi_g}{1 + s_i\lambda_g d_{kg}\phi_g}\right)^{x_{ig}} \\ &\quad \left(\frac{1}{1 + s_i\lambda_g d_{kg}\phi_g}\right)^{\phi_g^{-1}}. \end{aligned} \tag{5.4}$$

By (5.3) and (5.4), we have the following discriminant score for NBLDA:

$$\begin{aligned} \log P(y^* = k|\mathbf{x}^*) \ &= \ \sum_{g=1}^{G} X_g^* \left[\log d_{kg} - \log(1 + s^*\lambda_g d_{kg}\phi_g)\right] \\ &\quad - \sum_{g=1}^{G} \phi_g^{-1} \log(1 + s^*\lambda_g d_{kg}\phi_g) \\ &\quad + \log \pi_k + C, \end{aligned} \tag{5.5}$$

where C is a constant independent of k. We then assign the new observation $\mathbf{x}^*$ to class k that maximizes the quantity (5.5). Throughout this chapter, we estimate the prior probability π_k by n_k/n, where n_k is the sample size in class k. For balanced data, the prior probability is simplified as $\pi_k = 1/K$ for all $k = 1, \ldots, K$. For gene g, the total number of reads is $\lambda_g = \sum_{i=1}^{n} X_{ig}$, and the class difference d_{kg} can be estimated by $(\sum_{i \in C_k} X_{ig} + 1)/(\sum_{i \in C_k} s_i\lambda_g + 1)$. Estimation of the unknown parameters including s_i and ϕ_g will be discussed in Section 5.2.2.

To explore the relationship between the proposed NBLDA and the PLDA in Witten (2011), we assume that $s^*\lambda_g d_{kg}$ are bounded. When $\phi_g \to 0$, we have $\log(1 + s^*\lambda_g d_{kg}\phi_g) \to 0$ and $\phi_g^{-1} \log(1 + s^*\lambda_g d_{kg}\phi_g) = \log(1 + s^*\lambda_g d_{kg}\phi_g)^{\phi_g^{-1}} \to s^*\lambda_g d_{kg}$. Then

consequently,

$$\log P(y^* = k|\mathbf{x}^*) \approx \sum_{g=1}^{G} X_g^* \log d_{kg} - \sum_{g=1}^{G} s^* \lambda_g d_{kg} + \log \pi_k + C, \tag{5.6}$$

where the right hand of (5.6) is the discriminant score of PLDA. That is, the NBLDA classifier reduces to the PLDA classifier when there is little dispersion in the data. From this point of view, the proposed NBLDA can be treated as a generalized version of PLDA.

Since NBLDA contains the dispersion parameter which PLDA does not have, in what follows, we investigate how the dispersion changes their discriminant scores. We conduct a numerical comparison between NBLDA and PLDA. Two cases are considered in this chapter. The first case is that all genes have a common dispersion, and the second is that genes have different dispersions. Note that the classifiers (5.5) and (5.6) have the same terms: $\log \pi_k$ and C. Without loss of generality, we compute the discriminant scores only using the first two terms in (5.5) and (5.6), respectively. In the comparison study, we fix $X_g^* = 10$, $d_{kg} = 1.5$, $s^* = 1$, $\lambda_g = 10$ and $G = 500$. For the case of common dispersion, we set the dispersion ranging from 0 to 20. For the case of different dispersions, we let ϕ_g be i.i.d. random variables from a chi-squared distribution with the degrees of freedom ranging from 0.1 to 5.

Figure 5.1 exhibits the comparison results. The left panel shows the results for the case of common dispersion. Note that the discriminant score of PLDA is independent of the dispersion parameter and hence is a constant. For NBLDA, its discriminant score is a curve, and the slope is large for low dispersions and small for high dispersions. We discover that the discriminant score of NBLDA is sensitive to the dispersion. Even when the dispersion is very small, the difference between the two discriminant scores is significant. The right panel in Figure 5.1 shows the results for the case of different dispersions. The pattern of the right panel is similar to the left one except that the curve of NBLDA is not smooth. This suggests that when we analyze real data, we should first compute its average dispersion and then use such information to determine which classifier to use.

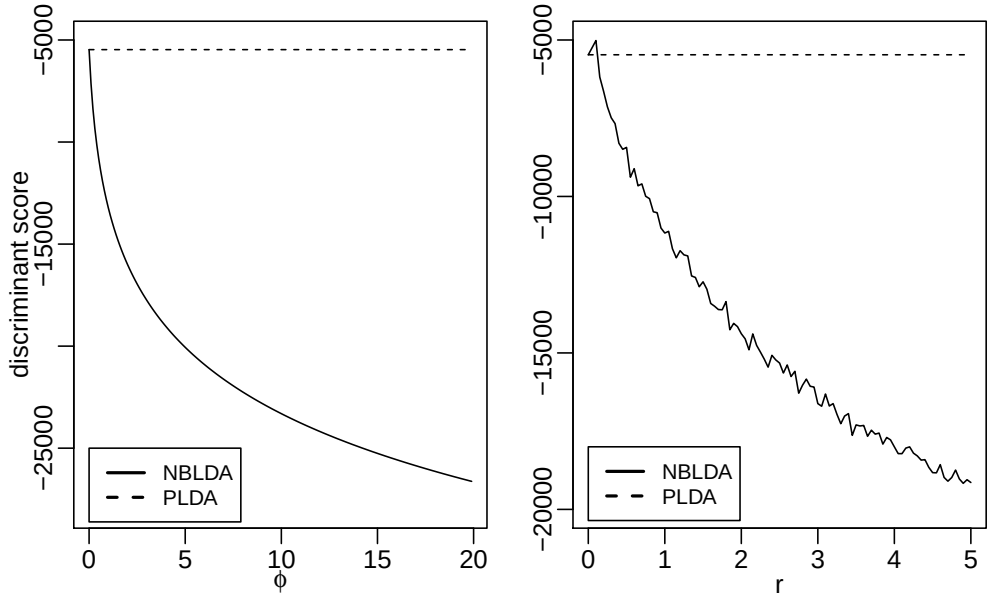

Figure 5.1: Numerical comparisons between NBLDA and PLDA. The left panel shows the results with a common dispersion ϕ. The right panel shows the results with different gene-specific dispersions ϕ_g which are i.i.d. random variables from a chi-squared distribution with r degrees of freedom. We compute the discriminant scores of NBLDA and PLDA for different ϕ and r.

