High-dimensional discriminant analysis and covariance matrix estimation

by

Yilei Wu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2017

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Dr. Yingying Fan Associate Professor, University of Southern California
Supervisor(s):	Dr. Yingli Qin Assistant Professor
	Dr. Mu Zhu Professor
Internal Member:	Dr. Pengfei Li Associate Professor
	Dr. Grace Yi Professor
Internal-External Member:	Dr. Yuying Li Professor, School of Computer Science

Author's declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapter 2 – 4 are based on research papers that are co-authored with my supervisors. My contributions include proposing the main methods, deriving the main theorems, developing and implementing the algorithms, performing the simulations, performing the real-data analyses, and writing the manuscripts. My supervisors gave me advice on these works. In addition, Chapter 2 and Chapter 3 are based on research papers that have been edited by my supervisors for publication.

Abstract

Statistical analysis in high-dimensional settings, where the data dimension p is close to or larger than the sample size n, has been an intriguing area of research. Applications include gene expression data analysis, financial economics, text mining, and many others. Estimating large covariance matrices is an essential part of high-dimensional data analysis because of the ubiquity of covariance matrices in statistical procedures. The estimation is also a challenging part, since the sample covariance matrix is no longer an accurate estimator of the population covariance matrix in high dimensions. In this thesis, a series of matrix structures, that facilitate the covariance matrix estimation, are studied.

Firstly, we develop a set of innovative quadratic discriminant rules by applying the compound symmetry structure. For each class, we construct an estimator, by pooling the diagonal elements as well as the off-diagonal elements of the sample covariance matrix, and substitute the estimator for the covariance matrix in the normal quadratic discriminant rule. Furthermore, we develop a more general rule to deal with nonnormal data by incorporating an additional data transformation. Theoretically, as long as the population covariance matrices loosely conform to the compound symmetry structure, our specialized quadratic discriminant rules enjoy low asymptotic classification error. Computationally, they are easy to implement and do not require large-scale mathematical programming.

Then, we generalize the compound symmetry structure by considering the assumption that the population covariance matrix (or equivalently its inverse, the precision matrix) can be decomposed into a diagonal component and a low-rank component. The rank of the low-rank component governs to what extent the decomposition can simplify the covariance/precision matrix and reduce the number of unknown parameters. In the estimation, this rank can either be pre-selected to be small or controlled by a penalty function. Under moderate conditions on the population covariance/precision matrix itself and on the penalty function, we prove some consistency results for our estimator. A blockwise coordinate descent algorithm, which iteratively updates the diagonal component and the low-rank component, is then proposed to obtain the estimator in practice.

In the end, we consider jointly estimating large covariance matrices of multiple categories. In addition to the aforementioned diagonal and low-rank matrix decomposition,

it is further assumed that there is some common matrix structure shared across the categories. We assume that the population precision matrix of category k can be decomposed into a diagonal matrix D, a shared low-rank matrix L, and a category-specific low-rank matrix $L^{(k)}$. The assumption can be understood under the framework of factor models — some latent factors affect all categories alike while others are specific to only one of these categories. We propose a method that jointly estimates the precision matrices (therefore, the covariance matrices) — D and L are estimated with the entire dataset whereas $L^{(k)}$ is estimated solely with the data of category k. An AIC-type penalty is applied to encourage the decomposition, especially the shared component. Under certain conditions on the population covariance matrices, some consistency results are developed for the estimators.

The performances in finite dimensions are shown through numerical experiments. Using simulated data, we demonstrate certain advantages of our methods over existing ones, in terms of classification error for the discriminant rules and Kullback–Leibler loss for the covariance matrix estimators. The proposed methods are also applied to real life datasets, including microarray data, stock return data and text data, to perform tasks, such as distinguishing normal from diseased tissues, portfolio selection and classifying webpages.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisors, Dr. Yingli Qin and Dr. Mu Zhu, for their continuous support and patient guidance over the past few years. I appreciate the remarkable ideas, the constructive advice, and the precious time they contributed to our meetings and works, as well as the effort they took to guide me through every step of becoming a researcher.

My gratitude extends to the rest of my examining committee members, Dr. Yingying Fan, Dr. Pengfei Li, Dr. Grace Yi and Dr. Yuying Li for taking their precious time to read and make comments on my thesis. The knowledge, that they generously share, will always be my asset.

Furthermore, my special thanks go to Dr. Stephen M.S. LEE. I would like to thank him for introducing me to the fascinating world of statistics, as well as teaching me the very first class in research. Without his help, support and guidance, I would not have even started my Ph.D. pursuit.

Many thanks go to my friends, who have made my time at Waterloo enjoyable.

Last but not least, I would like to express my thanks to my parents and Fei, for their unconditional love and constant care. I am also grateful for their understanding and support throughout my Ph.D. study.

Yilei Wu
October 26th, 2017, in Waterloo

To my family

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Challenges due to high-dimensionality

High-dimensional data have emerged in a variety of areas and become ever more ubiquitous. For example, it is typical that microarray gene expression data contain expression levels of tens of thousands of genes but a much smaller number of subjects. These data can be used for various purposes, such as identifying genes that are differentially expressed across samples, identifying subtypes of a disease, distinguishing tumors from normal tissues; studying treatment effects on gene expression and many others. See Butte (2002) for a comprehensive review.

Data from finance and economics can also be high-dimensional. In portfolio selection and risk management, hundreds of assets are to be considered for allocation. Since estimating the large covariance matrix of asset returns is necessary for asset allocation, there could be more than a hundred thousand parameters to be estimated. In forecasting, the number of predictors could be approximately the same as the number of observations. For example, the dataset studied by Stock and Watson (2012) consists of 195 quarterly observations on 143 U.S. macroeconomic times series.

Other sources of high-dimensional data include text mining, functional magnetic resonance imaging, computer vision, climatology, etc.

There are many challenges in analyzing high-dimensional data. One is the noise accumulation when estimating the mean vector. Although the sample mean of every dimension is consistent by itself, the accumulated noise of all dimensions could be large. Dimension reduction methods such as feature selection and projection have been proposed as solutions. See Fan and Fan (2008) for more details.

Another challenge is that the sample covariance matrix is no longer an accurate estimator of the population covariance matrix, because the number of unknown parameters grows quadratically with the dimensionality. Moreover, the sample covariance matrix is non-invertible when the number of features exceeds the sample size, while the inverse of the covariance matrix is crucial in many classical statistical methods such as Hotelling's T^2 test and Fisher's linear discriminant analysis.

In this thesis, we focus on the covariance matrix. We use the following two examples to illustrate the impact of inaccurate covariance matrix estimations and motivate new estimation methods.

1.1.1 Markowitz portfolio selection

In the classic Markowitz portfolio selection problem (Markowitz, 1952), we have the opportunity to invest in p assets and aim to decide the asset weights, so that a certain level of expected return is achieved and the "risk", which is described by the variance, is minimized. El Karoui (2010) showed that estimating the covariance matrix of the asset returns with the sample covariance matrix leads to risk underestimation.

Let $\boldsymbol{\mu}$ be the mean returns and Σ be the covariance matrix of returns of p assets. The vector $\boldsymbol{\mu}$ is p-dimensional, and Σ is a $p \times p$ matrix. Let $\mathbf{1}_p$ be the p-dimensional vector with all elements being one. The Markowitz problem is formulated as

$$\mathbf{w}_{\text{optimal}} = \underset{\mathbf{w} \in \mathbb{R}^p}{\arg\min}\ \mathbf{w}'\Sigma\mathbf{w} \quad \text{subject to} \quad \mathbf{w}'\boldsymbol{\mu} = \mu_0, \mathbf{w}'\mathbf{1}_p = 1, \tag{1.1}$$

in which μ_0 is the desired level of expected return.

In practice, $\boldsymbol{\mu}$ and Σ are unknown and need to be estimated. Let $\widehat{\boldsymbol{\mu}}$ and $\widehat{\Sigma}$ be the

sample mean and sample covariance matrix, respectively. Problem (1.1) becomes

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min}\ \mathbf{w}'\widehat{\Sigma}\mathbf{w} \quad \text{subject to} \quad \mathbf{w}'\widehat{\boldsymbol{\mu}} = \mu_0, \mathbf{w}'\mathbf{1}_p = 1. \tag{1.2}$$

Since we focus on the performance of $\widehat{\Sigma}$ as an estimator of Σ, we compare $\widehat{\mathbf{w}}$ with

$$\widetilde{\mathbf{w}} = \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min}\ \mathbf{w}'\Sigma\mathbf{w} \quad \text{subject to} \quad \mathbf{w}'\widehat{\boldsymbol{\mu}} = \mu_0, \mathbf{w}'\mathbf{1}_p = 1, \tag{1.3}$$

in which Σ is provided while $\boldsymbol{\mu}$ is estimated by the sample mean.

A result from random matrix theory stated that the largest eigenvalue of Σ will be overestimated by the largest eigenvalue of $\widehat{\Sigma}$; the smallest eigenvalue of Σ will be underestimated by the smallest eigenvalue of $\widehat{\Sigma}$; and the situation is worsen in the "large p" setting (Marčenko and Pastur, 1967). The impact of this phenomenon on the portfolio selection problem can be illustrated with a simplified case, where the population covariance matrix is the identity matrix. If Σ is the $p \times p$ identity matrix I_p, the risk in (1.3), $\widetilde{\mathbf{w}}'\Sigma\widetilde{\mathbf{w}}$, is always 1 regardless of the choice of the weights. However, as some eigenvectors of $\widehat{\Sigma}$ are associated with the underestimated eigenvalues that are smaller than 1, intuitively, (1.2) will tend to give a solution that is closer to these eigenvectors and underestimate the overall risk with $\widehat{\mathbf{w}}'\widehat{\Sigma}\widehat{\mathbf{w}} < 1$, especially when p is large.

Theorem 3.1 in El Karoui (2010) depicted the risk underestimation more rigorously. Under the assumption of normal distribution (not necessarily with identity covariance matrix) and $(p-2) < (n-1)$, they gave

$$\widehat{\mathbf{w}}'\widehat{\Sigma}\widehat{\mathbf{w}} = \widetilde{\mathbf{w}}'\Sigma\widetilde{\mathbf{w}}\ \frac{\chi^2_{n-p+1}}{n-1},$$

where $\widetilde{\mathbf{w}}'\Sigma\widetilde{\mathbf{w}}$ is statistically independent of χ^2_{n-p+1}.

When p has the same order of magnitude as n and they are both large, $\chi^2_{n-p+1}/(n-1)$ is approximately $1-(p-2)/(n-1)$. Therefore, "large p" results in risk underestimation if the sample covariance matrix is used in the Markowitz portfolio selection problem.

See El Karoui (2010) for a complete study on this problem.

1.1.2 High-dimensional classification

With gene expression data, discriminant analysis can be used to distinguish between tumors and normal tissues or classify malignancies into different classes. Unlike some financial data, whose dimension p is relatively close to the sample size n, gene expression data generally contain expression levels of tens of thousands of genes for at most dozens or hundreds of samples. Thus, the high-dimensional problem is more severe in this case.

In the following, we show that using the sample covariance matrix in high-dimensional linear discriminant analysis (LDA) could lead to large misclassification probability.

Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be the two classes of p-dimensional normal distributions $N(\boldsymbol{\mu}_1, \Sigma)$ and $N(\boldsymbol{\mu}_2, \Sigma)$, respectively. Discriminant analysis is the problem of assigning a newly observed vector $\mathbf{x}$ to one of these classes. We assume equal unconditional prior probabilities, and the optimal linear classifier is the Bayes rule (Anderson, 2003), which classifies $\mathbf{x}$ to class 1 if

$$Q = (\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0, \tag{1.4}$$

and to class 2 otherwise, in which $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$.

Assume that Σ has eigenvalues bounded away from 0 and $+\infty$, and let

$$c = \left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^{1/2}$$

be the Mahalanobis distance between the two classes. If $\mathbf{x} \sim N(\boldsymbol{\mu}_1, \Sigma)$, the probability of misclassification is

$$\mathbb{P}\left(Q < 0 | \mathbf{x} \in \mathcal{C}_1\right) = \Phi\left(-\frac{c}{2}\right), \tag{1.5}$$

in which $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The probability of misclassification of $\mathbf{x} \sim N(\boldsymbol{\mu}_2, \Sigma)$ is the same because of the symmetry; thus, we only focus on the probability of misclassifying $\mathbf{x} \sim N(\boldsymbol{\mu}_1, \Sigma)$.

In practice, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ have to be estimated from the training data $\mathbf{Y}$. Without loss of generality, we assume $\mathbf{Y}$ contains $n/2$ observations from both classes. After replacing

the parameters in (1.4) with their sample counterparts (sample mean and pooled sample covariance matrix), we have

$$\widehat{Q} = (\mathbf{x} - \widehat{\boldsymbol{\mu}})' \widehat{\Sigma}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2).$$

In the case of fixed p and $n \to +\infty$, according to Theorem 6.5.1 in Anderson (2003), if $\mathbf{x} \sim N(\boldsymbol{\mu}_1, \Sigma)$, $\widehat{Q}$ converges to $N(c^2/2, c^2)$ in distribution and

$$\mathbb{P}\left(\widehat{Q} < 0 | \mathbf{x} \in \mathcal{C}_1\right) \to \Phi\left(-\frac{c}{2}\right), \tag{1.6}$$

which is the Bayes risk. Thus, the rule with sample mean and sample covariance matrix is asymptotically optimal in low dimensions.

However, the situation changes when $p/n \to \infty$. Bickel and Levina (2004) showed the problem. As $\widehat{\Sigma}$ is non-invertible when $p > n$, they replaced $\widehat{\Sigma}^{-1}$ with $\widehat{\Sigma}^{-}$, its Moore-Penrose inverse. They considered a simple case $\Sigma = I_p$, and proved, for $\mathbf{x} \sim N(\boldsymbol{\mu}_1, \Sigma)$,

$$\mathbb{P}\left(\widehat{Q} < 0 | \mathbf{x} \in \mathcal{C}_1\right) \to \frac{1}{2}, \tag{1.7}$$

as $p/n \to \infty$. The convergence indicates that, asymptotically, the LDA rule using $\widehat{Q}$ could be random guess in the high-dimensional setting.

1.2 Outline of the thesis

Chapter 2 – 4 are based on three independent research papers. The research paper of Chapter 2 has been accepted by *Statistica Sinica*, the research paper of Chapter 3 has been submitted for publication, and the manuscript of Chapter 4 is in preparation for submission. The model assumption in Chapter 4 is a multiclass extension of that in Chapter 3, while the latter is a generalization of that in Chapter 2. In each of these chapters, we firstly review the relevant literature, and then we propose an innovative method and study its properties in detail. The most relevant summaries are presented at the end of these chapters, but the overall conclusion of the thesis is relegated to Chapter 5. Technical proofs of different chapters are contained in separate sections of the appendices. The notations

are kept consistent within each chapter and its corresponding technical proofs; however, for simplicity of presentation, some notations might be redefined across the chapters.

In Chapter 2, we tackle the problem that the quadratic discriminant rule, in its original format, does not perform well in the high-dimensional setting. To this end, we propose a substitute for each covariance matrix in the rule. Each substitute has the compound symmetry structure — its diagonal (off-diagonal) elements are the average of the diagonal (off-diagonal) elements of the population covariance matrix. This substitution immediately reduces the number of unknown parameters from $[(p+1)p]/2$ to 2 and dramatically simplifies the estimation task — estimation by sample counterparts is accurate in spite of the high dimension. The structure might seem stringent at first sight; however, we establish that the population covariance matrices only need to loosely conform to the compound symmetry structure to ensure nice performance of the altered discriminant rule. Under conditions on the structures of population covariance matrices and the information for differentiating two classes (e.g., differences between class means or class covariance matrices), we prove a low asymptotic misclassification rate. A special case, that ignores the correlations, is also studied. The simulation, in which a variety of matrices are experimented, empirically shows advantages of our methods. In real data analysis, we discover that some simple pre-processing steps could further improve the performances.

The aforementioned discriminant rules are also generalized to handle nonnormal data; the method is to transform the data to be normally distributed before applying the discriminant rules. Assuming the existence of such transformations is equivalent to assuming that the dependence structure of the data can be described by a normal copula model. To estimate the transformation for each dimension, a function, which is based on the marginal empirical cumulative distribution function, can be used. We establish that, the generalized rules have low asymptotic misclassification rates, if the conditions on population covariance matrices and between-class differences are placed on the transformed data, and p is controlled by $\exp(n^c)$, where c is a positive constant. Simulations show that incorporating the transformation is advantageous dealing with nonnormal data, although it might lead to slightly higher misclassification rates when the original data already distribute normally, because of the extra error introduced by the transformation estimation.

In Chapter 3, we generalize the compound symmetry structure and directly consider

large covariance matrix estimation, instead of its application in classification. The compound symmetry structure can be decomposed into a scaled identity matrix and a rank-1 matrix; as a generalization, we now assume that a covariance matrix can be decomposed into a diagonal matrix and a low-rank matrix — $\Sigma = D + L$, in which L has rank $r < p$. This assumption can be interpreted by a factor model — D represents the variance of independent errors and L represents the variance explained by r latent factors. The decomposition assumption reduces the number of unknown parameters because L can be written as $L = RR'$, in which R is a $p \times r$ matrix. In this way, the estimation is simplified and its accuracy could be improved. To obtain such an estimator, we consider an optimization problem that minimizes the summation of (i) the distance between the estimator and the sample covariance matrix and (ii) a penalty imposed on r. Then we establish a consistency property for the estimator, under some conditions of the penalty function. On the one hand, the penalty has to be strong enough so that r and the number of unknown parameters are encouraged to be small. On the other hand, it should not be too strong, otherwise, r might be smaller than the true rank, and this could lead to bias. In the simulations, our estimator is shown to be accurate, under various setups of the population covariance matrix. In fact, many of these covariance matrices are not covered by the theory; they either do not exactly satisfy the decomposition assumption or violate other conditions.

In Chapter 4, we extend the idea of diagonal and low-rank decomposition to jointly estimate covariance matrices of multiple categories. We assume that the covariance matrix of category k can be decomposed into a diagonal matrix D, a low-rank matrix L and another low-rank matrix $L^{(k)}$, i.e., $\Sigma^{(k)} = D + L + L^{(k)}$. In the decomposition, D and L are shared across categories while $L^{(k)}$ is category-specific. If we interpret the assumption with factor models, D represents the variance of independent errors, L represents the variance explained by factors with common effects across categories, and $L^{(k)}$ represents the variance explained by factors with category-specific effects. Exploiting the common matrix structure leads to fewer overall unknown parameters (therefore better estimation accuracy) than considering the diagonal and low-rank decomposition for every category separately. To be more specific, knowing that $L + L^{(k)}$ is the low-rank component of the decomposition of category k, we can see that when $\text{rank}(L) + \text{rank}(L^{(k)})$ is fixed, the

larger the rank of L, the smaller the total number of unknown parameters of all categories. To obtain such estimators, we minimize an objective function, which consists of (i) the distances between the estimators and the corresponding sample covariance matrices of all categories and (ii) an AIC-type penalty. An overall consistency result of the estimators is established by extending the proof of Chapter 3. Through simulations and real data analysis, we show the advantages of the joint estimation over separate estimations for multiple categories.

1.3 Main contributions

To begin with, some matrix structures, such as the compound symmetry structure and the joint diagonal and low-rank decomposition of multiple matrices, are studied for the first time to facilitate large covariance matrix estimation. We not only propose these matrix structures, but also conduct thorough research on their underlying interpretation, implementation, and theoretical and numerical properties. Similar matrix structures to the diagonal and low-rank decomposition of a single matrix have been considered by other researchers; however, we employ a different method to encourage such a decomposition.