5.2.2 Parameter Estimation

Note that the discriminant score in (5.5) involves two unknown parameters, size factor s^* and dispersion parameter ϕ_g.

Size factor estimation

Due to different sequencing depths, samples have different total numbers of reads. Hence a normalization of the read counts through a size factor is a necessary step for analyzing RNA-Seq data (Bullard et al.; 2010; Dillies et al.; 2013). To estimate the size factor s_i for the training data and the size factor s^* for the test data, we consider the following three procedures:

- *Total count*: Witten (2011) divided the total read counts of sample i by the total read counts of all samples to estimate the size factor of sample i. That is,

$$\hat{s}^* = \frac{\sum_{g=1}^{G} X_g^*}{\sum_{i=1}^{n}\sum_{g=1}^{G} X_{ig}},$$

$$\hat{s}_i = \frac{\sum_{g=1}^{G} X_{ig}}{\sum_{i=1}^{n}\sum_{g=1}^{G} X_{ig}}.$$

- *DESeq*: Anders and Huber (2010) first divided the read counts of sample i by the geometric mean of all samples' read counts, and then estimated the size factor by computing the median of those G values. Specifically, the size factors are estimated by

$$\hat{s}^* = \text{median}_g \frac{X_g^*}{(\prod_{i=1}^{n} X_{ig})^{1/n}},$$

$$\hat{s}_i = \text{median}_g \frac{X_{ig}}{(\prod_{l=1}^{n} X_{lg})^{1/n}}.$$

- *Upper quartile*: Bullard et al. (2010) proposed a robust method that uses the upper quartile of the read counts to estimate the size factors. Specifically, the size factors are estimated by

$$\hat{s}^* = \frac{q^*}{\sum_{i=1}^{n} q_i},$$

$$\hat{s}_i = \frac{q_i}{\sum_{i=1}^{n} q_i},$$

where q^* and q_i are the upper quartiles for the test data and sample i in the training data, respectively.

In our simulation studies, we find that the three methods provide little difference in the performance of classification. Hence, for brevity, we only report in the remainder of this chapter the simulation results based on the total count method.

Dispersion parameter estimation

Various methods for estimating the dispersion parameter ϕ_g have been proposed in the literature (Robinson and Smyth; 2008; Robinson et al.; 2010; Anders and Huber; 2010; Hardcastle and Kelly; 2010). A comparative study is also available in Landau and Liu (2013) where the authors investigated the influence of different dispersion parameter estimates on detecting differentially expressed genes in RNA-Seq data. More recently, Yu et al. (2013) proposed a shrinkage estimator for ϕ_g that shrinks the gene-specific estimation towards a target value. Specifically, the dispersion estimator is estimated by

$$\hat{\phi}_g = \delta\xi + (1-\delta)\tilde{\phi}_g, \tag{5.7}$$

where δ is a weight defined as

$$\delta = \frac{\sum_{g=1}^{G}\left\{\tilde{\phi}_g - (1/G)\sum_{g=1}^{G}\tilde{\phi}_g\right\}^2/(G-1)}{\sum_{g=1}^{G}\left(\tilde{\phi}_g - \xi\right)^2/(G-2)},$$

$\tilde{\phi}_g$ are the initial dispersion estimates obtained by the method of moments, and ξ is the target value calculated by minimizing the average squared difference between $\tilde{\phi}_g$ and $\hat{\phi}_g$. In this chapter, we use the estimator (5.7) to estimate the dispersion parameter.

5.3 Simulation Studies

In this section, we evaluate and compare the following classification methods:

- NBLDA,

- PLDA,
- Support vector machines (SVM),
- K-nearest neighbors (KNN).

For PLDA, we use the R package "PoiClaClu" provided in Witten (2011). For SVM, we use the R package "e1071" and choose the radial basis kernel in our simulation studies. For KNN, we choose $k = 1$, 3 and 5.

5.3.1 Simulation Design

We generate the data from the following negative binomial distribution:

$$(X_{ig}|y_i = k) \sim \text{NB}(s_i\lambda_g d_{kg}, \phi). \tag{5.8}$$

The total number of classes is $K = 2$, and both the training data and test data have n samples. In all G genes, the proportions of differentially expressed genes are 0.2, 0.4, 0.6, 0.8 and 1.0, which represents that 20%, 40%, 60%, 80% and 100% genes are differentially expressed, respectively. For the differentially expressed genes, we set $\log d_{kg} = z_{kg}$, where z_{kg} are i.i.d. random variables from the normal distribution $N(0, \sigma^2)$. For the constant genes, we set $d_{kg} = 1$. The size factors s_i are i.i.d. random variables from the uniform distribution on [0.2, 2.2]. The λ_g values are i.i.d. random variables from the exponential distribution with rate 0.04. Note that, for the sake of fairness, we have essentially followed the same simulation settings as those in Witten (2011). For the values of G, n, ϕ and σ, we specify them in Figures 5.2, 5.3 and 5.4.

To compare these methods, we compute the mean misclassification rates as follows: for each simulation, we generate n test samples and compute the following misclassification rate:

$$\frac{\text{the number of misclassified samples}}{n}.$$

We run 1,000 simulations, compute its mean, and then obtain the mean misclassification rate.