To encourage low-rank components in the (joint) diagonal and low-rank decomposition, we directly impose penalties on the ranks, e.g., the AIC-type penalties. To the best of our knowledge, we are the first to study this type of penalty in the context of high-dimensional covariance matrix estimation. Although other works have also considered low-rank matrices to facilitate high-dimensional covariance matrix estimation, they either did not employ penalties and chose the rank with other methods, or applied the nuclear norm to encourage a low rank component. Furthermore, due to the formulation of the penalty, we are able to establish corresponding consistency properties and efficient solvers.

Theoretical properties are established for these newly proposed methods. The conditions describe when these methods perform well, and the consistency results quantify the classification/estimation accuracy. (There could be situations, not covered by theories, where the methods still work well.) Some techniques and intermediate results of the proofs could also be considered contributions. For instance, we adapt the techniques

of establishing consistency for sparse estimators in Rothman et al. (2008) to handle our "low-rank" estimators. Moreover, some lemmas, that we prove as intermediate steps, are indeed standalone matrix properties.

Algorithms are developed for our methods. The algorithms for implementing the discriminant rules are highly efficient because of the simplicity of the compound symmetry. Although inverse matrices are part of the discriminant rules, the inverse of a compound symmetry structure is far less expensive than that of a general matrix. Algorithms of the (joint) diagonal and low-rank decomposition methods are not as simple and involve numerical optimization techniques. But they require low memory and are also considered efficient, given the difficult of handling low-rank matrices.

Our contributions will be elaborated in each of the following three chapters.

Chapter 2

High-dimensional Quadratic Discriminant Analysis

2.1 Introduction

In this chapter, we study discriminant analysis in high dimensions. Suppose a random vector $\mathbf{x} \in \mathbb{R}^p$, where p is very large, comes from either class 1 ($\mathcal{C}_1$) or class 2 ($\mathcal{C}_2$). On the training data, the class memberships of these vectors are labelled. The goal is to classify an unlabelled observation using a *discriminant rule* that is learned from the training data. To focus on the main issues, we shall assume that the unconditional prior probabilities of both classes are equal to 1/2. Otherwise, all discriminant rules mentioned can be modified simply by adding a constant to correct for the class imbalance.

For $i = 1, 2$, let $\boldsymbol{\mu}_i$ and Σ_i be the class mean and class covariance matrix, respectively. To determine the class label of $\mathbf{x}$, Fisher's *linear* discriminant rule (see, e.g., Anderson, 2003), which assumes $\Sigma_1 = \Sigma_2 = \Sigma$, classifies $\mathbf{x}$ to class 1 if

$$(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0, \tag{2.1}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1)/2$, and to class 2 otherwise. If the two covariance matrices cannot be taken to be identical, then the *quadratic* discriminant rule can be used, which classifies

$\mathbf{x}$ to class 1 if

$$\ln\left(|\Sigma_1|/|\Sigma_2|\right) + (\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \leq 0, \tag{2.2}$$

and to class 2 otherwise. Equation (2.2) is also the Bayes rule under the assumption that $\mathbf{x} \sim N(\boldsymbol{\mu}_i, \Sigma_i)$ if $\mathbf{x} \in \mathcal{C}_i$, and so is equation (2.1) when $\Sigma_1 = \Sigma_2$.

In practice, the parameters $\boldsymbol{\mu}_i$ and Σ_i are unknown and need to be estimated from training data. Let $\widehat{\boldsymbol{\mu}}_i$ and $\widehat{\Sigma}_i$ be the sample mean and sample covariance matrix of class i. They are conventionally used as estimators of $\boldsymbol{\mu}_i$ and Σ_i. The common covariance matrix in (2.1) is estimated by the pooled sample covariance matrix, $\widehat{\Sigma} = (n_1 + n_2 - 2)^{-1}[(n_1 - 1)\widehat{\Sigma}_1 + (n_2 - 1)\widehat{\Sigma}_2]$. When the dimension is high and the number of covariates p is close to or larger than the number of observations n, the sample covariance matrix is well-known to be a poor estimate of its population counterpart; in fact, it is often singular and cannot be directly plugged into the discriminant rules.

2.1.1 Linear discriminant analysis (LDA)

In recent years, many methods have been proposed in the literature for performing linear discriminant analysis (LDA) in high dimensions. For example, one can ignore the covariance terms and use just a diagonal matrix in (2.1) — these are referred to as "independence rules". Bickel and Levina (2004) showed that, if one simply uses the Moore-Penrose inverse of $\widehat{\Sigma}$, then the misclassification error of (2.1) converges to 1/2 as $p/n \to \infty$, whereas the independence rule is at least as good. These "independence rules" can, and often should, be applied in conjunction with feature selection. For instance, Fan and Fan (2008) pointed out that they can perform poorly by themselves due to noise accumulation in estimating the population centroids, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, in high-dimensional spaces. They proposed to select a subset of important features by performing two-sample t-tests before applying the independence rule. Based on similar considerations, Tibshirani et al. (2002) shrunk class centroids towards the overall center of the data in order to reduce noise, and also estimated Σ with a diagonal matrix.

Another popular approach in the literature is to impose sparsity assumptions. For example, Shao et al. (2011) assumed both Σ and the mean difference vector, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, to be

sparse, and estimated them by thresholding. Fan et al. (2013b) performed variable selection by "innovated thresholding" and "higher criticism thresholding" before carrying out LDA with the selected set of features. Hao et al. (2015) rotated the data to create sparsity prior to applying existent classifiers. Witten and Tibshirani (2011) applied a sparsity penalty in seeking out a projection direction that maximized the between-class variance. Notice that, for LDA, the (pooled) covariance matrix Σ affects classification only through the discriminant direction, $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Thus, various methods have been proposed to avoid the estimation of Σ itself — e.g., Fan et al. (2012) solved for the discriminant direction directly by minimizing the misclassification rate under a sparsity constraint; Mai et al. (2012) found the direction by solving a penalized linear regression problem; see also Cai and Liu (2011b).

2.1.2 Quadratic discriminant analysis (QDA)

The LDA rule (2.1) assumes that two classes share the same covariance matrix, which is challenging to test in high dimensions (see, e.g., Li and Chen, 2012; Cai et al., 2013a, and many others). If the null hypothesis, $H_0 : \Sigma_1 = \Sigma_2$, cannot be accepted for sound reasons, it may become necessary to consider quadratic discriminant analysis (QDA). However, because there are many more unknown parameters to estimate, QDA is much more challenging than LDA, especially in high dimensions, and much less work has been done about it.

As in the case of LDA, it is also natural to use just diagonal covariance matrices or to impose some sparsity conditions in order to regularize QDA. For example, diagonal quadratic discriminant analysis (DQDA) was studied by Dudoit et al. (2002), whereas Li and Shao (2015) suggested a sparse QDA (SQDA) procedure by thresholding not only the mean difference vector $\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2$, but also the covariance matrices $\widehat{\Sigma}_i$ and their difference $\widehat{\Sigma}_1 - \widehat{\Sigma}_2$. A more recent work on sparse QDA rule is based on the dimension reduction method, QUADRO, proposed by Fan et al. (2015). QUADRO constructs a quadratic projection $f(\mathbf{x}) = \mathbf{x}'\Omega\mathbf{x} - 2\boldsymbol{\delta}'\mathbf{x}$ by maximizing the Rayleigh quotient of f, which is the ratio of the variance explained by the class label to the remaining variance. The parameters, Ω and $\boldsymbol{\delta}$, are encouraged to be sparse by ℓ_1 penalties. The estimated projection can then

be used for classification. For example, the class label can be decided by the sign of $\mathbf{x}'\widehat{\Omega}\mathbf{x} - 2\widehat{\boldsymbol{\delta}}'\mathbf{x} - c$ for some thresholding constant c.

Friedman (1989) proposed regularized discriminant analysis (RDA) as a way to compromise between LDA and QDA. In particular, his proposal shrinks the sample class covariance matrix $\widehat{\Sigma}_i$ twice — once toward the pooled sample covariance matrix, $\widehat{\Sigma}$, and once again toward the diagonal matrix, $p^{-1}tr(\widehat{\Sigma}_i)I_p$, where $tr(\cdot)$ denotes the trace of a matrix and I_p is $p \times p$ identity matrix.

We shall refer to the quantity, $p^{-1}tr(\widehat{\Sigma}_i)I_p$, simply as the "trace estimator". It has been used in the literature for high-dimensional hypothesis testing and classification problems, and is closely related to our methods. One reason why the trace estimator is useful is that, under some mild conditions, $p^{-1}tr(\widehat{\Sigma}_i)$ can be shown to be a consistent estimator of $p^{-1}tr(\Sigma_i)$ even as $p \to \infty$.

For classification, Friedman's RDA (Friedman, 1989) clearly uses the trace estimator, as it shrinks the sample covariance matrix $\widehat{\Sigma}_i$ towards both the pooled covariance estimator $\widehat{\Sigma}$ and the trace estimator. Shrinking toward the trace estimator is one way to overcome the well-known bias in the sample covariance matrix, which inflates large eigenvalues and deflates smaller ones. The two directions of shrinkage are controlled by two separate tuning parameters, λ and γ, as follows:

$$\widehat{\Sigma}_i(\lambda) = \frac{(1-\lambda)(n_i - 1)\widehat{\Sigma}_i + \lambda(n_1 + n_2 - 2)\widehat{\Sigma}}{(1-\lambda)(n_i - 1) + \lambda(n_1 + n_2 - 2)},$$

and

$$\widehat{\Sigma}_i(\lambda, \gamma) = (1-\gamma)\widehat{\Sigma}_i(\lambda) + \gamma \left[p^{-1}tr(\widehat{\Sigma}_i(\lambda))I_p \right].$$

There are four extreme cases. When $\lambda = 0$ and $\gamma = 0$, RDA reduces to vanilla QDA. When $\lambda = 1$ and $\gamma = 0$, RDA amounts to LDA. When $\lambda = 1$ and $\gamma = 1$, RDA is equivalent to replacing $\widehat{\Sigma}$ in LDA with just the identity matrix — in this case, classification is based on comparing Euclidean distances $||\mathbf{x} - \widehat{\boldsymbol{\mu}}_i||^2$ instead of Mahalanobis distances $(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)'\widehat{\Sigma}^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_i)$, for $i = 1, 2$. When $\lambda = 0$ and $\gamma = 1$, RDA is equivalent to replacing $\widehat{\Sigma}_i$ in the QDA rule (2.2) with the trace estimator, $p^{-1}tr(\widehat{\Sigma}_i)I_p$.

For hypothesis testing, Bai and Saranadasa (1996) proposed a test statistic, which replaces the pooled sample covariance matrix $\widehat{\Sigma}$ in Hotelling's two-sample T^2-statistic with

the identity matrix I_p and uses just the squared Euclidean distance (rather than Mahalanobis distance) between the sample means for high-dimensional problems. However, to do so, a bias-correction term must be added that depends on $tr(\widehat{\Sigma})$. Chen and Qin (2010) generalized this to the case where $\Sigma_1 \neq \Sigma_2$ so using the pooled estimate $\widehat{\Sigma}$ is no longer appropriate.

Aoshima and Yata (2014) then followed up on these ideas and used them for classification. In particular, they substituted the identity matrix I_p for the sample covariance matrix $\widehat{\Sigma}$ in the LDA rule (2.1), and used the trace estimator in place of each $\widehat{\Sigma}_i$ in the QDA rule (2.2). These two rules are similar to two of the four extreme cases in Friedman's RDA, corresponding to $(\lambda, \gamma) = (1, 1)$ and $(\lambda, \gamma) = (0, 1)$, except for the aforementioned bias-correction terms involving $tr(\widehat{\Sigma})$ and $tr(\widehat{\Sigma}_i)$. They also investigated a few variants of their quadratic rule (Aoshima and Yata, 2015).

2.1.3 Handling nonnormal data

Compared with LDA, QDA is more sensitive to deviations from normality (Friedman, 1989). A common approach for relaxing the normality assumption is to assume that there exists a strictly monotone transformation for each dimension such that the transformed vector $\mathbf{x}$ follows a multivariate normal distribution given its class label (e.g., Lin and Jeon, 2003; Liu et al., 2009; Mai and Zou, 2015). After first estimating and then applying these transformations, Lin and Jeon (2003) performed classic LDA and QDA; Liu et al. (2009) estimated undirected graphical models; and Mai and Zou (2015) applied their direct method for sparse discriminant analysis (DSDA). In this work, we will also rely on this idea to generalize our methods.

2.1.4 Outline and summary of this chapter

One can view the trace estimator as the result of two operations: pooling the diagonal elements of each sample covariance matrix, and ignoring its off-diagonal elements. In this chapter, we take the idea of the trace estimator one step further, and introduce an estimator that also pools the off-diagonal elements. We will refer to the resulting QDA

rule as ppQDA (for having performed two pooling operations), and the QDA rule with the trace estimator as pQDA — a special case of our more general method. We will study their asymptotic performances (Section 2.2), and also generalize them to handle nonnormal data (Section 2.3). Our generalization is based on first estimating a set of nonparametric data transformations and then applying our methods to the transformed data. As such, we will refer to these generalized QDA rules as semiparametric ppQDA (Se-pQDA) and semiparametric pQDA (Se-pQDA), respectively. We will prove a result for Se-pQDA, but only demonstrate the performance of Se-ppQDA empirically; the proof of a similar result for Se-ppQDA is more complicated, and will be left for future research.

Here is a summary of our main contributions. First, while most existing high-dimensional discriminant analysis methods focus on LDA, we fill this gap by focusing on QDA. Second, the sample covariance matrix is inconsistent when the dimension is high but, instead of making sparsity assumptions, we reduce the number of unknown parameters by simplifying the matrix structure in a different way. Third, using more than just the trace estimator in the QDA rule, our proposed ppQDA rule allows us to make use of information about the correlations among different dimensions. Fourth, we relax the normality assumption for both ppQDA and pQDA, and establish theoretical results for all of them except Se-ppQDA, the semiparametric extension of ppQDA. Finally, because our methods are based on using a very simple matrix structure, all our methods are computationally feasible and easy to apply in practice.

We proceed as follows. In Section 2.2, we introduce our notation, and describe our main methods, ppQDA and pQDA. In Section 2.3, we propose semiparametric generalizations of our main methods for nonnormal data. In Section 2.4, we give an outline of the main proofs, but detailed proofs are relegated to the appendices. Section 2.5 contains extensive numerical experiments, and Section 2.6 shows a few real-data examples. Then, in Section 2.7, we provide some important discussions about the relative performance of our ppQDA rule to that of the Bayes decision rule, before we close with some concluding remarks in Section 2.8.

2.2 QDA by pooling elements of covariance matrices

Let $\{\mathbf{y}_{1k} : 1 \le k \le n_1\}$ and $\{\mathbf{y}_{2k} : 1 \le k \le n_2\}$ be training samples from p-dimensional normal distributions $N(\boldsymbol{\mu}_1, \Sigma_1)$ and $N(\boldsymbol{\mu}_2, \Sigma_2)$, respectively. That is, $\mathbf{y}_{1k} \in \mathcal{C}_1$ and $\mathbf{y}_{2k} \in \mathcal{C}_2$. In addition, all $\mathbf{y}_{ik}$s are assumed to be independent. Let y_{ijk} to denote the jth dimension of $\mathbf{y}_{ik}$, for $j = 1, \ldots, p$. In what follows, $\mathbf{x} \in \mathbb{R}^p$ is used to denote a generic feature vector observation *without* a class label, and our target is to classify $\mathbf{x}$ based on a rule learned from the training samples. The sample version of the QDA rule (2.2) is to classify $\mathbf{x}$ to class 1 if

$$\ln(|\widehat{\Sigma}_1|/|\widehat{\Sigma}_2|) + (\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)'\widehat{\Sigma}_1^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1) - (\mathbf{x} - \widehat{\boldsymbol{\mu}}_2)'\widehat{\Sigma}_2^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_2) \le 0, \tag{2.3}$$

and to class 2 otherwise, but this does not work when p is larger than or close to n. We propose to replace the sample covariance matrices in (2.3) with simpler alternatives. Our main idea is to simplify the matrix structure in order to reduce the number of unknown parameters. When there are fewer parameters, we can expect to estimate them consistently.

2.2.1 Some basic conditions

Before introducing the special matrix structure that we propose to use, we first describe some common conditions on the covariance matrices and sample sizes.

Let $\Sigma_{j_1 j_2}$ be the element of Σ in the j_1th row and j_2th column.

Condition 2.1. *With respect to p, $|\Sigma_{j_1 j_2}|$ is uniformly bounded by a constant c .*

Condition 2.1 places a bound on all the elements of Σ. Throughout the chapter, we shall assume that both Σ_1 and Σ_2 satisfy condition 2.1.

Let $\mathbf{1}_p = (1, 1, ..., 1)' \in \mathbb{R}^p$ and $Su(\Sigma) = \mathbf{1}_p'\Sigma\mathbf{1}_p$ be the summation of all elements in Σ. Condition 2.1 implies Condition 2.2 below.

Condition 2.2. *For both $i = 1, 2$, $tr(\Sigma_i) = O(p)$, $tr(\Sigma_i^2) = O(p^2)$ and $Su(\Sigma_i) = O(p^2)$.*

We also assume that the sample sizes n_i for the two classes are close.

Condition 2.3. *There exist* $n > 0$ *and constants* $0 < c_1 < c_2 < +\infty$ *such that* $c_1 < n_i/n < c_2$ *as* $n \to \infty$ *for both* $i = 1, 2$.

Condition 2.3 is equivalent to saying that $n_1 \asymp n_2$. The value n has the same order as n_1 and n_2; it will be used later in our theoretical statements, where we will often refer to the sample size in general, without specifying the classes.

2.2.2 Main method: ppQDA

We now describe our main idea. Given Σ_i, let

$$a_i = p^{-1}tr(\Sigma_i) \quad \text{and} \quad r_i = (p(p-1))^{-1}\left(Su(\Sigma_i) - tr(\Sigma_i)\right),$$

be the average of its diagonal elements and the average of its off-diagonal elements, respectively. Our main idea is to use the structured matrix,

$$A_i = \begin{pmatrix} a_i & r_i & \cdots & r_i \\ r_i & a_i & \cdots & r_i \\ \vdots & \vdots & \ddots & \vdots \\ r_i & r_i & \cdots & a_i \end{pmatrix} = (a_i - r_i)I_p + r_i \mathbf{1}_p \mathbf{1}_p',$$

which has uniform diagonal elements and uniform off-diagonal elements, in place of Σ_i, for $i = 1, 2$, in the quadratic discriminant rule (2.2).

Estimators of a_i and r_i, and hence of A_i as well, are based on the sample covariance matrix, i.e.,

$$\widehat{a}_i = p^{-1}tr(\widehat{\Sigma}_i), \quad \widehat{r}_i = (p(p-1))^{-1}\left(Su(\widehat{\Sigma}_i) - tr(\widehat{\Sigma}_i)\right),$$

and

$$\widehat{A}_i = (\widehat{a}_i - \widehat{r}_i)I_p + \widehat{r}_i \mathbf{1}_p \mathbf{1}_p'.$$

As both a_i and r_i are scalar parameters, their estimators $\widehat{a}_i$ and $\widehat{r}_i$ are consistent even when p is large.