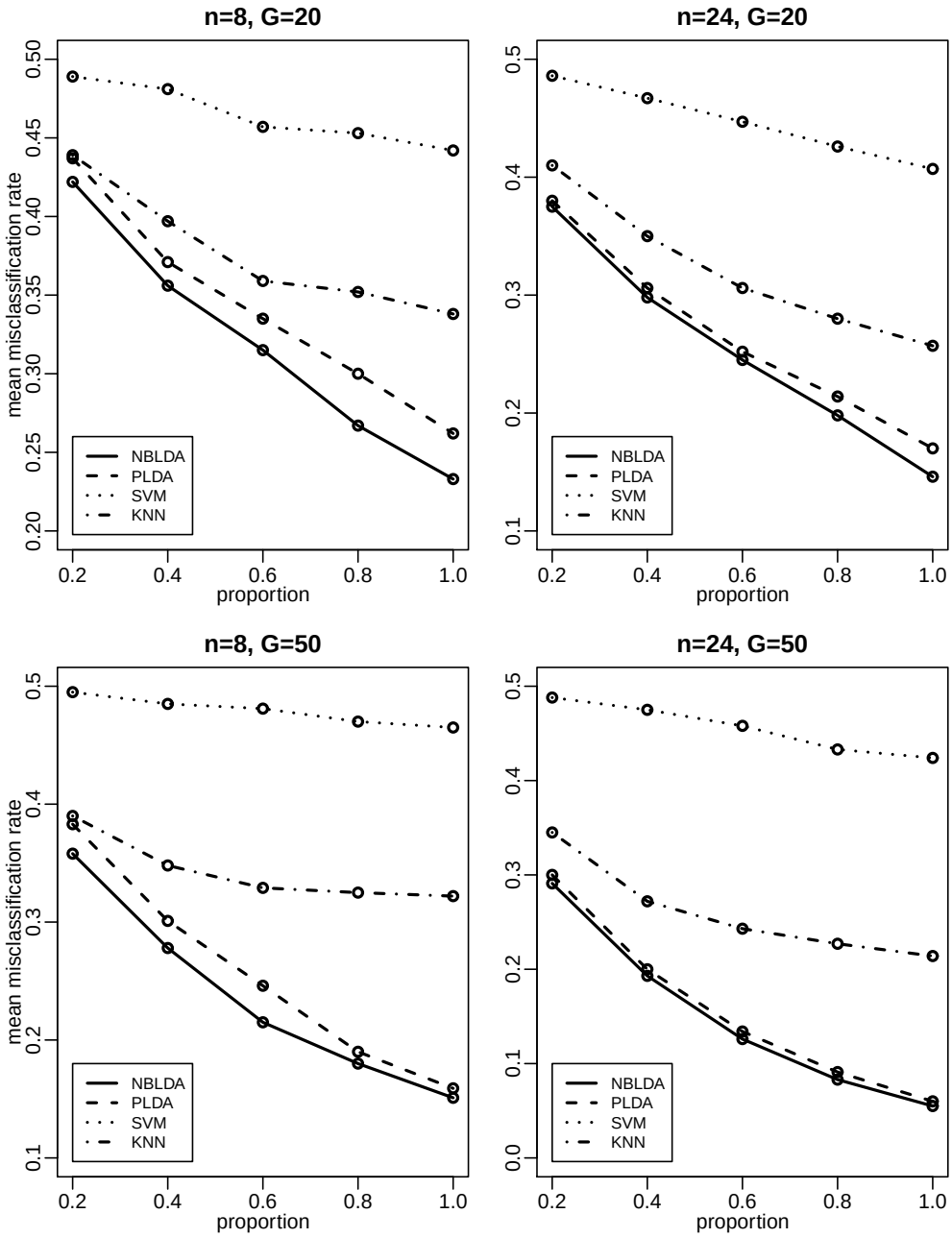

Figure 5.2: Mean misclassification rates for all four methods with $\phi = 20$ and $\sigma = 5$. The x-axis represents the proportion of differentially expressed genes. 20%, 40%, 60%, 80% and 100% differentially expressed genes are considered, respectively. These plots investigate the effect of proportion of differentially expressed genes.