Using $\widehat{A}_i$ to replace $\widehat{\Sigma}_i$, for $i = 1, 2$, in (2.3), we call the resulting decision rule the "ppQDA rule", where each "p" is short for "pooling" as constructing $\widehat{A}_i$ involves pooling both the diagonal and the off-diagonal elements of $\widehat{\Sigma}_i$. Specifically, the ppQDA rule classifies $\mathbf{x}$ to class 1 if

$$\widehat{Q} = \ln\left(|\widehat{A}_1|/|\widehat{A}_2|\right) + (\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)'\widehat{A}_1^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1) - (\mathbf{x} - \widehat{\boldsymbol{\mu}}_2)'\widehat{A}_2^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_2) \leq 0, \tag{2.4}$$

and to class 2 otherwise. Due to its special structure, the inverse of A_i can be directly calculated:

$$\widehat{A}_i^{-1} = (\widehat{a}_i - \widehat{r}_i)^{-1} I_p - \widehat{r}_i(\widehat{a}_i - \widehat{r}_i)^{-1}(\widehat{a}_i + (p-1)\widehat{r}_i)^{-1}\mathbf{1}_p\mathbf{1}_p'. \tag{2.5}$$

Hence, we see that no matrix inversion is required, which is also a highly desirable property, especially for large p.

Theoretically, we are able to establish that our simplified ppQDA rule has good classification performance under Condition 2.1, Condition 2.3 and some additional conditions on the matrices, A_i for $i = 1, 2$, given below:

Condition 2.4. *The population covariance matrices satisfy*

1. $a_i - r_i > \delta_i > 0$, $p[a_i + (p-1)r_i] > \delta_i' > 0$;
2. $|(a_1 - r_1) - (a_2 - r_2)| > \delta_0 > 0$;
3. $tr\left((A_i - \Sigma_i)^2\right) = o(p^2)$;
4. $\sum_{j=1}^{p}(v_{ij} - \bar{v}_i)^2 = o(p^2)$, *where* $(v_{i1}, v_{i2}, \ldots, v_{ip}) = \mathbf{1}_p'\Sigma_i$ — *i.e.,* v_{ij} *is* j*th column-sum of* Σ_i — *and* $\bar{v}_i = p^{-1}\sum_{j=1}^{p} v_{ij}$.

Theorem 2.1. *Let* $\widehat{R}_{n,p} = \mathbb{P}(\widehat{Q} > 0|\mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(\widehat{Q} \leq 0|\mathbf{x} \in \mathcal{C}_2)$ *be the misclassification probability of the ppQDA rule (2.4). If conditions 2.1, 2.3 and 2.4 hold, then,*

$$\lim_{p\to\infty, n\to\infty} \widehat{R}_{n,p} = 0.$$

Notice that, in Theorem 2.1, we do not need to restrict the rate with which p approaches infinity relative to how fast the sample size n increases, a common requirement

for high-dimensional problems. This is because the ppQDA rule, in effect, reduces each covariance matrix to just two scalar parameters, a_i and r_i, which can be consistently estimated regardless of how big the dimension p is. However, we will require a restriction of the aforementioned kind later in Section 2.3 as we extend our basic ideas to a semiparametric setting (see Remark 2.6 below).

While Theorem 2.1 establishes conditions under which the ppQDA rule can be nearly perfect asymptotically, we will also discuss in more detail below (Section 2.7) the factors that control how close the ppQDA rule can approach the Bayes decision rule when nearly perfect classification is not achievable.

The detailed proof of Theorem 2.1 is given in the Appendix, although we will give a brief outline of the proof in Section 2.4. Here in this section, we first make some important remarks about Conditions 2.4.

Remark 2.1. *As long as Σ_i is a positive definite matrix, the inequalities, $a_i - r_i > 0$ and $a_i + (p-1)r_i > 0$, in Condition 2.4-1 always hold, by the definition of a_i and r_i (see Lemma A.1, Appendix). In addition, the Condition 2.4-1 requires that both $a_i - r_i$ and $p[a_i + (p-1)r_i]$ be bounded away from 0, a degeneracy, even as the dimension gets high.*

Remark 2.2. *Condition 2.4-2 essentially requires that there is some difference between the two class covariance matrices, Σ_1 and Σ_2, so that the two classes can be separated. Generally for multivariate normal distributions, there are two sources of information that make classification possible: differences between the mean vectors (locations), and differences between the covariance matrices. Condition 2.4-2 is sufficient but not necessary, since it only requires some difference between the covariance matrices. If there is adequate signal in the mean vectors, e.g., if $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is fairly large, then Condition 2.4-2 can be relaxed. This will be discussed in more detail in the Appendix, after the proof of Lemma A.2. We choose to use a condition that is solely focused on the covariance matrices for two reasons. First, there are already many papers in the literature (see Section 2.1) about discriminant analysis and classification based on signals from the mean vectors alone. Second, our main idea — that of replacing Σ_i with A_i — is about dealing with large covariance matrices (by introducing a structural simplification). As a result, Condition 2.4-2 actually makes classification possible even if there is no location separation at all ($\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = 0$).*

Remark 2.3. *Both conditions 2.4-3 and 2.4-4 place a bound on the difference between the true covariance matrix Σ_i and its structural simplification A_i. Naturally, if the true covariance matrix Σ_i really does have the simplified structure A_i, then our proposed ppQDA rule will be trivially optimal. What makes our proposal useful and interesting, of course, is that it can perform well even when the true covariance matrix does not have exactly the special structure. Conditions 2.4-3 and 2.4-4 make it precise how much Σ_i can deviate from the structure that would be "ideal" for our proposal. In particular, Condition 2.4-3 means that the average of squared elementwise difference between Σ_i and A_i is $o(1)$. Condition 2.4-4 is similar to 2.4-3 except it is about the column sums of Σ_i, $v_{i1}, ..., v_{ip}$, instead of about its individual elements. Notice that the average column sum, $\bar{v}_i$, can be expressed as $Su(\Sigma_i)/p = a_i + (p-1)r_i$, which is also equal to the uniform column sum of A_i for every column. Thus, Condition 2.4-4 also means that the average squared difference between the column sums of Σ_i and those of A_i is $o(p)$. Here, it is important to note that some commonly used covariance structures do, in fact, satisfy these two conditions, including the autoregressive matrix such as $\Sigma_i(j_1, j_2) = \sigma_i^2 \rho_i^{|j_1 - j_2|}$ and the block diagonal matrix — provided that the block size q is $o(p)$. Of course, if Σ_i deviates a lot from the structural simplification, then both of these conditions can be violated. For example, if half of the off-diagonal entries in Σ_i are zero and the other half are 0.2, then it easily can be derived that $tr((A_i - \Sigma_i)^2) \geq 0.01p(p-1)$, so $tr((A_i - \Sigma_i)^2) \neq o(p^2)$ and Condition 2.4-3 no longer holds.*

2.2.3 Special case: pQDA

We also consider a special case, which uses just the trace estimator, $\widehat{a}_i I_p$, to replace $\widehat{\Sigma}_i$, $i = 1, 2$. We call this rule "pQDA" because only the diagonal elements of $\widehat{\Sigma}_i$ are pooled and the off-diagonal elements are simply "ignored". This rule classifies $\mathbf{x}$ to class 1 if

$$\widehat{Q}_0 = p \ln (\widehat{a}_1 / \widehat{a}_2) + \widehat{a}_1^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)'(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1) - \widehat{a}_2^{-1} (\mathbf{x} - \widehat{\boldsymbol{\mu}}_2)'(\mathbf{x} - \widehat{\boldsymbol{\mu}}_2) \leq 0, \tag{2.6}$$

and to class 2 otherwise.

Clearly, the trace estimator, $\widehat{a}_i I_p$ is a special case of $\widehat{A}_i$. But we can take advantage of the added special structure and derive a stronger and more interpretable result under a

different set of assumptions:

Condition 2.5. *The population covariance matrices satisfy*

1. *there exist positive constants c_3 and c_4 such that $c_3 < \lambda_{ij} < c_4$ for $i = 1, 2$ and $j = 1, \ldots, p$, where λ_{ij} is the jth eigenvalue of Σ_i;*

2. *there exists some positive constant c_5 such that $(a_{i_1}/a_{i_2} - \ln(a_{i_1}/a_{i_2}) - 1) + p^{-1}a_{i_2}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > c_5$ for $(i_1, i_2) = (1, 2)$ and $(2, 1)$.*

Theorem 2.2. *Let $\widehat{R}_{0,n,p} = \mathbb{P}(\widehat{Q}_0 > 0 | \mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(\widehat{Q}_0 \leq 0 | \mathbf{x} \in \mathcal{C}_2)$ be the misclassification probability of the pQDA rule (2.6). If conditions 2.1, 2.3, and 2.5 hold, then,*

$$\lim_{p \to \infty, n \to \infty} \widehat{R}_{0,n,p} = 0.$$

The proof of Theorem 2.2 is, by and large, similar to that of Theorem 2.1 and the details will be omitted. Below, we make some important remarks about Condition 2.5.

Remark 2.4. *Condition 2.5-1 requires that the Σ_is have bounded eigenvalues in order for pQDA to work. The reason why ppQDA does not require bounded eigenvalues is that, although both A_i and $a_i I_p$ have a similar structure (uniform diagonal elements and uniform off-diagonal elements), A_i has a spiked eigenvalue spectrum (provided that r_i does not degenerate to 0, the case of pQDA), whereas $a_i I_p$ has uniform eigenvalues. Boundedness can also be thought of as a different way of stating closeness. In addition, as $a_i I_p$ has uniform eigenvalues, it is intuitive that our pQDA rule will perform better if the true covariance matrix Σ_i has eigenvalues that are closer to each other.*

Remark 2.5. *As we mentioned before (Remark 2.2), in quadratic discriminant analysis, there are two sources of information that are useful for class separation. One is the difference in the mean vectors, and the other is the difference in the covariance matrices. In our pQDA rule, these two sources of information are parameterized by $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and a_1/a_2 or a_2/a_1, respectively. Condition 2.5-2 simply requires that there is sufficient combined information for class separation from both sources. Note that the expression $a_{i_1}/a_{i_2} - \ln(a_{i_1}/a_{i_2}) - 1$ achieves its minimum value of 0 when $a_{i_1} = a_{i_2}$. Hence, classification becomes easier the larger the difference is between a_1 and a_2 — regardless of whether $a_1 > a_2$ or $a_2 > a_1$.*

2.3 Generalization to deal with nonnormal data

As we briefly mentioned in Section 2.1, QDA often is more sensitive to violations of the normality assumption than is LDA. In this section, we investigate a semiparametric method to relax the normality assumption for the pQDA rule. The ppQDA rule can be generalized similarly, but the theoretical justification is much more tedious, although it requires no additional technique (more on this below in Remark 2.8). Thus, we will state generalized versions of both the ppQDA rule and the pQDA rule, as well as include both of them in our empirical studies (Sections 2.5 and 2.6), but we will only develop the theory for generalized pQDA.

For non-normal data, we follow a common approach in the literature (e.g., Lin and Jeon, 2003; Liu et al., 2009; Mai and Zou, 2015) and assume that

Condition 2.6. *there exist a set of strictly monotonic transformations*

$$h(\mathbf{y}) \equiv (h_1(y_1), h_2(y_2), ..., h_p(y_p))'$$

such that $h(\boldsymbol{y}_{ik}) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$ *for* $k = 1, \ldots, n_i$ *and* $i = 1, 2$.

This assumption is equivalent to using a Gaussian copula model to describe the dependence structure of multivariate observation $\mathbf{y}_{ik}$ (Lin and Jeon, 2003).

To test the validity of Condition 2.6, any high-dimensional normality test can be applied to the transformed data. However, testing normality in high dimensions is another complex research problem in itself. According to Lin and Jeon (2003), an alternative may be to check the classification results directly, as it is possible for a classification rule to work reasonably well even if the underlying normality assumption is violated.

Under this assumption, the generalization of ppQDA and pQDA is straight-forward. First, we obtain a nonparametric estimate of the transformations, say

$$\widehat{h}(\cdot) \equiv (\widehat{h}_1(\cdot_1), \widehat{h}_2(\cdot_2), ..., \widehat{h}_p(\cdot_p))',$$

from the training sample. Then, we apply ppQDA and pQDA to the transformed data, $\{\widehat{h}(\mathbf{y}_{ik}) : k = 1, ..., n_i; i = 1, 2\}$ and $\widehat{h}(\mathbf{x})$. We refer to these procedures as Se-ppQDA and Se-pQDA, respectively, where "Se" is short for "semiparametric".

In what follows, we will use the same notations as before to denote various distributional parameters and their estimates for the *transformed* data. For example, $\boldsymbol{\mu}_i$ and Σ_i will now denote the mean vector and covariance matrix of the transformed sample $\{h(\mathbf{y}_{ik}) : k = 1, ..., n_i\}$, while

$$\widehat{\boldsymbol{\mu}}_i = n_i^{-1} \sum_{k=1}^{n_i} \widehat{h}(\mathbf{y}_{ik}) \quad \text{and} \quad \widehat{\Sigma}_i = (n_i - 1)^{-1} \sum_{k=1}^{n_i} (\widehat{h}(\mathbf{y}_{ik}) - \widehat{\boldsymbol{\mu}}_i)(\widehat{h}(\mathbf{y}_{ik}) - \widehat{\boldsymbol{\mu}}_i)'$$

will denote the corresponding sample quantities based on the estimated transformation, $\widehat{h}$. Similarly, a_i, r_i (likewise $\widehat{a}_i, \widehat{r}_i$) will continue to denote, respectively, the average of the diagonal and off-diagonal elements of Σ_i (likewise $\widehat{\Sigma}_i$) — except Σ_i and $\widehat{\Sigma}_i$ are now covariance and sample covariance matrices of the *transformed* data.

2.3.1 Estimation of h

Let F_{ij} be the class-i marginal cumulative distribution function (CDF) for the jth dimension. Let σ_{ij}^2 be the variance of $h_j(y_{ij})$, i.e., σ_{ij}^2 is the jth diagonal element of Σ_i. Notice that each of the assumed transformations $h_j(\cdot)$ in Condition 2.6 must satisfy the following: if $u \sim F_{1j}$ and $v \sim F_{2j}$, then after transformation the marginal distributions of $h_j(u)$ and $h_j(v)$ can differ only up to a location-and-scale transform. Thus, we can set $\mu_{1j} = 0$ and $\sigma_{1j}^2 = 1$ for all $j = 1, \ldots, p$, without loss of generality. This, in turn, means that each h_j can be equivalently expressed as

$$h_j = \Phi^{-1} \circ F_{1j} \qquad \text{or} \qquad h_j = \sigma_{2j} \left(\Phi^{-1} \circ F_{2j} \right) + \mu_{2j}, \tag{2.7}$$

where Φ denotes the CDF of the standard normal.

This means the transformation h_j can be estimated using training samples from either class. Here, we will estimate it using data from class 1, i.e.,

$$\widehat{h}_j = \Phi^{-1} \circ \widehat{F}_{1j},$$

where $\widehat{F}_{1j}$ is an "edge-smoothed" version of the empirical CDF (e.g., Mai and Zou, 2015),

$$\widehat{F}_{1j}(t) = \begin{cases} 1 - \frac{1}{n_1^2}, & \text{if} \quad \widetilde{F}_{1j}(t) > 1 - \frac{1}{n_1^2}; \\ \widetilde{F}_{1j}(t), & \text{if} \quad \frac{1}{n_1^2} \leq \widetilde{F}_{1j}(t) \leq 1 - \frac{1}{n_1^2}; \\ \frac{1}{n_1^2}, & \text{if} \quad \widetilde{F}_{1j}(t) < \frac{1}{n_1^2}, \end{cases}$$

and $\widetilde{F}_{1j}$ is the actual empirical CDF, $\widetilde{F}_{1j}(t) = n_1^{-1} \sum_{k=1}^{n_1} \mathbf{1}\{y_{1jk} \leq t\}$. But our choice of using data from class 1 is entirely arbitrary. In practice, we recommend using data from the larger class in order to maximize estimation accuracy.

It is also possible to estimate the transformation h_j by making use of data from both classes. For example, Mai and Zou (2015) proposed such a pooled estimator for the special case in which the class covariance matrices are identical. A closer look at (2.7) suggests that a potential generalization of their pooled, two-sample estimator could be to take a weighted average of two different, one-sample estimators of h_j, e.g.,

$$\widehat{h}_j = \frac{n_1}{n}(\Phi^{-1} \circ \widehat{F}_{1j}) + \frac{n_2}{n}\left[\widehat{\sigma}_{2j}(\Phi^{-1} \circ \widehat{F}_{2j}) + \widehat{\mu}_{2j}\right],$$

where $\widehat{F}_{2j}$ is defined similarly as $\widehat{F}_{1j}$ above. To take full advantage of pooled estimation, one could obtain $\widehat{\sigma}_{2j}$ and $\widehat{\mu}_{2j}$ with a pooled method as well, as there is information about them not only in the transformed sample $\{\Phi^{-1}[\widehat{F}_{1j}(y_{2jk})]\}_{k=1}^{n_2}$ but also in $\{\Phi^{-1}[\widehat{F}_{2j}(y_{1jk})]\}_{k=1}^{n_1}$. However, since this is not the main focus of our study, we will not pursue this more complicated, pooled strategy in the current work.

2.3.2 Se-ppQDA and Se-pQDA

Since our estimated transformations $\widehat{h}_1, ..., \widehat{h}_p$ automatically make $\widehat{\boldsymbol{\mu}}_1 = \mathbf{0}$, the Se-ppQDA rule classifies $\mathbf{x}$ to class 1 if

$$\widehat{Q}_{\widehat{h}} = \ln\left(|\widehat{A}_1|/|\widehat{A}_2|\right) + \widehat{h}(\mathbf{x})'\widehat{A}_1^{-1}\widehat{h}(\mathbf{x}) - (\widehat{h}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_2)'\widehat{A}_2^{-1}(\widehat{h}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_2) \leq 0, \tag{2.8}$$

and to class 2 otherwise. Similarly, that $\widehat{\sigma}_{1j}^2 = 1$ for all $j = 1, ..., p$ implies $\widehat{a}_1 = p^{-1}tr(\widehat{\Sigma}_1) = 1$, so the Se-pQDA rule classifies $\mathbf{x}$ to class 1 if

$$\widehat{Q}_{\widehat{h},0} = p\ln\left(1/\widehat{a}_2\right) + \widehat{h}(\mathbf{x})'\widehat{h}(\mathbf{x}) - \widehat{a}_2^{-1}(\widehat{h}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_2)'(\widehat{h}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_2) \leq 0, \tag{2.9}$$

and to class 2 otherwise.

We are now ready to establish some theoretical results about the asymptotic performance of the Se-pQDA rule. While the idea behind Se-pQDA — first estimating the transformations and then applying pQDA to transformed data — is straight-forward, its performance is much more intricate to analyze than that of pQDA, being affected by not only the structural simplifications of the pQDA rule itself, but also the estimation quality of the p univariate transformations and that of the key model parameters for the transformed data.

Theorem 2.3. *Let* $\widehat{R}_{\widehat{h},0,n,p} = \mathbb{P}(\widehat{Q}_{\widehat{h},0} > 0|\mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(\widehat{Q}_{\widehat{h},0} \leq 0|\mathbf{x} \in \mathcal{C}_2)$ *be the misclassification probability of the Se-pQDA rule (2.9). Under Condition 2.6, if and conditions 2.1, 2.3, and 2.5 hold for the transformed data, then,*

$$\lim_{p\to\infty, n\to\infty} \widehat{R}_{\widehat{h},0,n,p} = 0,$$

provided that $p\exp(-Cn^{1/3-\theta}) \to 0$ *for some* $C > 0$ *and* $0 < \theta < 1/3$*, and that there exists some constant* $c_6 > 0$ *such that* $|\mu_{2j}| < c_6$ *for all* $j = 1, \ldots, p$*.*

Compared with Theorem 2.2 and aside from the obvious additional Condition 2.6, Theorem 2.3 requires two more assumptions, about which we will make some remarks below.