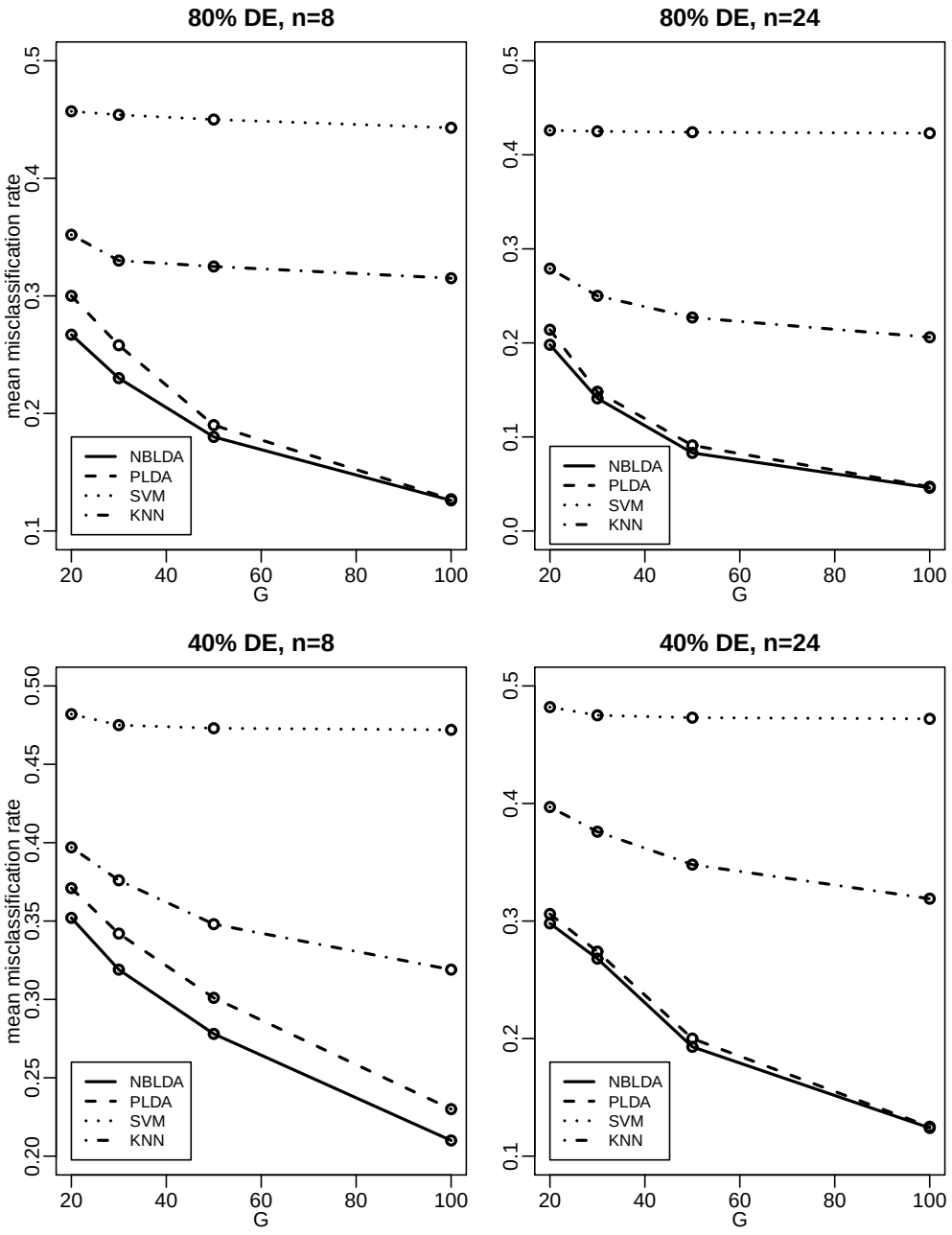

Figure 5.3: Mean misclassification rates for all four methods with $\phi = 20$ and $\sigma = 5$. "80% DE" means 80% genes are differentially expressed, and the same to "40% DE". This plot investigates the effect of numbers of genes.

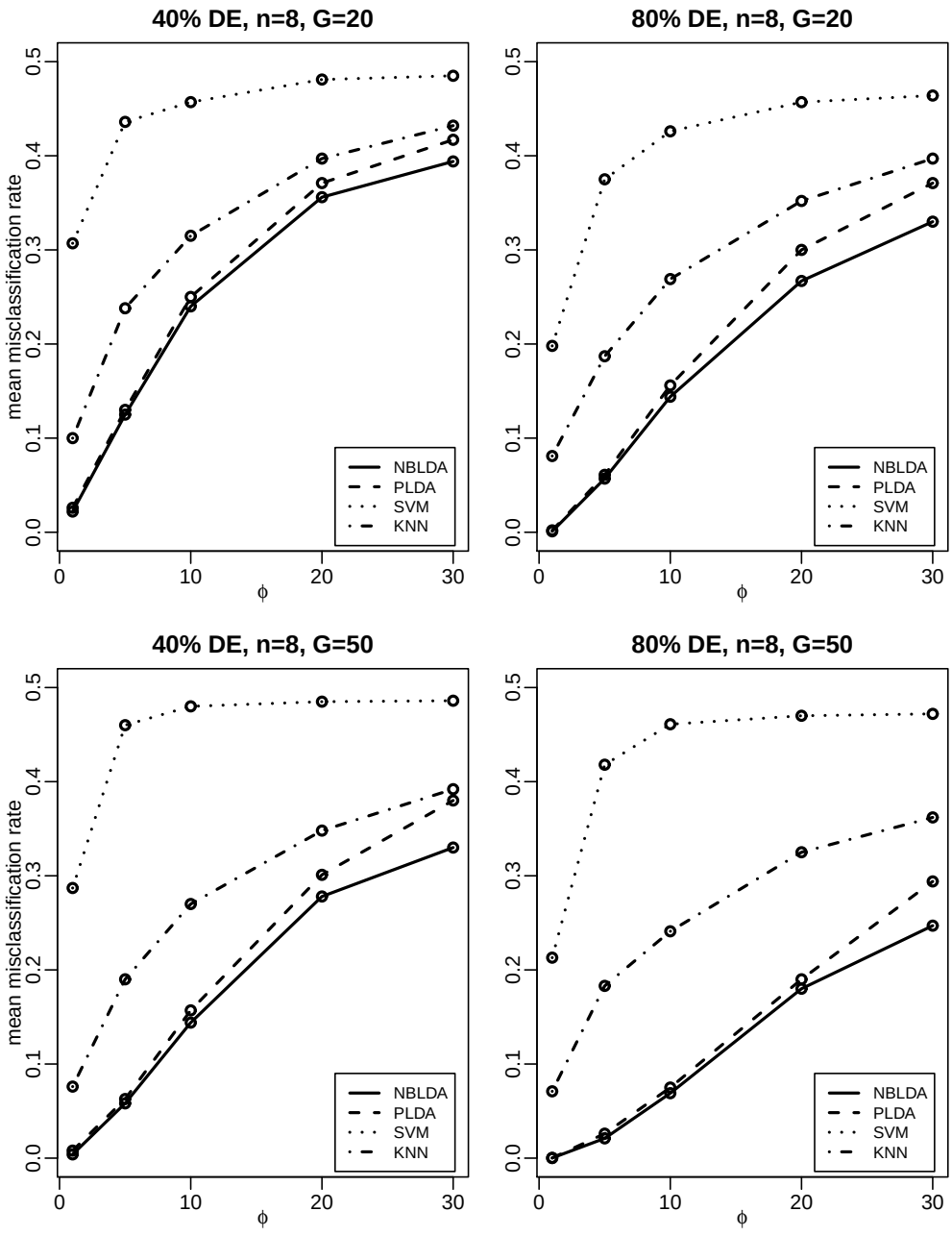

Figure 5.4: Mean misclassification rates for all four methods with $\sigma = 5$. "80% DE" means 80% genes are differentially expressed, and the same to "40% DE". This plot investigates the effect of overdispersion.

5.3.2 Simulation Results

Figure 5.2 investigates the effect of the proportion of differentially expressed genes on the mean misclassification rate. In general, with an increasing number of differentially expressed genes, both methods have decreased mean classification rates. NBLDA always outperforms the other three methods. In particular, when the sample size is small ($n = 8$), NBLDA has a significant improvement over the other approaches.

Figure 5.3 investigates the impact of the number of genes on the mean misclassification rate. We consider $G = 20$, 30, 50, and 100 for this investigation. From Figure 5.3, we observe that an increasing number of genes will lead to a lower misclassification rate. NBLDA shows its superiority over the other three methods, and the improvement is more significant when the sample size and the number of genes are smaller.

Figure 5.4 investigates the effect of overdispersion on the mean misclassification rate. We consider $\phi = 1$, 5, 10, 20 and 30 for this investigation. Figure 5.4 shows that a larger dispersion will result in a higher mean misclassification rate. Both NBLDA and PLDA perform better than SVM and KNN. When the overdispersion is not very high, NBLDA and PLDA have similar performance, with NBLDA slightly better than PLDA. When the overdispersion is high, however, the performance of NBLDA is much better than PLDA.