Remark 2.6. *Recall that, previously for ppQDA and pQDA, we did not need to control the rate with which* p *goes to infinity relative to that of* n*, but we do now for Se-pQDA. This is because we must now estimate* p *univariate transformations. To ensure that we can estimate these transformations reasonably well, the dimension* p *cannot grow too fast relative to the overall sample size* n*. More precisely, we require* $p\exp(-Cn^{1/3-\theta}) \to 0$ *for some* $C > 0$ *and* $0 < \theta < 1/3$ *as both* p *and* n *tend to infinity.*

Remark 2.7. *The additional assumption in Theorem 2.3 — that every* $|\mu_{2j}|$ *is bounded — is introduced to avoid some unnecessary technical difficulties in our proof. This added assumption does not really weaken our result. If* $|\mu_{2j}|$ *is very large, it only makes classification easier, and the more challenging (and hence more interesting) problem in practice*

occurs when the marginal signals are relatively weak. This is especially relevant as we have not made any sparsity assumptions about $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Situations in which signals from the mean vectors are relatively dense (see, e.g., Fan et al., 2013b) are only interesting when those signals are marginally faint.

2.4 Outline of proofs

In this section, we give a brief outline of the main proofs, but the actual proofs are given in the Appendix.

2.4.1 Theorems 2.1 and 2.2

To prove Theorem 2.1, we first prove it for Q, using the true parameters $\boldsymbol{\mu}_i, a_i, r_i$. This is essentially the population version of the ppQDA rule. To prove it for $\widehat{Q}$, the sample version, our main idea is to write $\widehat{Q}$ as $(\widehat{Q} - Q) + Q$ and prove that the quantity, $\widehat{Q} - Q$, is dominated by Q as $p, n \to \infty$, so that we can conclude

$$\mathbb{P}(\widehat{Q} > 0 | \mathbf{x} \in \mathcal{C}_1) - \mathbb{P}(Q > 0 | \mathbf{x} \in \mathcal{C}_1) = \mathbb{P}(\widehat{Q} - Q + Q > 0 | \mathbf{x} \in \mathcal{C}_1) - \mathbb{P}(Q > 0 | \mathbf{x} \in \mathcal{C}_1) \to 0$$

and likewise for $\mathbb{P}(\widehat{Q} \leq 0 | \mathbf{x} \in \mathcal{C}_2)$. The proof of Theorem 2.2 is very similar (and in fact, easier), even though their conditions are somewhat different.

2.4.2 Theorem 2.3

In a nutshell, Theorem 2.3 is proved in three steps. First, we prove it for $Q_{h,0}$, assuming that we know the transformation h as well as the true distributional parameters (e.g., $\boldsymbol{\mu}_i, \Sigma_i, A_i$, and so on) for the transformed data $h(\mathbf{y}_{ik})$. Then, we prove it for an intermediate quantity, $Q_{\widehat{h},0}$, which uses the estimated transformation $\widehat{h}$ but nonetheless still uses the true distributional parameters for the transformed data — again, $\boldsymbol{\mu}_i, \Sigma_i, A_i$, and so on. This intermediate quantity is perhaps somewhat difficult to conceptualize in practice — how can we have the true parameters for the transformed data if the transformation itself

is unknown and estimated? Here, it is important to keep in mind that this is merely a hypothetical entity used as a "stepping stone" for the theoretical proof; it has no intrinsic value in itself. Finally, we prove it for $\widehat{Q}_{\widehat{h},0}$.

The result for $Q_{h,0}$ can be obtained "for free" as a result of having proved Theorem 2.2 already by this point. To obtain the results for $Q_{\widehat{h},0}$ and subsequently for $\widehat{Q}_{\widehat{h},0}$, the key lies in being able to bound various probabilities that the difference is large between a quantity that depends on $h_j(x_j)$ and its counterpart that depends on $\widehat{h}_j(x_j)$ — say, $J(h_j(x_j))$ and $J(\widehat{h}_j(x_j))$. This is achieved using a similar set of techniques as used by Mai and Zou (2015). Specifically, the real line $\mathbb{R}$ is divided into four (4) different regions depending on whether $h_j(x_j)$ is

- less than $O(\sqrt{\ln n})$ distance away from 0,
- between $O(\sqrt{\ln n})$ and $O(\ln n)$ distance away from 0,
- between $O(\ln n)$ and $O(poly(n))$ distance away from 0 — where $poly(n)$ means "polynomial" in n, or
- more than $O(poly(n))$ distance away from 0;

and different bounds are obtained for each region. As we move through the four regions in the order listed above, the bounds on the difference, $|J(h_j(x_j))-J(\widehat{h}_j(x_j))|$, get successively looser, but the corresponding probabilities for $h_j(x_j)$ to fall into these regions also decrease.

Although we have used techniques from Mai and Zou (2015), it does *not* mean that our proofs are essentially the same as theirs. The main difference is that they assumed sparsity. In the final step when we move from $Q_{\widehat{h},0}$ to $\widehat{Q}_{\widehat{h},0}$, our proof is similar to theirs, but in the second step when we focus on $Q_{\widehat{h},0}$, our proof is considerably different. Specifically, the misclassification error of $Q_{\widehat{h},0}$ depends critically on how many $h_j(x_j)$ falls outside the first region described above. For Mai and Zou (2015), their sparsity assumption meant only a small number of those would affect their classification rule, and the resulting error could be controlled relatively easily. Without making any sparsity assumptions, however, all of those falling outside the first region will affect our classification rule, so we must carry out a more careful analysis respectively in each of the three other regions in order to control

our error. Another difference is that they focused on semiparametric *linear*, as opposed to *quadratic*, discriminant rules. As a result, many of our error/probability bounds are necessarily different from theirs.

Remark 2.8. *We are now ready to say more about establishing theoretical results for Se-ppQDA, having outlined our proof of Theorem 2.3 above. By and large, the required techniques remain the same, but since ppQDA uses a non-diagonal matrix (even though it is still a very simple one), we must now consider the interactions between* $h_j(x_j)$ *and* $h_{j'}(x_{j'})$ *for all* $j \neq j'$*. To do so, we must now divide* $\mathbb{R} \times \mathbb{R}$ *into* $4 \times 4 = 16$ *different regions, and obtain different bounds in each of them. This will undoubtedly be much more tedious, but the fundamental ideas are the same. Hence, we have decided not to pursue it at the present stage.*

2.5 Numerical experiments

In this section, we demonstrate the performance of pQDA, ppQDA, Se-pQDA and Se-ppQDA by simulation. Three other methods — DSDA (Mai et al., 2012), SSDA (Mai and Zou, 2015) and random forest (Breiman, 2001) — are included for comparison purpose. Both DSDA and SSDA are penalized linear discriminant rules, and the latter deals with nonnormal data; for these methods we used the `R` package `dsda`, provided by the authors of the methods. For random forest, we used the `R` package `randomForest` with a forest size of 1000; for all other parameters, we simply used their default values as further adjustments did not noticeably affect the performance.

We also include a benchmark classifier, in which the true covariance matrices (Σ_1, Σ_2) and the *sample* means $(\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2)$ are plugged into the QDA rule. Note that we used only the true covariance matrices — but *not* the true mean vectors — in the benchmark classifier, because we would like to focus on the effect of using our structured covariance matrices for classification, and to avoid letting the estimation of the mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ (an intricate problem on its own in high dimensions) unduly confound our performance evaluation.

For all our QDA procedures (i.e., pQDA, ppQDA, Se-pQDA, Se-ppQDA), we standardized the variance of each dimension j by the larger of the two within-class standard deviations, i.e., $\max\{\widehat{\sigma}_{1j}, \widehat{\sigma}_{2j}\}$. In the case of Se-pQDA and Se-ppQDA, such standardization was performed after first estimating and then applying the transformation h_j.

2.5.1 Different covariance matrices

We considered nine types of covariance matrix structures. The number of explanatory covariates was set to either $p = 400$ or $p = 800$. We use $M[1 : p_0, 1 : p_0]$ to denote the $p_0 \times p_0$ sub-matrix consisting of the first p_0 rows and columns of M. We set $p_0 = \lfloor 5p^{2/3} \rfloor$ to control how the sub-matrix increases with p. For the purpose of brevity, below we describe only the "interesting part" of our nine matrices; the elements not explicitly described are 1 if on the diagonal and 0 if on the off-diagonal.

M_1: The matrix M_1 contains an autoregressive $p_0 \times p_0$ sub-matrix, with $M_{1,j_1 j_2} = 0.2^{|j_1 - j_2|}$ for $j_1, j_2 \in \{1, \ldots, p_0\}$.

M_2: The matrix M_2 is a perturbed version of M_1. With probability $1/p_0$, each element $0.2^{|j_1 - j_2|}$ from $M_1[1 : p_0, 1 : p_0]$ is randomly replaced by $0.3^{|j_1 - j_2|}$. The matrices M_1 and M_2 therefore differ by approximately p_0 elements.

M_3: The matrix M_3 is block diagonal. Each diagonal block is a $q \times q$ matrix, $0.2\mathbf{1}_q\mathbf{1}_q' + 0.8I_q$, where q is chosen to be 4.

M_4: The matrix M_4 is a modified version of M_1. In particular, the sub-matrix $M_4[1 : p_0, 1 : p_0]$ is designed to have the same eigenvectors as $M_1[1 : p_0, 1 : p_0]$ but different, randomly generated eigenvalues. Let T be the orthogonal matrix containing the eigenvectors of $M_1[1 : p_0, 1 : p_0]$. Then, $M_4[1 : p_0, 1 : p_0] = T(\text{diag}\{\nu_1, \ldots, \nu_{p_0}\})T'$, where $\nu_j \overset{i.i.d}{\sim} \text{Uniform}(1, 2)$.

M_5: The matrix M_5 is simply $M_5 = 0.2\mathbf{1}_p\mathbf{1}_p' + 0.8I_p$.

M_6: The matrix $M_6 = M_5^{-1}$ is simply the inverse of M_5.

M_7: The matrix M_7 is a perturbed version of M_5. First, with probability 0.2, each off-diagonal element from the first five (5) rows and columns of $M_5[1 : p_0, 1 : p_0]$ is randomly replaced by zero (0) — call the resulting matrix B. Then, we let $M_7 = (B + \lambda I_p)/(1 + \lambda)$, where $\lambda = \max\{-\lambda_{\min}(B), 0\} + 0.05$ and $\lambda_{\min}(B)$ is the smallest eigenvalue of B, to ensure that M_7 is positive definite.

M_8: The matrix M_8 is also a perturbed version of M_5, except here the perturbations are made to the diagonal elements. Specifically, $M_8 = M_5 + \text{diag}\{\nu_1, \ldots, \nu_p\}$, in which $\nu_j \overset{i.i.d}{\sim} \text{Uniform}(0, 1)$ for $j \leq p_0$ and $\nu_j = 0.5$ for $j \geq p_0 + 1$.

M_9: The matrix M_9 is largely unstructured, with mostly small entries other than a few large ones. First, a baseline matrix B_0 is generated by randomly sampling each element from $\text{Uniform}(0, 0.2)$. Then, five (5) elements are randomly deleted and re-drawn from $\text{Uniform}(0.2, 0.8)$ instead. Finally, to ensure symmetry and positive-definiteness, we let $B = (B_0 + B_0')/2$ and $M_9 = (B + \lambda I_p)/(1 + \lambda)$, where $\lambda = \max\{-\lambda_{\min}(B), 0\} + 0.05$ and $\lambda_{\min}(B)$ is the smallest eigenvalue of B.

2.5.2 Simulated examples

Based on these nine different types of matrices, we created ten simulated classification examples. In all of them, the means of the two classes were taken to be $\boldsymbol{\mu}_1 = \mathbf{0}_p$ and $\boldsymbol{\mu}_2 = (3.5p^{-1/2}\mathbf{1}'_{0.6p}, \mathbf{0}'_{0.4p})'$. That is, the signal was spread out evenly among the first $0.6p$ dimensions. The magnitude of the signal in each dimension was controlled so that the between-class Euclidean distance did not change with p. The ten examples differed mostly by the covariance matrices of the two classes. In all cases, we also controlled the difference between the two within-class covariance matrices by a parameter $s \equiv 3p^{-1/2}$ (see below).

Example 1: $\Sigma_1 = M_1$, partly autoregressive, and $\Sigma_2 = \Sigma_1 + sI_p$.

Example 2: $\Sigma_1 = M_3$, block diagonal, and $\Sigma_2 = \Sigma_1 + sI_p$.

Example 3: $\Sigma_1 = M_4$, modified version of M_1, and $\Sigma_2 = \Sigma_1 + sI_p$. This example is designed to investigate a case in which the covariance matrices have eigenvalues that

are quite close to each other — one in which pQDA is expected to perform well (see Remark 2.4).

Example 4: $\Sigma_1 = M_1$, partly autoregressive, and $\Sigma_2 = M_2 + sI_p$, also partly autoregressive, but with some elements (both diagonal and off-diagonal ones) being different from those in Σ_1.

Example 5: $\Sigma_1 = \Sigma_2 = M_1$, partly autoregressive, and identical between the two classes. This example is designed to test the performance of our QDA rules when the LDA rule is optimal.

Example 6: $\Sigma_1 = M_5$, compound symmetry, and $\Sigma_2 = \Sigma_1 + sI_p$. This is an example in which ppQDA is expected to have an advantage over pQDA.

Example 7: $\Sigma_1 = M_6$, also compound symmetry, and $\Sigma_2 = \Sigma_1 + sI_p$. The matrix M_6 is different from M_5 in that it has negative off-diagonal elements that are close to 0 and is almost not positive definite.

Example 8: $\Sigma_1 = M_7$, compound symmetry with off-diagonal perturbations, and $\Sigma_2 = \Sigma_1 + sI_p$.

Example 9: $\Sigma_1 = M_8$, compound symmetry with diagonal perturbations, and $\Sigma_2 = \Sigma_1 + sI_p$.

Example 10: $\Sigma_1 = M_9$, unstructured, and $\Sigma_2 = \Sigma_1 + sI_p$.

2.5.3 Results

For all of our ten simulated examples, we used $n_1 = n_2 = 100$ training samples, and 1000 independent testing samples, respectively from $N(\boldsymbol{\mu}_1, \Sigma_1)$ and $N(\boldsymbol{\mu}_2, \Sigma_2)$. All simulations were repeated for 100 times, and the average misclassification rates on the testing samples were recorded, together with their standard errors.

Table 2.1 shows how the methods compared on the ten examples. Our suite of methods were generally better than DSDA, SSDA and random forest. This is not surprising as both

DSDA and SSDA assume sparsity and identical within-class covariance matrices, and the random forest does not make (or take advantage of) any specific distribution assumption. In each example, the best method statistically matched the benchmark classifier. Recall that, for the benchmark classifier, we used only the true covariance matrices but still kept using the sample rather than the population mean vectors, so it was possible sometimes for other methods to outperform it.

In Examples 1-4, the covariance matrices are better approximated by diagonal ones, so pQDA is expected to perform well, but we see that ppQDA performed reasonably well, too. This indicates that, whenever pQDA works, ppQDA is only slightly worse than, if not as good as, pQDA.

In Example 5, the two within-class covariance matrices are the same, so LDA is actually optimal, but we see that both pQDA and ppQDA still continued to perform well.

In Examples 6-7, the covariance matrices have exactly the compound symmetry structure, so naturally in these cases we see that ppQDA performed considerably better than all other methods.

In Examples 8-9, the covariance matrices no longer have exactly the compound symmetry structure, due to perturbations to the various off-diagonal (M_7, Example 8) and diagonal (M_8, Example 9) elements. In Example 10, the covariance matrices are largely unstructured, except that a few randomly selected entries are much larger than others. These examples were designed to test the robustness and sensitivity of ppQDA. In all of these cases, ppQDA maintained good performance and sometimes still commanded a considerable advantage over all other methods.

In Table 2.1, we see that both Se-pQDA and Se-ppQDA performed slightly worse than their counterparts without any nonlinear transformations. Clearly, estimating these extra transformations when they were unnecessary introduced additional errors.

We also transformed the data from these ten examples to be non-normally distributed and repeated our experiments. In particular, after data were first generated from $N(\boldsymbol{\mu}_1, \Sigma_1)$ and $N(\boldsymbol{\mu}_2, \Sigma_2)$, we applied one of six nonlinear transformations — $g_{(1)}(\cdot), \ldots, g_{(6)}(\cdot)$, as listed in Table 2.2 — in each dimension. The first $\lfloor p/6 \rfloor$ dimensions were transformed by $g_{(1)}$; dimensions $\lfloor p/6 \rfloor + 1$ to $2\lfloor p/6 \rfloor$ were transformed by $g_{(2)}$; and so on. All remaining

dimensions, from $6\lfloor p/6 \rfloor + 1$ to p, were left untransformed. Table 2.3 shows the result. When the data were non-normal, the advantages of Se-pQDA and Se-ppQDA over other methods became clear. The benchmark classifier in Table 2.3 is the same as the one in Table 2.1, and is equivalent to using the true transformations, true covariance matrices, and sample means.

2.6 Real data analysis

To test the performance of our methods with real data, we used a colon cancer dataset (Alon et al., 1999), available in the `R` package `rda` at `https://CRAN.R-project.org/package=rda`, and a malaria dataset (Ockenhouse et al., 2006), available at `http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2362`. For our various QDA procedures, variables were standardized in the same manner as described in Section 2.5. For Se-pQDA and Se-ppQDA, the transformations $h_1, h_2, ..., h_p$ were estimated based on training data from the larger class (specifically, the "tumor" class for the colon cancer data, and the "infected" class for the malaria data), and any pre-processing operations (e.g., pre-screening, if applicable, and variable standardization) were performed after the transformation.

2.6.1 Colon cancer data

Alon et al. (1999) studied the colon cancer dataset by performing cluster analysis on both genes and tissues. The dataset consists of $n_1 = 40$ tumor and $n_2 = 22$ normal colon tissues. The original dataset contained more than $6,500$ features (genes), but the one available in the `rda` package contains only $2,000$ features with the highest minimal intensities across samples, which were used by Alon et al. (1999) in their cluster analysis. The dataset was randomly split into a training set (2/3) and a testing set (1/3). All discriminant rules were estimated from the training data and then applied to the testing data. This process was repeated 100 times.

Table 2.4 shows the average and median misclassification rates, together with their respective standard errors, from the 100 replications. Our pQDA and ppQDA rules were

Table 2.1: Average misclassification rates (%) and their standard errors. Data are generated from $N(\boldsymbol{\mu}_1, \Sigma_1)$, $N(\boldsymbol{\mu}_2, \Sigma_2)$.