For real biomedical research in which RNA-Seq technology is used, it is common that thousands or tens of thousands of genes are measured simultaneously. We perform a gene selection procedure to screen the informative genes before applying a classification rule to RNA-Seq data. By doing gene selection, we rule out the noise as much as possible so that the variance of the discriminant score is reduced, and consequently we have an increased interpretability. For more details, see Section 5.4.

5.4 Real Data Analysis

We first describe four data sets. The first three are RNA-Seq data and the last one is a chromatin immunoprecipitation (ChIP) sequencing data set.

- *Liver and kidney data* (Marioni et al.; 2008). There are two classes in this data set. One class contains 7 technical replicates which come from a liver sample. The other class contains 7 technical replicates which come from a kidney sample. A total of 22,925 genes are measured in this data set. The data set is available as a Supplementary File in Marioni et al. (2008).

- *Yeast data* (Nagalakshmi et al.; 2008). The data set contains two library preparations: random haxamer (RH) and oligo (dT), which are treated as two classes in this paper. In each class, three samples are included: one original sample, its technical replicate, and its biological replicate. A total of 6,874 genes are quantified in this data set. The data set is available as a Supplementary File in Anders and Huber (2010).

- *Cervical cancer data* (Witten et al.; 2010). Two groups of samples are contained in this data set. One is the nontumor group which includes 29 samples, and the other one is the tumor group which includes 29 samples. There are 714 microRNAs in this data set. This data set is available in Gene Expression Omnibus (GEO) Datasets with access number GSE20592.

- *Transcription factor binding data* (Kasowski et al.; 2010). This data set contains 10 classes with a total of 39 samples. 19,061 binding regions are included in this data set and those regions are treated as distinct features. This data set is available as a Supplementary File in Anders and Huber (2010).

5.4.1 Gene Selection

The BSS/WSS method, which is proposed by Dudoit et al. (2002), is a common gene selection method and has been widely used in the literature (Lee et al.; 2005b; Pang et al.; 2009; Huang et al.; 2010). This method computes the ratio of the sum of squares between groups to the sum of squares within groups for each gene, and selects genes whose ratios are in the top. However, this method assumes the data to be normally distributed so that it may not be suitable for RNA-Seq data.

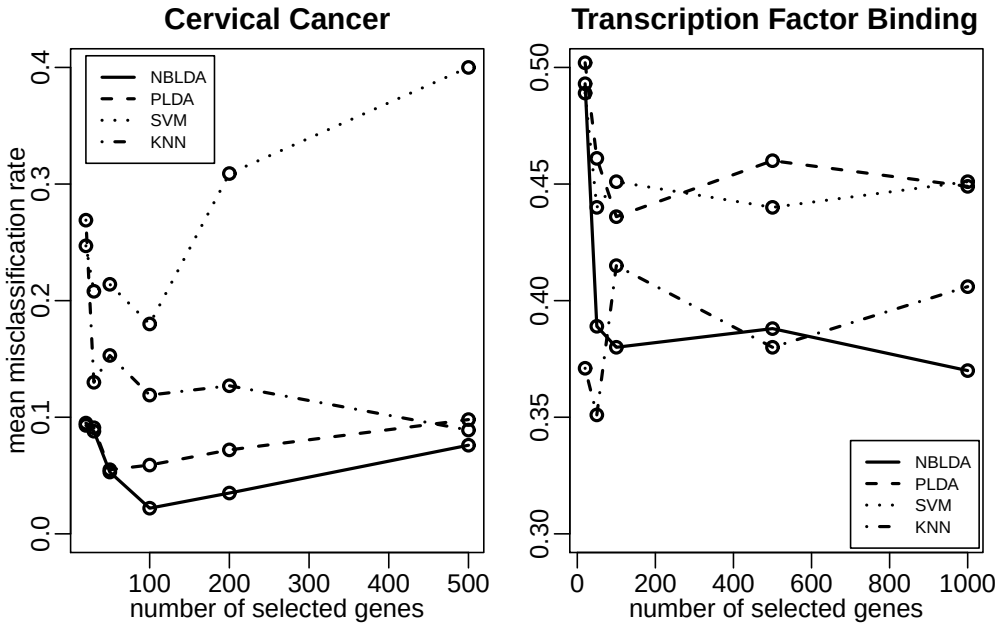

Figure 5.5: Mean misclassification rates for Cervical cancer data and Transcription factor binding data.

Witten (2011) proposed a screening method to select genes for RNA-Seq data. Since gene g will be deleted from the classification rule, $d_{kg} = 1$, they shrink the estimate of d_{kg} towards 1 by using soft-thresholding to perform the gene selection procedure. However, this method can not be applied to our discriminant analysis because the dispersion is involved in our discriminant rule. For the negative binomial distribution, edgeR (Robinson and Smyth; 2008; Robinson et al.; 2010) has been proposed to detect differentially expressed genes in RNA-Seq data. This method first estimates the gene-wise dispersions by maximizing the combination of gene-specific conditional likelihood and common conditional likelihood, and then replaces the hypergeometric distribution in Fisher's exact test by the negative binomial distribution to construct an exact test. In this chapter, we use edgeR to perform the gene selection procedure, which is available in Bioconductor (www.bioconductor.org).

5.4.2 Results

We first conduct the gene selection procedure using edgeR and obtain G genes for further analysis. We then randomly split the sample into two sets: the training set and the test set. The training set is used to construct the classifier and the test set is used to compute the misclassification rate. We repeat the whole procedure 1,000 times and compute the mean misclassification rate for the four methods, NBLDA, PLDA, SVM, and KNN, respectively.

The comparison results are shown in Figure 5.5. Because the mean misclassification rates of the four methods are all zeros for Liver and kidney data and Yeast data, we only show the results for other two data sets in Figure 5.5. For Cervical cancer data, 52 samples are assigned to the training set and 6 samples to the test set. A total of 20, 30, 50, 100, 200 and 500 genes are selected, respectively. Among all approaches we consider in this chapter, our proposed NBLDA has the lowest misclassification rate. A big improvement over the other approaches can be observed when more than 50 genes are selected. For Transcription factor binding data, to conduct the binary classification, we randomly assign 30 samples to the training set and the remaining 9 samples to the test set. We choose 20, 50, 100, 500 and 1,000 genes, respectively for this data set. In Figure 5.5, we observe that NBLDA also outperforms PLDA for Transcription factor binding data. Expect when the number of genes is small, NBLDA has a better or comparable performance than the other three methods.

Finally, we estimate the average dispersion of the two data sets to check if it also supports our comparison results made in the previous paragraph. The simplest way for estimating the dispersion is to use the method of moments. However, this estimate may not be reliable (sometimes is a negative value) when the sample size is small. Landau and Liu (2013) and Yu et al. (2013) recently reviewed several dispersion estimation methods. For Cervical cancer data and Transcription factor binding data, we compute their average dispersions using the method in Yu et al. (2013) and present the estimates in Table 5.1. We note that both data sets possess a considerably high average dispersion when the number of selected genes is not very large. This, together with the numerical comparison in Figure 5.1, explains why NBLDA provides a better

performance than PLDA for these two data sets.

Table 5.1: The average dispersions for Cervical cancer data and Transcription factor binding data, where "G" represents the number of top genes selected by edgeR.

Data sets	G=20	G=50	G=100	G=500
Cervical cancer	25.71	24.42	19.02	11.03
Transcription factor binding	8.12	5.71	4.48	2.86

5.5 Discussion

Next generation sequencing technology has been widely applied in biomedical research and RNA-Seq begins to replace the microarray technology gradually in recent years. Since RNA-Seq data are nonnegative integers, differing from that of microarray data, it is necessary to develop methods that are well suited for RNA-Seq data. Two discrete distributions, the Poisson distribution and negative binomial distribution, are commonly used in the literature to model RNA-Seq data. Compared to the Poisson distribution, the negative binomial distribution allows its variance to exceed its mean and is more suitable for the situations when biological replicates are available. Nevertheless, the negative binomial model is more complicated than the Poisson model as the additional dispersion parameter also needs to be estimated.

In this chapter, we have proposed an NBLDA classifier using the negative binomial model. Our simulation results show that our proposed NBLDA has a better performance than PLDA in the presence of moderate or high dispersions. When there is little dispersion in the data, NBLDA is also comparable to PLDA. We have further explored the relationship between NBLDA and PLDA, and investigated the impact of dispersion on the discriminant score of NBLDA by conducting a numerical comparison. It is worth noting that even for a small dispersion, the two discriminant scores can be rather different. This suggests that for real RNA-Seq data with moderate or high dispersion, NBLDA may be a more appropriate method than PLDA.

Note that the true dispersions are unlikely to be known in practice. Therefore, we propose to first estimate the average dispersion using some novel estimation methods in the recent literature. Second, if the estimated average dispersion is small, we use PLDA; and otherwise we use NBLDA.

We note that the independence assumption in Witten (2011) and in this chapter is very restrictive. For real gene expression data sets, it may not be realistic to assume that all genes are independent of each other. In our future study, we would like to incorporate the network information of pathways or gene sets to further improve the performance of classification. The clustering of sequencing data is also an important issue in biomedical research. Hence, another possible future work is to extend the clustering method in Witten (2011) to follow the negative binomial model. To conclude, our proposed method is general and can be applied to other next generation sequencing data sets including ChIP-Seq data.

Chapter 6

Summary

In this thesis, we considered estimating the high-dimensional covariance matrix and its determinant. And we applied high-dimensional covariance matrix estimation to Hotelling's tests. Besides, we considered the linear discriminant analysis for RNA-Sequencing data.

In Chapter 2, we proposed an optimal shrinkage estimation of the covariance matrices. This method estimated many covariance matrices simultaneously and to estimate one covariance matrix we shrink its sample covariance matrix towards the pooled sample covariance matrix through a shrinkage parameter. Some properties of the optimal shrinkage parameter and its estimation method were given. Simulation studies and real data analysis are conducted to investigate the performance of our methods.

In Chapter 3, we proposed a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. For high-dimensional small sample size data, the diagonal Hotelling's tests are better than the unscaled Hotelling's tests and the regularized Hotelling's tests. However, sample variance is an unreliable variance estimator for limited observations. Therefore, we used optimal shrinkage variance estimations to improve the performance of the diagonal Hotelling's test. The improvements were shown in our simulation studies. Consequently, we suggested using shrinkage-based diagonal Hotelling's tests to test the significance of gene sets with small sample sizes. Furthermore, if the number of genes in the gene sets is not large, the scaled chi-

squared null distribution is recommended.

In Chapter 4, we compared a total of nine methods for estimating the log-determinant, or equivalently the determinant, of high-dimensional covariance matrix. Three types of covariance structures were considered in our simulation studies. Overall, there is no single method that dominates other methods under all considered circumstances. In general, the sample size and the correlation of the data sets have a great impact on the accuracy of estimation. Consequently, we may select appropriate estimation methods according to the sample size and the prior information on the correlation structure of the covariates. In the situations, however, when such prior information is not available, we recommend to use POCTE to estimate the determinant of high-dimensional covariance matrix.

In Chapter 5, we proposed an NBLDA classifier using the negative binomial model. Our simulation results showed that our proposed NBLDA has a better performance than PLDA in the presence of moderate or high dispersions. When there is little dispersion in the data, NBLDA is also comparable to PLDA. We further explored the relationship between NBLDA and PLDA, and investigated the impact of dispersion on the discriminant score of NBLDA by conducting a numerical comparison. It is worth noting that even for a small dispersion, the two discriminant scores can be rather different. This suggests that for real RNA-Seq data with moderate or high dispersion, NBLDA may be a more appropriate method than PLDA. Note that the true dispersions are unlikely to be known in practice. Therefore, we proposed to first estimate the average dispersion using some novel estimation methods in the recent literature. Second, if the estimated average dispersion is small, we use PLDA; and otherwise we use NBLDA.

In the future, some work needs to be further studied. For instance, in Chapter 2, some structure assumptions, such as sparsity and conditional sparsity, can be added to the covariance matrices, and the consistency of the estimators is worthy of investigating. The assumption of normal population in Chapter 3 is restrictive and the non-normal distributions can be considered. The simulations in Chapter 4 can be seen as a comparison for covariance matrix estimation and which loss function

provides to be the best candidate for evaluating the estimation methods for the covariance matrix can be investigated. In Chapter 5, we can consider to incorporate the network information of pathways or gene sets to further improve the performance of classification.

Bibliography

Abadir, K. M. and Magnus, J. R. (2005). *Matrix Algebra*, Cambridge University Press.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data, *Genome Biology* **11**: R106.