	Example	pQDA	ppQDA	Se-pQDA	Se-ppQDA	DSDA	SSDA	RF	Benchmark
$p = 400$	1	**13.5(0.11)**	14.3(0.12)	14.1(0.12)	15.3(0.13)	32.3(0.26)	34.7(0.26)	24.6(0.13)	13.7(0.11)
	2	**13.7(0.11)**	14.7(0.12)	14.2(0.12)	15.6(0.13)	32.4(0.21)	35.1(0.24)	25.1(0.12)	14.1(0.11)
	3	**20.8(0.12)**	21.0(0.14)	21.2(0.11)	22.0(0.11)	35.6(0.31)	38.5(0.34)	30.4(0.14)	20.5(0.13)
	4	**13.6(0.09)**	14.5(0.11)	14.3(0.10)	15.5(0.12)	32.0(0.20)	34.9(0.30)	24.8(0.15)	13.5(0.10)
	5	**20.5(0.11)**	22.3(0.15)	21.9(0.12)	24.7(0.15)	31.7(0.26)	34.5(0.28)	26.8(0.13)	24.9(0.14)
	6	38.3(0.41)	**14.0(0.10)**	36.5(0.40)	16.4(0.11)	38.8(0.26)	38.4(0.25)	36.4(0.28)	13.0(0.08)
	7	13.4(0.10)	**0.00(0.00)**	15.7(0.11)	2.70(0.06)	28.0(0.19)	33.1(0.30)	25.9(0.14)	0.00(0.00)
	8	33.8(0.46)	**16.7(0.12)**	30.1(0.46)	17.7(0.13)	38.4(0.29)	38.8(0.25)	34.9(0.23)	6.50(0.07)
	9	39.3(0.35)	26.1(0.14)	37.1(0.35)	**25.7(0.12)**	42.4(0.23)	42.4(0.23)	39.0(0.17)	24.8(0.12)
	10	23.1(0.36)	**9.40(0.09)**	18.4(0.30)	11.0(0.11)	35.5(0.26)	36.1(0.25)	28.2(0.15)	5.50(0.06)
$p = 800$	1	**16.7(0.11)**	17.8(0.13)	17.1(0.12)	18.6(0.14)	36.8(0.21)	40.1(0.30)	29.7(0.12)	17.4(0.10)
	2	**17.2(0.12)**	18.2(0.14)	17.7(0.14)	19.2(0.15)	37.4(0.30)	40.5(0.27)	29.9(0.13)	17.8(0.11)
	3	25.6(0.13)	26.1(0.15)	**25.1(0.14)**	26.6(0.15)	40.8(0.33)	43.6(0.25)	35.5(0.12)	24.4(0.12)
	4	**16.6(0.11)**	17.7(0.11)	17.1(0.10)	18.7(0.11)	36.7(0.20)	39.7(0.25)	29.5(0.13)	17.4(0.11)
	5	**24.3(0.14)**	26.0(0.16)	26.2(0.15)	29.7(0.16)	36.5(0.35)	40.0(0.3)	31.7(0.13)	28.7(0.13)
	6	41.7(0.34)	**18.2(0.12)**	40.5(0.30)	20.0(0.12)	42.7(0.25)	43.1(0.25)	40.4(0.26)	17.0(0.10)
	7	18.9(0.12)	**0.00(0.00)**	20.2(0.13)	3.80(0.07)	37.2(0.33)	41.3(0.31)	32.1(0.13)	0.00(0.00)
	8	36.7(0.44)	22.0(0.13)	32.9(0.48)	**21.8(0.13)**	43.1(0.23)	43.1(0.21)	39.1(0.18)	8.40(0.08)
	9	42.9(0.25)	30.5(0.14)	40.8(0.30)	**29.5(0.14)**	45.4(0.20)	45.5(0.21)	42.7(0.18)	29.4(0.12)
	10	28.0(0.38)	**16.0(0.12)**	22.6(0.32)	16.6(0.12)	40.6(0.23)	40.8(0.23)	34.0(0.17)	0.60(0.02)

Table 2.2: List of non-linear transformations.

$g_{(1)}(y) = y^3$	$g_{(2)}(y) = \exp(y)$
$g_{(3)}(y) = \arctan(y)$	$g_{(4)}(y) = \Phi(y)$
$g_{(5)} = (y+1)^3$	$g_{(6)} = \arctan(2y)$

comparable with DSDA, which gave the best result on the same dataset as reported by a comprehensive review paper (Mai, 2013), but computationally our methods were much less expensive. For this dataset, the Se-pQDA and Se-ppQDA rules did not perform as well, but neither did SSDA, a clear indication that the extra data transformations $h_1, h_2, ..., h_p$ were unnecessary and having to estimate them only brought in extra estimation error.

2.6.2 Malaria data

The malaria dataset consists of $n_1 = 49$ infected and $n_2 = 22$ healthy samples. For each sample, expression levels are available for $22,283$ genes. The data was randomly split into a training set and a testing set, with a sample-size ratio of approximately 1:1. Afterwards, the genes were screened on the training set and the $p = 5000$ most significant ones were kept for discriminant analysis. The significance level for the screening test was decided by the smaller of two p-values, one from a two-sample t-test and another from an F-test of equal variance. Again, this process was repeated 100 times.

The rough pre-screening step was used to avoid excessive noise accumulation, as our theory for the semiparametric QDA classifiers (Theorem 2.3) requires that p does not grow too fast relative to the sample size n, due to the need to estimate p distinct univariate transformations — see Remark 2.6.

Table 2.5 reports the average and median misclassification rates, together with their respective standard errors. We can see that, for this dataset, the pQDA and ppQDA rules did not perform well, and neither did DSDA, but our Se-pQDA and Se-ppQDA rules produced the best results, with the SSDA trailing slightly behind. This suggests that not only were these data nonnormal, but there were also signals that linear classifiers could not capture. This is precisely the kind of situations in which our methods are useful.

Table 2.3: Average misclassification rates (%) and their standard errors. Data are generated from $N(\boldsymbol{\mu}_1, \Sigma_1)$, $N(\boldsymbol{\mu}_2, \Sigma_2)$, and then transformed by $g_{(1)}(\cdot), \ldots, g_{(6)}(\cdot)$.

	Example	pQDA	ppQDA	Se-pQDA	Se-ppQDA	DSDA	SSDA	RF	Benchmark
$p = 400$	1	19.7(0.11)	20.3(0.11)	**14.1(0.10)**	15.4(0.11)	37.0(0.31)	34.6(0.31)	24.5(0.13)	13.7(0.11)
	2	20.1(0.09)	20.8(0.09)	**14.5(0.11)**	15.8(0.12)	37.3(0.30)	34.9(0.26)	24.9(0.14)	14.1(0.11)
	3	26.9(0.11)	26.9(0.12)	**20.9(0.12)**	21.6(0.13)	40.3(0.22)	38.5(0.32)	30.3(0.12)	20.5(0.13)
	4	20.0(0.10)	20.6(0.10)	**14.1(0.11)**	15.4(0.11)	37.2(0.31)	35.0(0.33)	24.7(0.15)	13.5(0.10)
	5	25.9(0.16)	27.2(0.16)	**21.9(0.14)**	24.8(0.18)	41.0(0.28)	34.2(0.26)	26.8(0.13)	24.9(0.14)
	6	38.9(0.20)	30.4(0.15)	36.8(0.38)	**16.6(0.13)**	45.1(0.26)	38.6(0.25)	36.3(0.28)	13.0(0.08)
	7	21.8(0.11)	14.2(0.13)	15.6(0.12)	**2.80(0.06)**	36.9(0.29)	33.0(0.40)	25.7(0.14)	0.00(0.00)
	8	34.8(0.17)	28.6(0.10)	30.4(0.46)	**17.8(0.12)**	44.6(0.22)	38.6(0.26)	35.4(0.23)	6.50(0.07)
	9	39.5(0.21)	34.8(0.13)	36.9(0.40)	**25.7(0.13)**	47.0(0.17)	42.0(0.19)	39.6(0.18)	24.8(0.12)
	10	24.9(0.15)	19.0(0.09)	18.0(0.30)	**11.0(0.11)**	40.0(0.19)	36.0(0.21)	27.8(0.14)	5.50(0.06)
$p = 800$	1	22.5(0.11)	23.1(0.11)	**17.6(0.12)**	19.2(0.12)	43.5(0.30)	40.2(0.28)	29.6(0.13)	17.4(0.10)
	2	22.6(0.12)	23.1(0.13)	**17.6(0.13)**	19.1(0.15)	43.9(0.32)	40.7(0.30)	30.0(0.14)	17.8(0.11)
	3	29.9(0.12)	30.0(0.12)	**25.1(0.14)**	26.3(0.15)	46.0(0.21)	44.0(0.24)	35.7(0.11)	24.4(0.12)
	4	22.2(0.12)	22.7(0.13)	**17.1(0.13)**	18.6(0.14)	43.3(0.27)	40.7(0.30)	29.8(0.12)	17.4(0.11)
	5	29.1(0.12)	30.1(0.15)	**26.1(0.12)**	29.6(0.15)	46.0(0.25)	40.5(0.34)	31.9(0.13)	28.7(0.13)
	6	41.7(0.26)	33.8(0.16)	40.8(0.32)	**19.9(0.11)**	48.2(0.13)	42.6(0.26)	40.7(0.23)	17.0(0.10)
	7	25.3(0.10)	17.0(0.13)	20.3(0.13)	**4.00(0.07)**	44.4(0.31)	41.5(0.31)	32.3(0.14)	0.00(0.00)
	8	36.9(0.20)	31.4(0.12)	32.7(0.46)	**21.9(0.12)**	47.8(0.17)	43.0(0.23)	39.1(0.17)	8.40(0.08)
	9	42.1(0.21)	37.7(0.14)	40.2(0.37)	**29.2(0.15)**	48.7(0.17)	45.0(0.19)	42.6(0.17)	29.4(0.12)
	10	28.9(0.14)	24.0(0.10)	22.2(0.32)	**16.4(0.12)**	46.1(0.23)	41.1(0.25)	33.8(0.15)	0.60(0.02)

Table 2.4: Colon cancer data. Average and median misclassification rates and their standard errors. Standard errors for the median are obtained by bootstrapping.

Method	pQDA	ppQDA	Se-pQDA	Se-ppQDA	DSDA	SSDA
Average(%)	15.1(0.57)	15.2(0.58)	16.8(0.67)	16.6(0.66)	15.2(0.59)	19.6(0.79)
Median(%)	13.6(1.87)	13.6(2.06)	13.6(2.20)	13.6(2.10)	13.6(1.25)	18.2(1.10)

Table 2.5: Malaria data. Average and median misclassification rates and their standard errors. Standard errors for the median are obtained by bootstrapping.

Method	pQDA	ppQDA	Se-pQDA	Se-ppQDA	DSDA	SSDA
Average(%)	8.46(0.67)	6.91(0.59)	4.00(0.31)	3.69(0.30)	8.50(0.50)	4.90(0.42)
Median(%)	7.14(1.36)	5.71(0.84)	2.86(0.74)	2.86(0.32)	8.57(0.65)	5.71(1.09)

2.7 Discussion

So far our theoretical results have focused on establishing conditions under which our proposed methods (e.g., ppQDA, pQDA, Se-pQDA) can have nearly perfect performance asymptotically. In reality, of course, perfect classification is not always possible, in which case we would like to know how well our methods can perform relative to the Bayes decision rule. In this section, we will provide some answers to this question for ppQDA.

To do so, we further simplify the situation by focusing on a special case where there is no signal for classification in the class means, i.e., $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$. As we have already stated earlier (see Remark 2.2), since there are already many papers in the literature about classification based on signals from the mean vectors alone, and since our main idea of replacing Σ_i with A_i is "only" about dealing with large covariance matrices, we think it actually makes things clearer if we concentrate on just the covariance matrices and ignore the mean vectors.

We will also focus on the population version of the ppQDA rule. This is justified since

we already proved (see Section 2.4) the dominance of the population quantity Q over $\widehat{Q}-Q$ as $p, n \to \infty$. However, our proof has assumed condition 2.4, but this section is primarily concerned with situations in which asymptotically perfect classification is not achievable, so it would be desirable if this dominance could be established without condition 2.4-2. Indeed, this is possible, provided that some mild modifications are made to conditions 2.4-3 and 2.4-4. Specifically, instead of the difference between A_i and Σ_i being simply $o(p^2)$, now its order must also depend on how much signal there is for classification, as measured by $(a_{i_1} - r_{i_1})/(a_{i_2} - r_{i_2})$ for $(i_1, i_2) = (1, 2)$ and $(2, 1)$. A detailed proof is omitted, as the technique is similar to that used in the proof of Theorem 2.1.

2.7.1 The Bayes decision rule versus ppQDA

Let A_1, A_2, Σ_1 and Σ_2 be defined as in Section 2.2. Under the assumption that $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, the quantity that drives (population) ppQDA, using the true (as opposed to estimated) parameters, is

$$Q = \ln\left(|A_1|/|A_2|\right) + \mathbf{x}'A_1^{-1}\mathbf{x} - \mathbf{x}'A_2^{-1}\mathbf{x},$$

whereas the Bayes decision rule is driven by

$$Q_B = \ln\left(|\Sigma_1|/|\Sigma_2|\right) + \mathbf{x}'\Sigma_1^{-1}\mathbf{x} - \mathbf{x}'\Sigma_2^{-1}\mathbf{x}.$$

Clearly, the performance of ppQDA will be close to that of the Bayes rule if $\Sigma_i \approx A_i$ for both $i = 1, 2$, but we will argue below that this need not necessarily be the case.

To see this, suppose first that $\mathbf{x} \in \mathcal{C}_1$. Then, for any matrix B, we have

$$\begin{aligned}\mathbb{E}(\mathbf{x}'B\mathbf{x}|\mathbf{x} \in \mathcal{C}_1) = \mathbb{E}[tr(\mathbf{x}'B\mathbf{x})|\mathbf{x} \in \mathcal{C}_1] = \mathbb{E}[tr(B\mathbf{x}\mathbf{x}')|\mathbf{x} \in \mathcal{C}_1] \\ = tr[B\mathbb{E}(\mathbf{x}\mathbf{x}'|\mathbf{x} \in \mathcal{C}_1)] = tr(B\Sigma_1),\end{aligned}$$

which immediately implies

$$\mathbb{E}(Q_B|\mathbf{x} \in \mathcal{C}_1) \;=\; \ln|\Sigma_2^{-1}\Sigma_1| + p - tr(\Sigma_2^{-1}\Sigma_1), \tag{2.10}$$

and

$$\mathbb{E}(Q|\mathbf{x} \in \mathcal{C}_1) \;\; = \;\; \ln|A_2^{-1}A_1| + tr(A_1^{-1}\Sigma_1) - tr(A_2^{-1}\Sigma_1). \tag{2.11}$$

But the inverse formula for $\widehat{A}_i$, given in equation (2.5), applies to A_i as well, so we can write

$$tr(A_i^{-1}\Sigma_1) = [(a_i - r_i)^{-1}]tr(\Sigma_1) - [r_i(a_i - r_i)^{-1}(a_i + (p-1)r_i)^{-1}]tr(\mathbf{1}_p\mathbf{1}_p'\Sigma_1).$$

However, the definition of A_1 implies $tr(\Sigma_1) = tr(A_1)$ and

$$tr(\mathbf{1}_p\mathbf{1}_p'\Sigma_1) = tr(\mathbf{1}_p'\Sigma_1\mathbf{1}_p) = Su(\Sigma_1) = Su(A_1) = tr(\mathbf{1}_p'A_1\mathbf{1}_p) = tr(\mathbf{1}_p\mathbf{1}_p'A_1).$$

This means $tr(A_i^{-1}\Sigma_1) = tr(A_i^{-1}A_1)$ so that (2.11) can be further reduced to

$$\mathbb{E}(Q|\mathbf{x} \in \mathcal{C}_1) \;\; = \;\; \ln|A_2^{-1}A_1| + p - tr(A_2^{-1}A_1). \tag{2.12}$$

Together, equations (2.12) and (2.10) are highly suggestive of the possibility that, given $\mathbf{x} \in \mathcal{C}_1$, the performance of ppQDA can be close to that of the Bayes rule as long as $A_2^{-1}A_1$ is close to $\Sigma_2^{-1}\Sigma_1$ in the sense that

$$tr(A_2^{-1}A_1) \approx tr(\Sigma_2^{-1}\Sigma_1) \quad \text{and} \quad |A_2^{-1}A_1| \approx |\Sigma_2^{-1}\Sigma_1|,$$

whereas each A_i need not be close to Σ_i in itself.

Moreover, for two $p \times p$, symmetric, positive-definite matrices U, V, we can define the function,

$$\phi(U,V) \;\; = \;\; \Big|\ln|V^{-1}U| + p - tr(V^{-1}U)\Big|,$$

as one way to measure their difference — notice that $\phi(U,V) = 0$ if $U = V$, and the absolute value is needed because, for any $p \times p$, symmetric, positive-definite matrix M with eigenvalues $\lambda_1, ..., \lambda_p$, the function $\ln|M| + p - tr(M) = \sum(\ln\lambda_j + 1 - \lambda_j) \leq 0$ with equality only when $\lambda_j = 1$ for all j; see also Remark 2.5. For $\mathbf{x} \in \mathcal{C}_1$, our analysis above shows that, on average, the Bayes rule and the ppQDA rule are simply using the same $\phi(\cdot,\cdot)$ function to measure the differences between a different set of matrices — (Σ_1, Σ_2) for the Bayes rule and (A_1, A_2) for ppQDA.

Combined with arguments similar to those we used to prove Theorem 2.1 (see Section 2.4), our analysis above also suggests that, for $\mathbf{x} \in \mathcal{C}_1$, the performance of ppQDA can be asymptotically close to that of the Bayes rule if

$$\frac{\phi(\Sigma_1, \Sigma_2) - \phi(A_1, A_2)}{\phi(\Sigma_1, \Sigma_2)} \sim o(1)$$

as $p \to \infty$.

The same argument applies to the case of $\mathbf{x} \in \mathcal{C}_2$, except that, in this case, the differences are measured by $\phi(A_2, A_1)$ and $\phi(\Sigma_2, \Sigma_1)$ instead of by $\phi(A_1, A_2)$ and $\phi(\Sigma_1, \Sigma_2)$. Thus, we define the symmetric difference measure,

$$\varphi(U, V) \quad = \quad \phi(U, V) + \phi(V, U),$$

and conjecture that the relative performance of our ppQDA rule to that of the Bayes rule depends very much on the quantity,

$$\Delta \equiv \frac{\varphi(\Sigma_1, \Sigma_2) - \varphi(A_1, A_2)}{\varphi(\Sigma_1, \Sigma_2)}, \tag{2.13}$$

and whether $\Delta \to 0$ as $p \to \infty$. We present some empirical evidence below to support this observation.

2.7.2 Empirical evidence

In this section, we re-examine *some* examples from Section 2.5 to see (i) how the quantity Δ, given in (2.13), changes with p; and (ii) how it relates to the overall misclassification error.

Not all examples from Section 2.5 are included because some of them — in particular, examples 5, 6, 7 — do not contribute any information to either question (i) or question (ii) above. In example 5, $\Sigma_1 = \Sigma_2$, which means $\varphi(\Sigma_1, \Sigma_2) = 0$, so Δ is not well defined. In examples 6 and 7, $\Sigma_i = A_i$ for both $i = 1, 2$, which means $\varphi(\Sigma_1, \Sigma_2) - \varphi(A_1, A_2) = 0$, so $\Delta = 0$ as well. We also remove classification signals contained in the location parameters by setting $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, and focus on signals contained in the covariance matrices alone.

For question (i), Table 2.6 shows that the quantity, Δ, generally decreases with p. For question (ii), Figure 2.1 shows that small values of Δ are highly predictive of small gaps between the performance of ppQDA and that of the Bayes rule.

Table 2.6: The quantity Δ versus p.