Bai, Z. D. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem, *Statistica Sinica* **6**: 311–329.

Baldi, P. and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes, *Bioinformatics* **17**: 509–519.

Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, *Journal of Machine Learning Research* **9**: 485–516.

Bchetnia, M., Tremblay, M.-L., Leclerc, G., Dupérée, A., Powell, J., McCuaig, C., Morin, C., Legendre-Guillemin, V. and Laprise, C. (2012). Expression signature of epidermolysis bullosa simplex, *Human Genetics* **131**: 393–406.

Bickel, P. J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, *Bernoulli* **10**: 989–1010.

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

Bullard, J. H., Purdom, E., Hansen, K. D. and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments, *BMC Bioinformatics* **11**: 94.

Cai, T., Liang, T. and Zhou, H. (2013). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions, *arXiv preprint arXiv:1309.0482* .

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation, *Journal of the American Statistical Association* **106**: 672–684.

Cai, T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding, *The Annals of Statistics* **40**: 2014–2042.

Cai, T. and Zhou, H. (2012). Optimal rates of convergence for sparse covariance matrix estimation, *The Annals of Statistics* **40**: 2389–2420.

Chen, L. S., Paul, D., Prentice, R. L. and Wang, P. (2011). A regularized Hotellings T^2 test for pathway analysis in proteomic studies, *Journal of the American Statistical Association* **106**: 1345–1360.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing, *The Annals of Statistics* **38**: 808–835.

Cui, X., Hwang, J. T. G., Qiu, J., Blades, N. J. and Churchill, G. A. (2005). Improved statistical tests for differential gene expression by shrinking variance components estimates, *Biostatistics* **6**: 59–75.

Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples, *Biometrics* **16**: 41–50.

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J. et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, *Briefings in Bioinformatics* **14**: 671–683.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A., Jhangri, G., Einecke, G., Famulski, K., Halloran, P., and Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS, *BMC Bioinformatics* **8**: 242.

Dong, S., Nutt, C. L., Betensky, R. A., Stemmer-Rachamimov, A. O., Denko, N. C., Ligon, K. L., Rowitch, D. H. and Louis, D. N. (2005). Histology-based expression profiling yields novel prognostic markers in human glioblastoma, *Journal of Neuropathology and Experimental Neurology* **64**: 948–955.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association* **97**: 77–87.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes, *The Annals of Applied Statistics* **1**: 107–129.

Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models, *The Annals of Statistics* **39**: 3320–3356.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion), *Journal of the Royal Statistical Society: Series B* **75**: 603–680.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*, Chapman and Hall.

Fisher, T. J. and Sun, X. (2011). Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix, *Computational Statistics & Data Analysis* **55**: 1909–1918.

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

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A. et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**: 531–537.

Guionnet, A. (2009). *Large Random Matrices: Lectures on Macroscopic Asymptotics*, Springer.

Guo, Y., Hastie, T. and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays, *Biostatistics* **8**: 86–100.

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

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

Hardcastle, T. J. and Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data, *BMC Bioinformatics* **11**: 422.

Hotelling, H. (1931). The generalization of Student's ratio, *Annals of Mathematical Statistics* **2**: 360–378.

Huang, S., Tong, T. and Zhao, H. (2010). Bias-corrected diagonal discriminant rules for high-dimensional classification, *Biometrics* **66**: 1096–1106.

Karoui, N. E. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices, *The Annals of Statistics* **36**: 2717–2756.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E. et al. (2010). Variation in transcription factor binding among humans, *Science* **328**: 232–235.

Kaur, S., Archer, K. J., Devi, M. G., Kriplani, A., Strauss, J. F. and Singh, R. (2012). Differential gene expression in granulosa cells from polycystic ovary syndrome patients with and without insulin resistance: identification of susceptibility gene sets through network analysis, *Journal of Clinical Endocrinology & Metabolism* **97**: E2016–E2021.

Kuster, D. W., Merkus, D., Kremer, A., van IJcken, W. F., de Beer, V. J., Verhoeven, A. J. and Duncker, D. J. (2011). Left ventricular remodeling in swine after myocardial infarction: a transcriptional genomics approach, *Basic Research in Cardiology* **106**: 1269–1281.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation, *The Annals of Statistics* **37**: 42–54.

Landau, W. M. and Liu, P. (2013). Dispersion estimation and its effect on test performance in RNA-Seq data analysis: A simulation-based comparison of methods, *PLOS ONE* **8**: e81415.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, *Journal of Empirical Finance* **10**: 603–621.

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

Lee, J. W., Lee, J. B., Park, M. and Song, S. H. (2005a). An extensive comparison of recent classification tools applied to microarray data, *Computational Statistics & Data Analysis* **48**: 869–885.

Lee, J. W., Lee, J. B., Park, M. and Song, S. H. (2005b). An extensive comparison of recent classification tools applied to microarray data, *Computational Statistics & Data Analysis* **48**: 869–885.

Letac, G. and Massam, H. (2004). All invariant moments of the Wishart distribution, *Scandinavian Journal of Statistics* **31**: 295–318.

Li, J. and Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data, *Statistical Methods in Medical Research* **22**: 519–536.

Lin, B., Zhang, L. and Chen, X. (2014). LFCseq: a nonparametric approach for differential expression analysis of RNA-seq data, *BMC Genomics* **15**(Suppl 10): S7.

Lorenz, D. J., Gill, R. S., Mitra, R. and Datta, S. (2014). Using RNA-seq data to detect differentially expressed genes, *Statistical Analysis of Next Generation Sequencing Data*, Springer, pp. 25–49.

Maciejewski, H. (2014). Gene set analysis methods: statistical models and methodological differences, *Briefings in Bioinformatics* **15**: 504–518.

Mardis, E. R. (2008). Next-generation DNA sequencing methods, *Annual Review of Genomics and Human Genetics* **9**: 387–402.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays, *Genome Research* **18**: 1509–1517.

Mokry, M., Hatzis, P., Schuijers, J., Lansu, N., Ruzius, F. P., Clevers, H. and Cuppen, E. (2012). Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes, *Nucleic Acids Research* **40**: 148–158.