Example	$p = 100$	$p = 400$	$p = 800$	$p = 1000$
1	0.1624	0.1312	0.1112	0.1051
2	0.1728	0.1367	0.1160	0.1094
3	0.0973	0.0468	0.0305	0.0268
4	0.1566	0.1304	0.1091	0.1051
8	0.4911	0.4026	0.3267	0.3237
9	0.1228	0.0966	0.0720	0.0702

Remark 2.9. *In this section, we have focused on the special case where $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$. For the more general case where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \neq \mathbf{0}$, similar arguments can be carried through, except equations (2.10) and (2.12) will each contain an extra term — respectively,*

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad \textit{and} \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'A_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

But we can still arrive at the same conclusions, provided that *we re-define the function ϕ as*

$$\phi(U, V) \;=\; \Big|\ln|V^{-1}U| + p - tr(V^{-1}U)\Big| + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'V^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Then, the function

$$\begin{aligned}\varphi(U, V) \equiv \phi(U, V) + \phi(V, U) = \\ \Big|\ln|V^{-1}U| + p - tr(V^{-1}U)\Big| + \Big|\ln|U^{-1}V| + p - tr(U^{-1}V)\Big| \\ + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\left(V^{-1} + U^{-1}\right)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\end{aligned}$$

is still a symmetric measure of difference between two classes, except it now measures differences not only between U and V but also between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ — e.g., $\varphi(U, V) = 0$ if and only if both $U = V$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. This is very much analogous to condition 2.5-2 for Theorem 2.2.

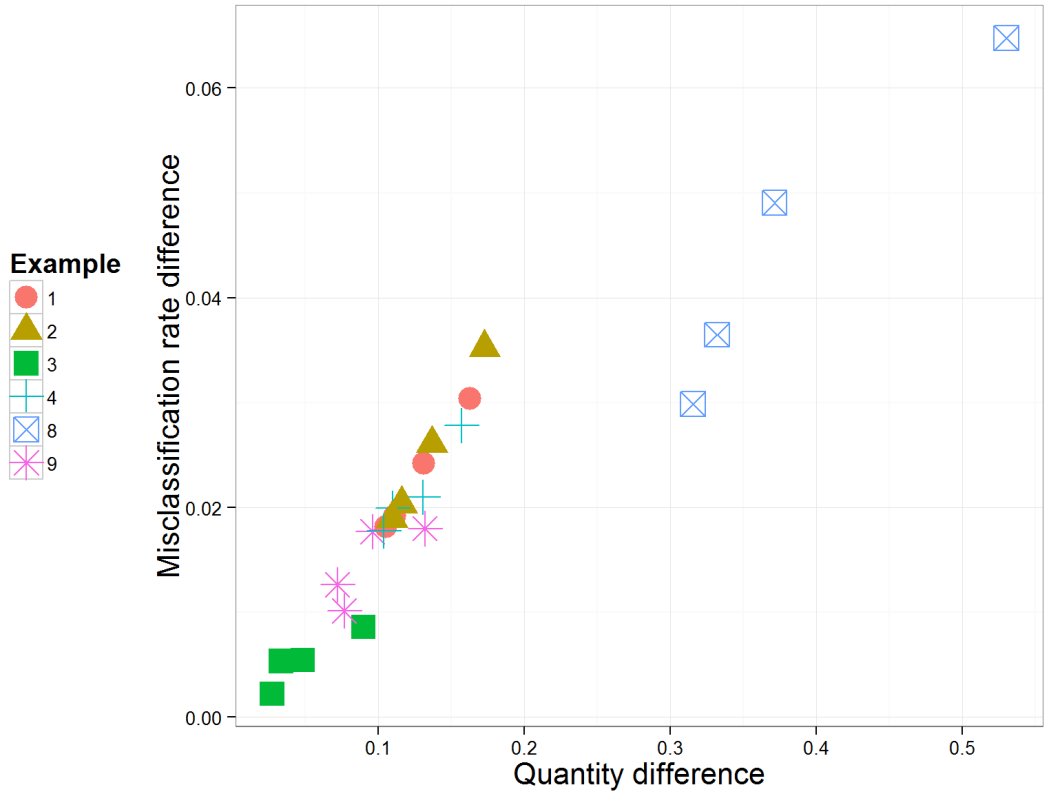

Figure 2.1: The difference, $\widehat{e}(Q) - \widehat{e}(Q_B)$, versus Δ, where $\widehat{e}(Q)$ denotes a Monte Carlo estimate (based on 100 test samples) of $e(Q) \equiv \mathbb{P}(Q > 0|\mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(Q \leq 0|\mathbf{x} \in \mathcal{C}_2)$, the misclassification error of the ppQDA rule, and likewise for $\widehat{e}(Q_B)$.

2.8 Conclusion

We have proposed two simple rules — namely, ppQDA and pQDA — to perform quadratic discriminant analysis for high-dimensional data, and generalized both rules by using a semiparametric transformation in order to handle data that do not necessarily follow the normal distribution. Desirable theoretical properties have been established for ppQDA, pQDA, and Se-pQDA — the semiparametric extension of pQDA. The performances of our specialized quadratic discriminant rules are comparable to, if not better than, other high-dimensional discriminant analysis methods in many numerical experiments and several real-data examples.

Unlike many existing high-dimensional discriminant analysis methods that focus on LDA, our methods aim at performing QDA, which allows us to exploit the difference between covariance matrices from separate classes and use it for classification. The sample covariance matrix is inconsistent when the dimension is high. Whereas most methods address this difficulty by imposing sparsity conditions, we do so by simplifying the structure of covariance matrices while still trying to capture some subtle information from across all dimensions. The special matrix structure that we use can be viewed as a generalization of the trace estimator, which has been used in high-dimensional hypothesis-testing as well as classification problems. Specifically, we pool not only the diagonal elements but also the off-diagonal ones in each covariance matrix, so as to obtain some information about the correlations among different dimensions. As a result, our easy-to-apply discriminant rules enjoy very low computational costs. The sparsity approach can be quite unstable for weak signals, and is more suitable for dealing with cases with just a few strong signals. Our approach is more attractive for cases with many weak signals.

Because of the complexity of the problem, at this point it is difficult to imagine that there could be a universally optimal discriminant analysis method for high-dimensional data. Almost every method can enjoy some advantages under certain circumstances. Due to noise accumulation, the performance of our methods could certainly deteriorate when there are a large number of useless covariates, but so would most methods. Due to the special matrix structure that we use, which has a common set of diagonal elements and a common set of off-diagonal ones, one may also expect that our discriminant rules may not

perform too well if the marginal variances across different dimensions are vastly different, or if some dimensions are very highly correlated while others have little correlation. In practice, however, these two problems can be alleviated by pre-screening and properly pre-processing the data, as we already have seen in Section 2.6. Our current main interest lies in the question of what other special matrix structures we can exploit for high-dimensional QDA. Prominent candidates must allow us to capture more information in each covariance matrix (than what can be captured by just two scalars a_i, r_i), but still have a relatively small number of "easily estimable" parameters.

Chapter 3

High-dimensional Covariance Matrix Estimation using a Diagonal and Low-rank Decomposition

3.1 Introduction

In this chapter, we focus on the covariance matrix estimation itself, instead of one of its applications, discriminant analysis. The emphasis is laid on the estimation accuracy rather than the misclassification rate. The method to be proposed is inspired by, and a generalization of, the compound symmetry structure considered in Chapter 2.

3.1.1 High-dimensional covariance matrix estimation

Before proceeding with the detail of our method, we conduct a brief review on covariance/precision matrix estimation in high dimensions.

The simplest estimator can be built using a scaled identity or a diagonal matrix as a substitute for the sample covariance matrix S. It is well-known that the sample covariance matrices tend to overestimate the large eigenvalues and underestimate the small eigenvalues

of the population covariance matrix; this bias can be corrected by shrinking the sample covariance matrix towards a scaled identity matrix, e.g., $\mathrm{tr}(S)I_p$ (Friedman, 1989). An optimal weight for the convex linear combination between the sample covariance matrix and the identity matrix has been proposed and studied by Ledoit and Wolf (2004). Ignoring the correlations and preserving only the diagonal part of S is a long-established practice in the high-dimensional classification, often referred to as the independence rule or the "naive Bayes classifier"; it has been demonstrated to outperform Fisher's linear discriminant rule under certain conditions (Dudoit et al., 2002; Bickel and Levina, 2004; Fan and Fan, 2008).

Apart from the scaled identity matrix and the diagonal matrix, other structured estimators have also been proposed. Methods such as banding (Bickel and Levina, 2008) and tapering (Furrer and Bengtsson, 2007) are useful when the covariates have a natural ordering (Rothman et al., 2008). Cai et al. (2013b) studied banding and tapering estimators in estimating large Toeplitz covariance matrices, which arise in the analysis of stationary time series.

Another popular assumption is sparse covariance or precision matrices. Sparse covariance matrix estimators can be obtained by either thresholding or regularization. Thresholding has been studied by Bickel and Levina (2008) and Cai and Liu (2011a), and applied in discriminant analysis by Shao et al. (2011) and Li and Shao (2015). To encourage sparsity, Rothman (2012) and Xue et al. (2012) imposed lasso-type penalties on the covariance matrix. Sparsity is a good assumption for the precision matrix in many applications, e.g., for Gaussian data zeros in the precision matrix suggest conditional independence; it can be achieved directly by imposing an ℓ_1 penalization on the precision matrix (Yuan and Lin, 2007; Rothman et al., 2008; Banerjee et al., 2008; Friedman et al., 2008; Lam and Fan, 2009; Cai et al., 2011) or indirectly through regularized regression (Meinshausen and Bühlmann, 2006; Rocha et al., 2008; Yuan, 2010; Sun and Zhang, 2013).

In the context of high-dimensional data analysis, it is reasonable to assume that the variance of the observed data can be explained by a small number of latent factors; thus, factor models can be applied to reduce the number of parameters in covariance matrix estimation, too. Assuming observable factors and independent error terms, Fan et al. (2008) proposed a covariance matrix estimator by estimating the loading matrix with regression and the covariance matrix of the error terms with a diagonal matrix. This

method was generalized by Fan et al. (2011) so that the error covariance was not necessarily diagonal, but it was assumed to be sparse and estimated with thresholding techniques. Fan et al. (2013a) then considered the case where the factors are unobservable. Assuming the number of latent factors (k) to be known, they performed PCA on the sample covariance matrix, kept the first k principal components to estimate the covariance matrix of the latent factors, and thresholded the remaining principal components to estimate a sparse covariance matrix for the error terms.

A related matrix structure is called "spiked covariance matrix", that is, the covariance matrix has only a few eigenvalues greater than one and can be decomposed into a low-rank matrix plus an identity matrix (Johnstone, 2001). Cai et al. (2015) proposed a sparse spiked covariance matrix estimator. In addition to the spiked structure, they assumed that the matrix spanned by the eigenvectors of the low-rank component has a small number of nonzero rows, which in turn constrains the covariance matrix to have a small number of rows and columns containing nonzero off-diagonal entries.

Chandrasekaran et al. (2012) proposed a latent variable method for Gaussian graphical model selection, based on the conditional independence interpretation of zero off-diagonals in the precision matrix. Assuming the observable and latent variables are jointly distributed as Gaussian, they showed that, if one assumes (i) the conditional precision matrix of the observables given the latent factors is sparse and (ii) the number of latent factors is small, then the marginal precision matrix of the observables must consist of a sparse component plus a low-rank component. The authors then considered a penalized likelihood approach to estimate such a marginal precision matrix, using the ℓ_1-norm to regularize the sparse component and the nuclear-norm to regularize the low-rank component. They also derived some consistency results for their estimator in the operator norm. Taeb and Chandrasekaran (2017) extended this framework to allow the incorporation of covariates.

A comprehensive review has been provided by Cai et al. (2016b), in which they also compared some of the aforementioned methods in terms of their respective convergence rates.

3.1.2 Outline and summary of this chapter

In this chapter, we make the explicit structural assumption that the population covariance/precision matrix can be decomposed into a diagonal plus a low-rank matrix, in order to facilitate the estimation of large covariance/precision matrices in high dimensions. In Section 3.2, we discuss this main model assumption in more detail.

While this model assumption is similar (but not identical) to some of the works reviewed in Section 3.1.1, the main difference is that we do not rely on nuclear norm regularization to promote low-rank-ness; instead, we directly impose a penalty on the matrix rank itself. In Section 3.3 and Section 3.4, we present estimators of the covariance/precision matrix under this model assumption, and show that estimation consistency can be achieved with a proper choice of the penalty function.

As is often the case, our estimators are characterized, or defined, as solutions to various optimization problems. In Section 3.5, we describe an efficient blockwise coordinate descent algorithm for solving the main optimization problem. In particular, given the low-rank component, the diagonal component can be obtained by solving a relatively cheap semi-definite program; given the diagonal component, the low-rank component actually can be obtained analytically. Since optimization with nuclear-norm constraints is still computationally burdensome for large matrices, we think our approach, which avoids nuclear-norm regularization, can be especially attractive.

In Section 3.6 and Section 3.7, we demonstrate the performances of our method with various simulations and an analysis of some real financial data. All proofs are relegated to the appendices.

3.1.3 Notations

We use $\mathbf{R}^{p_1 \times p_2}$ to denote the set of $p_1 \times p_2$ matrices, $\mathbf{S}^p$ to denote the set of symmetric $p \times p$ matrices, $\mathbf{S}^p_+$ to denote the subset of matrices $\subset \mathbf{S}^p$ which are positive semi-definite, and $\mathbf{S}^p_{++}$ to denote the subset of those which are strictly positive definite. Sometimes, another superscript is added to denote a restriction on the rank, for example, $\mathbf{S}^{p,r}$ is used

to denote the subset of matrices in $\mathbf{S}^p$ with rank $\leq r$, and likewise for $\mathbf{S}_+^{p,r}$, $\mathbf{S}_{++}^{p,r}$. For the corresponding sets of diagonal matrices, we replace $\mathbf{S}$ with $\mathbf{D}$, e.g., $\mathbf{D}^p$, $\mathbf{D}_+^p$, and $\mathbf{D}_{++}^p$.

For any $A \in \mathbf{S}^p$, we use $\mathrm{tr}(A)$ to denote its trace, $|A|$ to denote its determinant, and $\lambda_{\max}(A), \lambda_{\min}(A)$ to denote its largest and smallest eigenvalues. Furthermore, we use $\|A\|_F = \{\mathrm{tr}(A^T A)\}^{1/2}$ to denote its Frobenius norm, $\|A\|_* = \mathrm{tr}\{(A^T A)^{1/2}\}$ to denote its nuclear norm (which is equivalent to the sum of its singular values), $\|A\|_{op} = \{\lambda_{\max}(AA^T)\}^{1/2}$ to denote its operator norm, and $\|A\|_1 = \sum_{i,j} |A_{ij}|$ to denote its ℓ_1 norm. The function $\mathrm{diag}(\cdot)$ converts a vector to a diagonal matrix by setting the diagonal elements to be the input vector and a matrix to a vector by extracting the diagonal elements.

3.2 Problem set-up and model assumption

Consider a random sample $X = (x_1, \ldots, x_n)$, in which $x_1, \ldots, x_n$ are independently and identically distributed p-variate random vectors from the multivariate normal distribution with population mean 0 and population covariance matrix Σ_0. (We assume that the data have been centered in order to focus on the covariance matrix estimation problem alone, but it is important to point out that, in high dimensions, even estimating the mean vector is an intricate problem and much research has been conducted to address it.) The sample covariance matrix S, is a natural estimator of Σ_0 if p is fixed and $n \to \infty$, but it can perform badly when p is close to or larger than n, so some additional structural constraints are needed in order to facilitate estimation. We study a particular type of such structural constraints.

The main model assumption in our work here is that the population covariance matrix, $\Sigma_0 \in \mathbf{S}_{++}^p$, can be decomposed as

$$\Sigma_0 = L_{\Sigma_0} + D_{\Sigma_0},$$

in which $L_{\Sigma_0} \in \mathbf{S}_+^{p,r_0}$ is a low-rank matrix for some $r_0 \leq p$, and $D_{\Sigma_0} \in \mathbf{D}_{++}^p$ is a diagonal matrix.

Such a decomposition is always possible as long as $r_0 \leq p$, but only for reasonably small r_0 is the assumed decomposition interesting and valuable for estimating large covariance

matrices. Thus, for a particular matrix Σ_0, we define r_0 as the smallest among all attainable ranks of L_{Σ_0} after the decomposition, i.e., $r_0 = \text{rank}(L^*)$ in which

$$\begin{aligned} L^* &= \arg\min_{L} \ \text{rank}(L), \\ \text{subject to} \qquad & L + D = \Sigma_0, \ L \in \mathbf{S}_+^p, \ D \in \mathbf{D}_{++}^p. \end{aligned} \tag{3.1}$$

As a solution of (3.1), the matrix L^* itself might not be unique, but the optimal value r_0 is.

How should one understand this model assumption conceptually? As our first intuition, the assumption can be viewed as a generalization of the compound symmetry structure

$$\begin{bmatrix} a & b & \cdots & b \\ b & a & \cdots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \cdots & a \end{bmatrix}$$

with $a > b$, which was exploited earlier in Chapter 2 as a special structure to facilitate quadratic discriminant analysis in high dimensions. Notice that covariance matrices having the compound symmetry structure above can be decomposed into a rank-one matrix plus a scaled identity matrix,

$$b1_p1_p^{^T} + (a-b)I_p,$$

in which 1_p is a vector of ones and I_p is the $p \times p$ identity matrix. Therefore, the compound symmetry structure can be seen as a special case of the "diagonal + low-rank" decomposition.

The proposed decomposition also coincides with the factor analysis model and enjoys a nice interpretation. It is equivalent to assuming that the observed random vector x depends on a potentially smaller number of latent factors, i.e., $x = Rz + \epsilon$, in which z is some unobserved r_0-dimensional random vector from a normal distribution with mean 0 and variance I_{r_0}, R is an unobserved $p \times r_0$ loading matrix, and ϵ is a p-dimensional vector of independently distributed error terms with zero mean and finite variance, $\text{cov}(\epsilon) = \Psi$. Under the given structure, it is straight-forward to see that $\text{cov}(x) = RR^{^T} + \Psi$, in which $RR^{^T} \in \mathbf{S}_+^{p,r_0}$ is a low-rank matrix and $\Psi \in \mathbf{D}_{++}^p$ is a diagonal matrix. For our purpose,

we are not interested in estimating the loading matrix or analyzing the latent factors; we merely exploit the special structure to help us estimate Σ_0. This purely "utilitarian" use of the factor model is also the reason why we can define r_0 simply as the smallest attainable rank in the "diagonal + low-rank" decomposition.

Finally, we can also think of the "diagonal + low-rank" assumption as an alternative to the popular sparsity assumption to facilitate the estimation of large covariance matrices. Numerous methods with lasso-type penalties assume a large number of zero off-diagonal entries in Σ_0; undoubtedly *some* of these sparse structures can be represented as the sum of a low-rank matrix (i.e., with many empty rows and columns) and a diagonal matrix. The rank constraint is also somewhat analogous to the sparsity constraint. Specifically, the rank of L_{Σ_0} is the number of its non-zero eigenvalues, so low-rank means its spectrum (i.e., set of eigenvalues) is sparse. Like the sparsity constraint, a rank constraint also reduces the total number of parameters to be estimated, as lower ranks of L_{Σ_0} imply more linearly dependent columns and rows in L_{Σ_0}.

3.3 Precision matrix estimation with fixed rank

3.3.1 The estimation method

Our main model assumption can be equivalently imposed either on the covariance or on the corresponding precision matrix. Let $\Theta_0 = \Sigma_0^{-1}$ be the corresponding precision matrix. To understand the structure of Θ_0 when Σ_0 has the aforementioned "diagonal + low-rank" decomposition, we notice by a result of Henderson and Searle (1981) that

$$\begin{aligned}(L_{\Sigma_0} + D_{\Sigma_0})^{-1} &= -D_{\Sigma_0}^{-1}\left(I_p + L_{\Sigma_0} D_{\Sigma_0}^{-1}\right)^{-1} L_{\Sigma_0} D_{\Sigma_0}^{-1} + D_{\Sigma_0}^{-1} \\ &\triangleq -L_0 + D_0, \end{aligned} \tag{3.2}$$

in which $L_0 \in \mathbf{S}_+^{p,r_0}$ and $D_0 \in \mathbf{D}_{++}^p$, because the product of several matrices has rank at most equal to the minimum rank of all the individual matrices in the product, and the inverse of a matrix in $\mathbf{D}_{++}^p$ is still in $\mathbf{D}_{++}^p$. Therefore, we see that the precision matrix Θ_0 has an equivalent decomposition.

With this in mind, we will henceforth concentrate on estimating the precision matrix rather than the covariance matrix. This is in line with various recent literatures on covariance matrix estimation; the precision matrix is also the more "natural" variable for maximizing the Gaussian-likelihood and the more "direct" quantity to use in many statistical procedures such as discriminant analysis.

Other than the main "diagonal + low-rank" condition, our theoretical results will also require a "bounded eigenvalue" condition (see Condition 3.1 below), which is purely technical but common in the literature. Thus, our entire set of conditions about the population covariance/precision matrix is as follows:

Condition 3.1. *There exist constants $c_1, c_2 > 0$ such that $c_1 \leq \lambda_{\min}(\Sigma_0) \leq \lambda_{\max}(\Sigma_0) \leq c_2$, or equivalently, $c_2^{-1} \leq \lambda_{\min}(\Theta_0) \leq \lambda_{\max}(\Theta_0) \leq c_1^{-1}$, uniformly with respect to p.*

Condition 3.2. *For some $r_0 = o(p)$, the population covariance matrix $\Sigma_0 \in \mathbf{S}^p_{++}$ can be decomposed as $\Sigma_0 = L_{\Sigma_0} + D_{\Sigma_0}$, where $L_{\Sigma_0} \in \mathbf{S}^{p,r_0}_{+}$ and $D_{\Sigma_0} \in \mathbf{D}^p_{++}$; or equivalently, the precision matrix $\Theta_0 \in \mathbf{S}^p_{++}$ can be decomposed as $\Theta_0 = -L_0 + D_0$, where $L_0 \in \mathbf{S}^{p,r_0}_{+}$ and $D_0 \in \mathbf{D}^p_{++}$.*

In this section, we shall first consider a simple version of the problem, in which the rank of L_0 is pre-specified. We will consider the more general version of the problem later in Section 3.4. One pragmatic reason for first considering a simple (and perhaps somewhat unrealistic) version of the problem is because our main result regarding the more general version and our computational algorithm for solving it are both based on results that we shall derive in this section for the simple version.

For the simple version, a natural precision matrix estimator is

$$\begin{aligned} (\widehat{\Theta}_r, \widehat{L}_r, \widehat{D}_r) &= \arg\min_{\Theta} \{\mathrm{tr}(\Theta S) - \log|\Theta|\}, \\ \text{subject to} \quad & \Theta = -L + D,\ \Theta \in \mathbf{S}^p_{+},\ L \in \mathbf{S}^{p,r}_{+},\ D \in \mathbf{D}^p, \end{aligned} \tag{3.3}$$

in which r is a pre-specified constant. The objective function is the negative log-likelihood of the normal distribution, up to a constant. Let

$$\mathbf{F}_r = \{\Theta \in \mathbf{S}^p_{++} \mid L \in \mathbf{S}^{p,r}_{+}, D \in \mathbf{D}^p_{++} \text{ and } \Theta = -L + D\}$$

denote the search space of the optimization problem given in (3.3). In Sections 3.3.2 and 3.3.3 below, we will establish theoretical results to the following effects: (i) if the pre-specified constant $r \geq r_0$, then the true precision matrix $\Theta_0 \in \mathbf{F}_r$, but if r is much larger than r_0, the search space can be "too large" and solving (3.3) will become inefficient for estimating Θ_0; (ii) if the pre-specified constant $r < r_0$, then $\Theta_0 \notin \mathbf{F}_r$, and the gap between $\widehat{\Theta}_r$ and Θ_0 will depend on the distance between Θ_0 and the search space $\mathbf{F}_r$.

Remark 3.1. *In (3.3), it is unnecessary to explicitly restrict Θ or D to be positive definite. The $-\log|\Theta|$ term in the objective function and the constraint $\Theta \in \mathbf{S}_+^p$ together will guarantee $\Theta \in \mathbf{S}_{++}^p$. In addition, as $\Theta = -L + D$ and $L \in \mathbf{S}_+^{p,r}$, we will also automatically have $D \in \mathbf{D}_{++}^p$, for Θ may not be in $\mathbf{S}_{++}^p$ otherwise.*

Remark 3.2. *The non-uniqueness of $\widehat{L}_r$ and $\widehat{D}_r$ is inconsequential for our purposes; our results and discussions below only depend on $\widehat{\Theta}_r$ being a feasible minimizing solution to (3.3).*

3.3.2 The conservative case: $r \geq r_0$

To pre-specify the rank of L_0, denoted by r, it is generally advisable to err on the conservative side by choosing it to be large enough so that one can be more or less sure that $r \geq r_0$.

Theorem 3.1. *Under Conditions 3.1 and 3.2, if $r \geq r_0$ and $\widehat{\Theta}_r$ is a solution of (3.3), then*

$$\|\widehat{\Theta}_r - \Theta_0\|_F = O_p\left\{\max(a_{n,p,r}, b_{n,p})\right\},$$

in which

$$a_{n,p,r} = r^{1/2}(p/n)^{1/2}, \quad b_{n,p} = \left\{(p\log p)/n\right\}^{1/2}.$$

The true rank, r_0, may be fixed and finite, or it may diverge to infinity with p and n. Since Theorem 3.1 concerns the case of $r \geq r_0$, if $r_0 \to \infty$, then r must necessarily also go to infinity. Hence, finite choices of $r \geq r_0$ are only possible if r_0 is also finite. If r_0 is finite and we choose a finite $r \geq r_0$, the consistency of $\widehat{\Theta}_r$ is driven by $b_{n,p}$, whose order is greater than that of $a_{n,p,r}$, and the theorem basically suggests that choosing $r \geq r_0$ conservatively

will not hurt estimation in any fundamental way. Otherwise if we must choose a diverging r, it becomes possible for the convergence rate to be driven by $a_{n,p,r}$, and the theorem basically implies that the estimator $\widehat{\Theta}_r$ will be less efficient for larger, more conservative, choices of r.

3.3.3 The aggressive case: $r < r_0$

What if one errs on the aggressive side by choosing r to be too small so that $r < r_0$? Let

$$d_r = \min_{\Theta \in \mathbf{F}_r} \|\Theta - \Theta_0\|_F$$

be the distance from Θ_0 to the search space $\mathbf{F}_r$. When $r \geq r_0$, $d_r = 0$. When $r < r_0$, the true precision matrix Θ_0 is no longer in the search space $\mathbf{F}_r$, and $d_r > 0$. Under such circumstances, it is still possible to achieve the same level of performance provided that d_r is not too large.

Theorem 3.2. *Under Conditions 3.1 and 3.2, if $r < r_0$, $d_r = O\{\max(a_{n,p,r_0}, b_{n,p})\}$, and $\widehat{\Theta}_r$ is a solution of (3.3), then*

$$\|\widehat{\Theta}_r - \Theta_0\|_F = O_p\left\{\max(a_{n,p,r_0}, b_{n,p})\right\},$$

in which

$$a_{n,p,r_0} = r_0^{1/2}(p/n)^{1/2}, \quad b_{n,p} = \{(p \log p)/n\}^{1/2}.$$

While the proof itself is given in the appendices, the main reason why Theorem 3.2 holds is as follows. Let $\Theta_r \in \mathbf{F}_r$ be the matrix closest to Θ_0 such that $\|\Theta_r - \Theta_0\|_F = d_r$. It can be shown that $\widehat{\Theta}_r$, as the solution to maximizing the likelihood function in the search space $\mathbf{F}_r$, will be close to Θ_r. So, if d_r is small, $\widehat{\Theta}_r$ will also be reasonably close to Θ_0. More importantly, the condition $d_r = O\{\max(a_{n,p,r_0}, b_{n,p})\}$ requires the distance d_r to be of order $\max(a_{n,p,r_0}, b_{n,p})$, which, by Theorem 3.1, is also the order of the estimation error when the rank r is correctly set to be r_0. As a result, the error caused by Θ_0 being away from $\mathbf{F}_r$ is relatively small and does not increase the order of the estimation error.

According to Theorem 3.2, we require $d_r \to 0$ for $\widehat{\Theta}_r$ to be a consistent estimator. Here we provide an example of such d_r. Let $\Theta_0 = I_p - avv^T$, where v is a p-vector with the first

q elements being 1 and the rest being 0, and a is a positive real number. Let $a < 1/q$ to ensure that Θ_0 is positive definite. If we set $r = 0$, the closest diagonal matrix to Θ_0 is the one that contains the diagonal elements of Θ_0, thus, $d_r = a\sqrt{q(q-1)}$. If $a = o(1/q)$, we have $d_r \to 0$.

However, by definition d_r is also a lower bound for the estimation error,

$$\|\widehat{\Theta}_r - \Theta_0\|_F \quad \geq \quad d_r,$$

which means, not surprisingly, that $\widehat{\Theta}_r$ will cease to be a consistent estimator of Θ_0 if d_r is large.

3.3.4 Discussion

To summarize what we have presented so far, although the optimization problem (3.3) is straight-forward and easy to implement (see Section 3.5), it is generally not possible to specify r accurately. An inaccurate choice of r can be harmful in two ways: (1) A conservative choice of $r > r_0$ leads to slower convergence and less estimation efficiency. (2) An aggressive choice of $r < r_0$ can ruin the consistency of $\widehat{\Theta}_r$, because it can enlarge the distance between Θ_0 and the search space $\mathbf{F}_r$.

In the next section, we introduce a rank penalty to circumvent these problems. However, our main result below (Theorem 3.3) as well as the main computational algorithm (Section 3.5) are both heavily based on the results (Theorems 3.1 and 3.2) that we have obtained so far in this section.

3.4 Precision matrix estimation with rank penalty

3.4.1 The estimation method

One way to avoid having to specify the rank of the low-rank component L is by adding a penalty on the rank of L to the objective function in (3.3). That is, instead of (3.3), we

can solve the following optimization problem:

$$\begin{aligned}(\widehat{\Theta}, \widehat{L}, \widehat{D}) &= \underset{\Theta, L, D}{\arg\min} \left[\mathrm{tr}(\Theta S) - \log|\Theta| + \tau\{\mathrm{rank}(L)\}\right], \\ \text{subject to} \quad & \Theta = -L + D,\ \Theta \in \mathbf{S}_+^p,\ L \in \mathbf{S}_+^p,\ D \in \mathbf{D}^p, \end{aligned} \tag{3.4}$$

where $\tau(\cdot)$ is a monotonically increasing penalty function.

In the literature, it is popular to impose rank restrictions on a matrix by penalizing its nuclear norm. There are some advantages to directly penalizing its rank. Let $\widehat{\Theta}_r$ denote the solution to (3.3). Clearly, if we fix $\mathrm{rank}(L) = r$ in (3.4), its solution becomes $\widehat{\Theta} = \widehat{\Theta}_r$. This means $\widehat{\Theta}$ can only be one of $\{\widehat{\Theta}_r \mid r = 1, \ldots, p\}$, which will have a direct implication on how (3.4) can be solved in practice. In particular, we shall see in Section 3.5 below that, for fixed r, $\widehat{\Theta}_r$ can be obtained by a relatively efficient blockwise coordinate descent algorithm, in which the update of L given D can be achieved analytically, and the update of D given L is a relatively cheap semi-definite program.

In this section, however, we shall concentrate on the key question of how to choose the penalty function $\tau(\cdot)$ in order to ensure that $\widehat{\Theta}$ is a good estimator of Θ_0. Our answer is that it must satisfy the following two conditions:

Condition 3.3. *If $r < r_0$ and $d_r / \max(a_{n,p,r_0}, b_{n,p}) \to \infty$, then $|\tau(r) - \tau(r_0)|/d_r^2 \to 0$.*

Condition 3.4. *If $r > r_0$ and $r / \max(r_0, \log p) \to \infty$, then $a_{n,p,r}^2 / |\tau(r) - \tau(r_0)| \to 0$.*

These conditions are quite technical, and readers will find a concrete example of $\tau(\cdot)$, to be provided later in Section 3.4.3, much easier to grasp. Our main result is that, with a penalty function that satisfies Conditions 3.3 and 3.4, the solution of (3.4) will be a good estimator of Θ_0.

Theorem 3.3. *Under Conditions 3.1, 3.2, 3.3 and 3.4, if $\widehat{\Theta}$ is a solution of (3.4), then*

$$\|\widehat{\Theta} - \Theta_0\|_F = O_p\left\{\max(a_{n,p,r_0}, b_{n,p})\right\},$$

in which

$$a_{n,p,r_0} = r_0^{1/2}(p/n)^{1/2}, \quad b_{n,p} = \left\{(p \log p)/n\right\}^{1/2}.$$

Comparing the conclusion of Theorem 3.3 with that of Theorem 3.1, we can see that the convergence rates of the two methods, whether using a penalty on rank(L) or a pre-specified rank for L, are similar. The only difference is that the convergence rate of the former depends on the true rank r_0, as long as the penalty function $\tau(\cdot)$ is chosen appropriately, while the convergence rate of the latter depends on the presumed rank r.

3.4.2 Technical conditions on the penalty function

To understand Conditions 3.3 and 3.4, and how they are essential to Theorem 3.3, let us partition the set $\{r \mid r \neq r_0\}$ into four disjoint pieces:

$$\begin{aligned}
\mathbf{A}_1 &= \{r \mid r < r_0, d_r/\max(a_{n,p,r_0}, b_{n,p}) \to \infty\}, \\
\mathbf{A}_2 &= \{r \mid r < r_0, d_r = O[\max(a_{n,p,r_0}, b_{n,p})]\}, \\
\mathbf{A}_3 &= \{r \mid r > r_0, r = O[\max(r_0, \log p)]\}, \\
\mathbf{A}_4 &= \{r \mid r > r_0, r/\max(r_0, \log p) \to \infty\}.
\end{aligned}$$

Notice that, by definition, for any $r_i \in \mathbf{A}_i$ $(i = 1, 2, 3, 4)$, we have $r_1 < r_2 < r_0 < r_3 < r_4$.

Together, Theorem 3.1 and Theorem 3.2 have already established the convergence rate of $\widehat{\Theta}_r$ to be $\max(a_{n,p,r_0}, b_{n,p})$ for $r \in \mathbf{A}_2 \cup \mathbf{A}_3 \cup \{r_0\}$. A penalty function that satisfies Conditions 3.3 and 3.4 will ensure that the solution to (3.4) cannot be in the set $\{\widehat{\Theta}_r \mid r \in \mathbf{A}_1 \cup \mathbf{A}_4\}$.

Specifically, as $\|\widehat{\Theta}_r - \Theta_0\|_F \geq d_r$, any $\widehat{\Theta} \in \{\widehat{\Theta}_r \mid r \in \mathbf{A}_1\}$ cannot achieve the convergence rate given in Theorem 3.3, but Condition 3.3 ensures that such a $\widehat{\Theta}$ will not be chosen by (3.4). To see this, if $r \in \mathbf{A}_1$, we have

$$\operatorname{tr}(\widehat{\Theta}_r S) - \log|\widehat{\Theta}_r| \geq \operatorname{tr}(\widehat{\Theta}_{r_0} S) - \log|\widehat{\Theta}_{r_0}|,$$

and

$$\tau(r) < \tau(r_0).$$

The first inequality encourages the optimization problem (3.4) to favor a solution with rank(L) $= r_0$ while the second inequality encourages it to favor one with a smaller rank,

r. Condition 3.3 will ensure that $\tau(r_0) - \tau(r)$ is relatively small so that the influence from the penalty function (the second inequality above) will remain relatively weak. Likewise, by Theorem 3.1, any $\widehat{\Theta} \in \{\widehat{\Theta}_r \mid r \in \mathbf{A}_4\}$ cannot achieve the convergence rate given in Theorem 3.3, either, but Condition 3.4 will ensure that, for $r \in \mathbf{A}_4$, $\tau(r) - \tau(r_0)$ is sufficiently large so that the influence from the penalty function is strong enough to outweigh the fact that the first inequality above has now switched direction for $r \in \mathbf{A}_4$.

3.4.3 A concrete example

At this point, it will help greatly to see a concrete example of penalty functions that satisfy Conditions 3.3 and 3.4. Given n observations from a p-dimensional multivariate Gaussian model, when rank(L) in (3.4) is r, where $r \leq p$, Akaike (1987) defined the Akaike information criterion (AIC) as

$$\text{AIC}(r) \;=\; \frac{1}{n}\left[(-2)\sum_{i=1}^{n}\ell(x_i) + \{2p(r+1) - r(r-1)\}\right], \tag{3.5}$$

where $\ell(x)$ denotes the log-density function.

However, the AIC penalty has to be modified to satisfy Condition 3.3 and Condition 3.4. We let

$$\tau(r) \;=\; \delta_{n,p}\left\{2p(r+1) - r(r-1)\right\}/n, \tag{3.6}$$

in which

$$\delta_{n,p} \;\to\; \infty, \tag{3.7}$$

and

$$\delta_{n,p} \;=\; o\left\{d_r^2 n/(r_0 p)\right\} \quad \text{for all} \quad r \in \mathbf{A}_1. \tag{3.8}$$

We can see that (3.6) is essentially a scaled version of the AIC penalty. The condition (3.7) on the scaling factor $\delta_{n,p}$ means that the penalty (3.6) is larger than the AIC penalty asymptotically.

For all $r \in \mathbf{A}_1$, $d_r^2/(r_0 p/n) \to \infty$ by definition, so (3.8) does not contradict with (3.7); it is also equivalent to

$$\delta_{n,p} \;=\; o\left[\min_{r \in \mathbf{A}_1}\{d_r^2 n/(r_0 p)\}\right].$$

To verify that (3.6) satisfies Condition 3.3 and Condition 3.4, notice that

$$\tau(r) - \tau(r_0) \;=\; \delta_{n,p}(r - r_0)\{2p - (r + r_0 - 1)\}/n.$$

On the one hand, any given $r < r_0$ such that $d_r / \max(a_{n,p,r_0}, b_{n,p}) \to \infty$ is in the set $\mathbf{A}_1$ and

$$\begin{aligned} |\tau(r) - \tau(r_0)|/d_r^2 &= \delta_{n,p}(r - r_0)\{2p - (r + r_0 - 1)\}/(d_r^2 n) \\ &= o\left[(r - r_0)\{2p - (r + r_0 - 1)\}/(r_0 p)\right] \\ &= o(1), \end{aligned}$$

so Condition 3.3 is satisfied. On the other hand, any given $r > r_0$ such that $r / \max(r_0, \log p) \to \infty$ is in the set $\mathbf{A}_4$ and

$$\begin{aligned} a_{n,p,r}^2/|\tau(r) - \tau(r_0)| &= rp/[\delta_{n,p}(r - r_0)\{2p - (r + r_0 - 1)\}] \\ &= o(1), \end{aligned}$$

so Condition 3.4 is satisfied.

3.4.4 Discussion

The convergence rate given by Theorem 3.3 applies both to finite r_0 and to r_0 that may diverge to infinity with p and n. If r_0 is fixed and finite, the consistency of $\widehat{\Theta}$ is driven by $b_{n,p} = [(p \log p)/n]^{1/2}$, whose order is greater than that of a_{n,p,r_0}; otherwise, it is possible for the convergence rate to be driven by $a_{n,p,r_0} = r_0^{1/2}(p/n)^{1/2}$ — e.g., if r_0 goes to infinity faster than does $\log p$.