Morozova, O., Hirst, M. and Marra, M. A. (2009). Applications of new sequencing technologies for transcriptome analysis, *Annual Review of Genomics and Human Genetics* **10**: 135–151.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing, *Science* **320**: 1344–1349.

Nakayama, R., Nemoto, T., Takahashi, H., Ohta, T., Kawai, A., Seki, K., Yoshida, T., Toyama, Y., Ichikawa, H. and Hasegawa, T. (2007). Gene expression analysis of soft tissue sarcomas: characterization and reclassification of malignant fibrous histiocytoma, *Modern Pathology* **20**: 749–759.

Newton, M. A., Quintana, F. A., Den Boon, J. A., Sengupta, S. and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis, *The Annals of Applied Statistics* **1**: 85–106.

Pang, H., Tong, T. and Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data, *Biometrics* **65**: 1021–1029.

Park, J. and Nag Ayyala, D. (2013). A test for the mean vector in large dimension and small samples, *Journal of Statistical Planning and Inference* **143**: 929–943.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S. and Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature* **415**: 436–442.

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence, *Electronic Journal of Statistics* **5**: 935–980.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics* **26**: 139–140.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion with applications to SAGE data, *Biostatistics* **9**: 321–332.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices, *Biometrika* **99**: 733–740.

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

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical Applications in Genetics and Molecular Biology* **4**: 32.

Searcy, J. L., Phelps, J. T., Pancani, T., Kadish, I., Popovic, J., Anderson, K. L., Beckett, T. L., Murphy, M. P., Chen, K. C., Blalock, E. M., Landfield, P. W.,

Porter, N. M. and Olivier, T. (2012). Long-term pioglitazone treatment improves learning and attenuates pathological markers in a mouse model of alzheimer's disease, *Journal of Alzheimer's Disease* **30**: 943–961.

Shen, Y., Lin, Z. and Zhu, J. (2011). Shrinkage-based regularization tests for high-dimensional data with application to gene set analysis, *Computational Statistics & Data Analysis* **55**: 2221–2233.

Si, Y. and Liu, P. (2013). An optimal test with maximum average power while controlling FDR with application to RNA-Seq data, *Biometrics* **69**: 594–605.

Srivastava, M. S. (2009). A test for the mean vector with fewer observations than the dimension under non-normality, *Journal of Multivariate Analysis* **100**: 518–532.

Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension, *Journal of Multivariate Analysis* **99**: 386–402.

Srivastava, M. S., Katayama, S. and Kano, Y. (2013). A two sample test in high dimensional data, *Journal of Multivariate Analysis* **114**: 349–358.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean, *Annals of the Institute of Statistical Mathematics* **16**: 155–160.

Sun, L., Hui, A.-M., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R. et al. (2006). Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain, *Cancer Cell* **9**: 287–300.

Tan, K. M., Petersen, A. and Witten, D. (2014). Classification of RNA-seq data, *Statistical Analysis of Next Generation Sequencing Data*, Springer, pp. 219–246.

Tao, T. (2012). *Topics in Random Matrix Theory*, American Mathematical Society.

Tong, T., Chen, L. and Zhao, H. (2012). Improved mean estimation and its application to diagonal discriminant analysis, *Bioinformatics* **28**: 531–537.

Tong, T., Feng, Z., Hilton, J. S. and Zhao, H. (2013). Estimating the proportion of true null hypotheses using the pattern of observed p-values, *Journal of Applied Statistics* **40**: 1949–1964.

Tong, T., Jang, H. and Wang, Y. (2012). James-Stein type estimators of variances, *Journal of Multivariate Analysis* **107**: 232–243.

Tong, T., Wang, C. and Wang, Y. (2014). Estimation of variances and covariances for high-dimensional data: a selective review, *WIREs Computational Statistics* **6**: 255–264.

Tong, T. and Wang, Y. (2007). Optimal shrinkage estimation of variances with applications to microarray data analysis, *Journal of the American Statistical Association* **102**: 113–122.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings National Academic Science* **98**: 5116–5121.

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics* **10**: 57–63.

Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices, *Journal of the American Statistical Association* **103**: 340–349.

Warton, D. I. (2011). Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations, *Biometrics* **67**: 116–123.

Witten, D. M. (2011). Classification and clustering of sequencing data using a Poisson model, *The Annals of Applied Statistics* **5**: 2493–2518.

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems, *Journal of the Royal Statistical Society: Series B* **71**: 615–636.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant, *Journal of the Royal Statistical Society: Series B* **73**: 753–772.

Witten, D., Tibshirani, R., Gu, S. G., Fire, A. and Lui, W. (2010). Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls, *BMC Biology* **8**: 58.

Wright, G. W. and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments, *Bioinformatics* **19**: 2448–2455.

Wu, H., Wang, C. and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-Seq data, *Biostatistics* **14**: 232–243.

Wu, Y., Genton, M. G. and Stefanski, L. A. (2006). A multivariate two-sample mean test for small sample size and missing data, *Biometrics* **62**: 877–885.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior, *The Annals of Statistics* **22**: 1195–1211.

Yin, J. and Li, H. (2013). Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization, *Journal of Multivariate Analysis* **116**: 365–381.

Yu, D., Huber, W. and Vitek, O. (2013). Shrinkage estimation of dispersion in Negative Binomial models for RNA-Seq experiments with small sample size, *Bioinformatics* **29**: 1275–1282.

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

Zhan, F., Barlogie, B., Arzoumanian, V., Huang, Y., Williams, D. R., Hollmig, K., Pineda-Roman, M., Tricot, G., van Rhee, F., Zangari, M. et al. (2007).

Gene-expression signature of benign monoclonal gammopathy evident in multiple myeloma is linked to good prognosis, *Blood* **109**: 1692–1700.

Zhang, J. and Xu, J. (2009). On the k-sample Behrens-Fisher problem for high-dimensional data, *Science in China Series A: Mathematics* **52**: 1285–1304.

Zhou, Y., Xia, K. and Wright, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data, *Bioinformatics* **27**: 2672–2678.

CURRICULUM VITAE

Academic qualifications of the thesis author, Mr. DONG Kai:

- Received the degree of Bachelor of Science (Statistics) from Minzu University of China, July 2009.
- Received the degree of Master of Science (Probability and Mathematical Statistics) from Beijing University of Technology, July 2012.

August 2015