One can better assess our convergence rate here in the Frobenius norm by comparing it with the convergence rate of the "sparse precision matrix estimator" given by Rothman

et al. (2008). Their convergence rate in the Frobenius norm is $\{(p+s)(\log p)/n\}^{1/2}$, in which s is the number of nonzero off-diagonal entries in the population precision matrix. For fixed s, their rate becomes $\{(p \log p)/n\}^{1/2}$ and is the same as our rate ($b_{n,p}$) for fixed r_0.

That these convergence rates are of a comparable order provides another argument that the low-rank assumption can be regarded as an analogue of the sparsity assumption for estimating high-dimensional covariance/precision matrices, except that it encourages a slightly different matrix structure.

3.5 A blockwise coordinate descent algorithm

We now describe a computational algorithm for solving the optimization problem (3.4). As we have pointed out in Section 3.4, the solution to (3.4) can only be one of $\{\widehat{\Theta}_r \mid r = 0, 1, ..., p\}$. In principle, this means we can simply solve (3.3) for all $r \in \{0, 1, ..., p\}$ and choose the one that minimizes the objective function (3.4). In practice, it is usually sufficient, and not impractical, to do this only on a subset of $\{0, 1, ..., p\}$, say $\mathbb{Z}_r$.

That is, we first obtain a series of fixed-rank estimators, $\widehat{\Theta}_r$, by solving (3.3) for each $r \in \mathbb{Z}_r$. Then, we use the penalty function (3.6), given in Section 3.4.3, and evaluate the objective function (3.4) at each $\{\widehat{\Theta}_r \mid r \in \mathbb{Z}_r\}$, and the one that minimizes the objective function (3.4) is taken as the solution, $\widehat{\Theta}$. As we do not have an explicit expression for $\delta_{n,p}$, it is treated in practice as a tuning parameter and selected by minimizing the objective function on a separate, validation data set.

For each $r \in \mathbb{Z}_r$, $\widehat{\Theta}_r$ is obtained by solving the fixed-rank optimization problem (3.3) with a blockwise coordinate descent algorithm, which iteratively updates L and D (see Algorithm 1). For fixed D, we can actually solve for L analytically; this provides an enormous amount of computational saving. The validity of line 4, the analytic update of L given D, is established by Lemma B.4 in the appendices. For fixed L, we solve a log-determinant semi-definite program over D, e.g., using the SDPT3 solver (Tütüncü et al., 2003) available as part of the YALMIP toolbox (Lofberg, 2004) in Matlab; the fact that D is diagonal means the semi-definite program here is one of the cheapest kinds to solve.

To initialize the blockwise coordinate descent algorithm for each $r \in \mathbb{Z}_r$, we suggest arranging all $r \in \mathbb{Z}_r$ in ascending order and solving for each $\widehat{\Theta}_r$ sequentially, using the last solution as a "warm start" for finding the next solution. To be more specific, for $r^{(1)} < r^{(2)} < ... \in \mathbb{Z}_r$, we suggest using the diagonal component of $\widehat{\Theta}_{r^{(k-1)}}$, namely $\widehat{D}_{r^{(k-1)}}$, as the initial point ($D^{(0)}$ in Algorithm 1, line 2) for obtaining $\widehat{\Theta}_{r^{(k)}}$. To initialize the algorithm for the very first $\widehat{\Theta}_{r^{(1)}}$, we suggest using the solution of (3.3) corresponding to $r = 0$; taking $r = 0$ means there is no low-rank component, so we have an analytical solution, $D^{(0)} = \widehat{D}_0 = \text{diag}\{s_{11}^{-1}, \ldots, s_{pp}^{-1}\}$, where s_{jj} is the jth diagonal element of the sample covariance matrix S. Our experience from running many numerical experiments shows that obtaining $\widehat{\Theta}_r$ in such a sequential manner is much more efficient than obtaining each $\widehat{\Theta}_r$ independently with random "cold start" initialization.

Remark 3.3. *We think Lemma B.4, the analytic update of L given D, is a useful piece of contribution on its own. It can be used to obtain other "low-rank + something" type of decompositions of precision matrices, as the low-rank step (line 4 of the algorithm) does not depend on D being diagonal. For example, one can assume that D is a sparse matrix and the coordinate descent algorithm (Algorithm 1) can still be applied, as long as one modifies the D step (line 14) to include a sparsity penalty such as $\|D\|_1 = \sum_{i,j} |D_{ij}|$, although we generally will expect the resulting D step to become more computationally expensive than it is when D is diagonal.*

3.6 Simulation

3.6.1 Simulation settings

In this section, we compare four different estimators of the covariance/precision matrix: the sample covariance matrix (S); a simple diagonal estimator (D_S), which keeps only the diagonal elements of S and sets all off-diagonal elements to zero; the graphical lasso (Glasso) by Friedman et al. (2008); and our method (DL). The graphical lasso is implemented with the `R` package `glasso`.

Algorithm 1: Blockwise coordinate descent algorithm for solving (3.3).

$f_{old} = \infty$;
Initialize $D = D^{(0)}$;
while do
  $L = D^{1/2}UVU^TD^{1/2}$, in which
    $V = \text{diag}\{1 - 1/\max(w_1, 1), \ldots, 1 - 1/\max(w_r, 1)\}$;
    $U = [\ u_1 \ \ \ldots \ \ u_r\ ]$;
    $w_1, \ldots, w_r$ denote the r largest eigenvalues of $D^{1/2}SD^{1/2}$.;
    $u_1, \ldots, u_r$ denote the corresponding eigenvectors;
  $f_{new} = \text{tr}\{(D - L)S\} - \log|D - L|$;
  if $|f_{new} - f_{old}| < tol$ **then**
    return D, L ;
  end
  $f_{old} = f_{new}$;
  Minimize $\text{tr}\{(D - L)S\} - \log|D - L|$ over D by solving a log-determinant semi-definite program.
end

Using a training sample size of $n = 100$, we generated data from p-dimensional ($p = 50, 100, 200$) normal distributions with mean 0 and the following five population covariance matrices:

Example 1: The matrix Σ_1 is compound symmetric, $\Sigma_1 = (0.2)1_p 1_p^T + (0.8)I_p$.

Example 2: The matrix Σ_2 is "diagonal + low-rank", $\Sigma_2 = I_p + RR^T$, where $R \in \mathbf{R}^{p\times 5}$ and all of its elements are independently sampled from the Uniform$(0, 1)$ distribution.

Example 3: The matrix Σ_3 is block diagonal, consisting of 5 identical blocks $B = (0.2)1_q 1_q^T + (0.8)I_q$, where $q = p/5$.

Example 4: The matrix Σ_4 is almost "diagonal + low-rank" but with some perturbations. First, a "diagonal + low-rank" matrix is created, $B_0 = I_p + RR^T$, where $R \in \mathbf{R}^{p\times 3}$ and all of its elements are independently sampled with probability 0.8 from the Uniform$(0, 1)$ distribution and set to 0 otherwise. Next, a perturbation matrix $B_1 \in \mathbf{R}^{p\times p}$ is created, whose elements are independently sampled with probability 0.05 from the Uniform$(-0.05, 0.05)$ distribution and set to 0 otherwise. Then, the perturbation matrix B_1 is symmetrized before being combined with B_0 to obtain $B = \left\{B_0^{-1} + (B_1 + B_1^T)/2\right\}^{-1}$. Finally, we let $\Sigma_4 = B + \delta I_p$, with $\delta = |\min(\lambda_{\min}(B), 0)| + 0.05$, to ensure it is positive definite.

Example 5: The matrix Σ_5 is designed to have a sparse inverse. First, a baseline matrix $B_0 \in \mathbf{R}^{p\times p}$ is created where all of its elements are set to 0.5 with probability 0.5 and 0 otherwise. Then, it is symmetrized and made positive definite before being inverted: $B = B_0 + B_0^T$, $\delta = |\min(\lambda_{\min}(B), 0)| + 0.05$, and $\Sigma_5 = (B + \delta I_p)^{-1}$.

Each population covariance matrix in the first three examples can be decomposed into a low-rank plus a diagonal matrix. Let the decomposition be $\Sigma_k = L_{\Sigma_k} + D_{\Sigma_k}$ for $k = 1, 2, 3$; then, $L_{\Sigma_1} \in \mathbf{S}_+^{p,1}$ and $L_{\Sigma_2}, L_{\Sigma_3} \in \mathbf{S}_+^{p,5}$. Example 4 is used to test the robustness of our method; starting from a "diagonal + low-rank" matrix, we randomly perturbed

approximately 10% of the elements in the corresponding precision matrix. Example 5 is used to illustrate the performance of our method in a situation that is ideal to the graphical lasso, where the corresponding precision matrix is sparse.

Tuning parameters are selected by minimizing the negative log-likelihood function on a separate validation data set of size 100. For the graphical lasso, the tuning parameter was selected from $\{0.01, 0.03, 0.05, 0.07, 0.09,\ 0.11, 0.15, 0.20\}$. For our method, we used $\mathbb{Z}_r = \{1, 3, 5, 7, 9\}$, and the tuning parameter $\delta_{n,p}$ was selected from $\{0.6, 0.8, 1.0, 1.2, 1.4\}$. Recall from Section 3.5 that only the size of $\mathbb{Z}_r$ affects our computational time, not the number of tuning parameters we evaluate.

3.6.2 Estimation accuracy

As Rothman et al. (2008), we evaluated the estimation accuracy with the Kullback–Leibler loss,

$$L_{KL}\left(\widehat{\Theta}, \Theta_0\right) = \text{tr}\left(\Theta_0^{-1}\widehat{\Theta}\right) - \log\left|\Theta_0^{-1}\widehat{\Theta}\right| - p. \tag{3.9}$$

When $\widehat{\Theta} = \Theta_0$, the true precision matrix, the loss achieves its minimum of zero. For the graphical lasso and our method, the estimated precision matrix $\widehat{\Theta}$ could be directly plugged into the loss function (3.9); for S and D_S, the estimated covariance matrix needed to be inverted first. Thus, we could not evaluate the loss for S when $p = 100$ and $p = 200$, because it was non-invertible.

Table 3.1 reports the average Kullback–Leibler loss over 100 replications and its standard error. Not surprisingly, the sample covariance matrix S was the worst estimator; the diagonal estimator D_S was better in most cases, but not as good as the other two methods. In the first four examples, our method outperformed the graphical lasso. In Example 5, an ideal case for the graphical lasso in which the population precision matrix was sparse, our method performed slightly worse than, but still remained largely competitive against, the graphical lasso.

Table 3.1: Average (standard error) of Kullback–Leibler loss over 100 replications.

		S	D_S	Glasso	DL
Example 1	$p = 50$	37.59 (0.311)	9.058 (0.011)	2.618 (0.016)	0.980 (0.019)
	$p = 100$	NA	20.09 (0.017)	5.496 (0.029)	1.983 (0.026)
	$p = 200$	NA	42.73 (0.024)	11.39 (0.050)	3.893 (0.040)
Example 2	$p = 50$	37.44 (0.331)	36.80 (0.019)	4.148 (0.024)	2.751 (0.030)
	$p = 100$	NA	80.70 (0.043)	9.469(0.044)	5.708 (0.043)
	$p = 200$	NA	170.0 (0.071)	20.38 (0.082)	11.85 (0.060)
Example 3	$p = 50$	37.67 (0.341)	5.417 (0.011)	3.080 (0.026)	3.247 (0.038)
	$p = 100$	NA	14.40 (0.016)	7.643 (0.038)	6.103 (0.046)
	$p = 200$	NA	34.72 (0.022)	16.48 (0.074)	12.00 (0.076)
Example 4	$p = 50$	37.52 (0.316)	26.21 (0.017)	3.522 (0.023)	2.028 (0.022)
	$p = 100$	NA	33.00 (0.019)	7.534 (0.040)	3.917 (0.036)
	$p = 200$	NA	136.4 (0.062)	16.35 (0.066)	9.044 (0.057)
Example 5	$p = 50$	37.57 (0.312)	42.80 (0.020)	8.267 (0.034)	9.949 (0.046)
	$p = 100$	NA	78.15 (0.028)	22.03 (0.047)	24.13 (0.073)
	$p = 200$	NA	180.1 (0.035)	59.89 (0.096)	61.08 (0.123)

3.6.3 Rank recovery

We also investigated how well r_0 was recovered by comparing the 10 largest eigenvalues of $\widehat{L}$ with those of L_0, the low-rank component of the population precision matrix. According to (3.2), L_0 can be derived as

$$L_0 = D_{\Sigma_0}^{-1}\left(I + L_{\Sigma_0} D_{\Sigma_0}^{-1}\right)^{-1} L_{\Sigma_0} D_{\Sigma_0}^{-1}.$$

For Examples 1–3, the components L_{Σ_0} and D_{Σ_0} could be obtained directly from the set-up. For Example 4, because of the perturbation, the components were only approximate: $L_{\Sigma_4} \approx RR^T$ where $R \in \mathbf{R}^{p\times 3}$, and $D_{\Sigma_4} \approx I_p$. We skip Example 5 here because the true covariance/precision matrix does not have a corresponding low-rank component.

As the results were similar for different values of p, we only present here those for $p = 100$. In Figure 3.1, the 10 largest eigenvalues of L_0 and of $\widehat{L}$ are plotted. For $\widehat{L}$, the bigger dots in the middle are the averages over 100 replications; the smaller dots above and below are the values, (average) $\pm$ (1.96)(standard error). We can see that on average our method successfully identified the nonzero eigenvalues, or the rank, of L_0.

3.7 Real data analysis

To showcase a real application of our method to high-dimensional covariance/precision matrix estimation, we discuss the classic Markowitz portfolio selection problem (Markowitz, 1952). In this problem, we have the opportunity to invest in p assets, and the aim is to determine how much to invest in each asset so that a certain level of expected return is achieved while the overall risk is minimized. To be more specific, let μ be the mean returns of p assets and Σ, their covariance matrix. Let 1_p be the p-dimensional vector $(1, 1, ..., 1)^T$. Then, the Markowitz problem is formulated as

$$\widehat{w} = \underset{w\in\mathbf{R}^p}{\arg\min}\ w^T\Sigma w \quad \text{subject to} \quad w^T\mu = \mu_0, w^T 1_p = 1, \tag{3.10}$$

in which $\mathbf{R}^p$ is a space of p-vectors, w is a vector of asset weights, μ_0 is the desired level of expected return, and $w^T\Sigma w$ is the variance of the portfolio, which quantifies the investment risk.

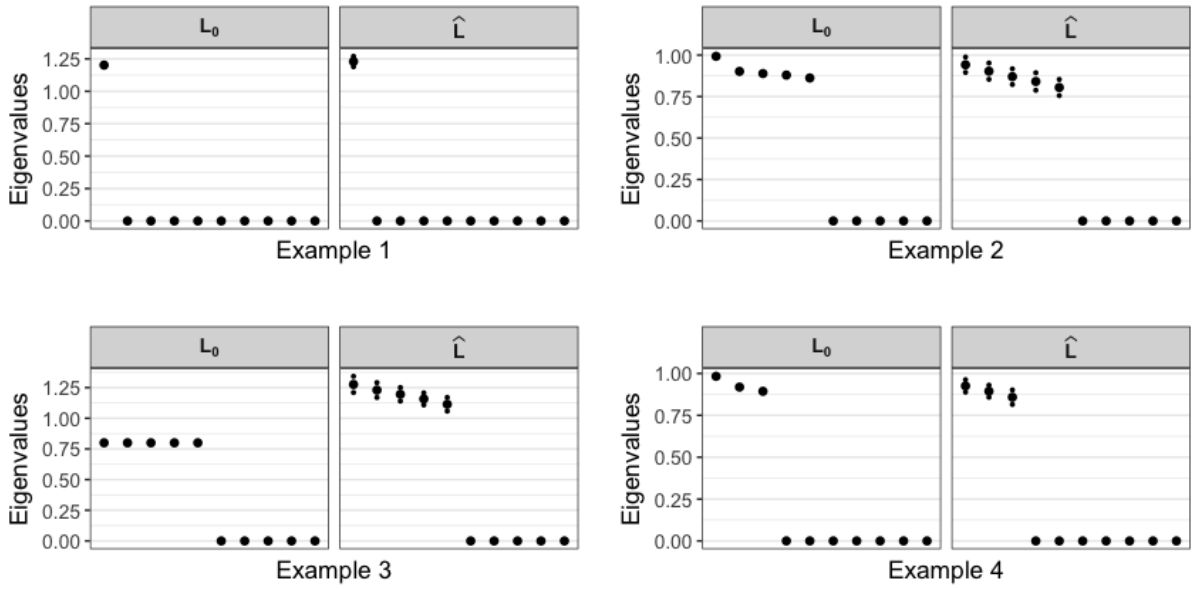

Figure 3.1: Comparison of the 10 largest eigenvalues of L_0 and those of $\widehat{L}$ [(average) $\pm$ (1.96)(standard error)].

In practice, μ and Σ can be estimated respectively by the sample mean and the sample covariance matrix before the optimization problem (3.10) is solved, provided that the sample size n is much larger than the dimension p; in high dimensions, however, solving (3.10) with the sample covariance matrix often leads to undesirable *risk underestimation* (El Karoui, 2010). Instead, different estimators of the covariance matrix can be used, such as those we have studied in the previous section: namely, the diagonal estimator (D_S), the graphical lasso (Glasso), and our method (DL).

To compare these different covariance matrix estimators for solving the Markowitz problem, we used monthly stock return data of companies in the S&P100 index from January 1990 to December 2007, as did Xue et al. (2012). This dataset contains $p = 67$ companies that remained in the S&P100 throughout this entire period; for each stock, there are $12 \times (2007 - 1990 + 1) = 216$ monthly returns.

For each month starting in January 1996, we first constructed a portfolio by solving the Markowitz problem using an estimated μ and Σ from the preceding $n = 72$ monthly returns, and a target return of $\mu_0 = 1.3\,\%$. The performance of the resulting portfolio was then measured by its return in that month. For any given estimator of Σ, a total of $12 \times (2007 - 1996 + 1) = 144$ portfolios were constructed and evaluated in this manner.

We used three-fold cross-validation to choose the tuning parameters for both the graphical lasso and our method. Each time, portfolios were constructed based on two-thirds of the training data (48 months), and the tuning parameter that maximized the average return on the remaining one-third of the training data (24 months) was selected. For the graphical lasso, the tuning parameter was selected from $\{0.2, 0.4, \ldots, 3.0\}$. For our method, we chose from the same set of tuning parameters, and the candidate ranks we considered, $\mathbb{Z}_r$, consisted of all even numbers between 2 and 28.

Table 3.2 shows the results. Again, the sample covariance matrix was noticeably outperformed by all of the other three methods. Our method (DL) was better than D_S in terms of both the average return and the overall volatility (standard error). Comparing with the graphical lasso, although our average return was slightly lower, our portfolio had much lower volatility, and hence a higher Sharpe ratio, a popular measure of overall portfolio performance in finance defined as $[\text{mean}(x - x_b)]/[\text{stdev}(x - x_b)]$, where x is the

Table 3.2: Average, standard error, and Sharpe ratio of monthly portfolio returns, January 1996 to December 2007. All numbers are expressed in %.

	S	D_S	Glasso	DL
Average	0.70	1.32	1.42	1.41
Standard Error	13.2	5.08	5.13	4.73
Sharpe ratio	5.30	26.0	27.7	29.8

portfolio's and x_b is the risk-free rate of return. For this demonstration here, we simply took $x_b = 0$ to be constant.

3.8 Conclusion

In this chapter, we have proposed a high-dimensional covariance/precision matrix estimation method that decomposes the covariance/precision matrix into a low-rank plus a diagonal matrix. This structural assumption can be understood as being driven by a factor model and as an alternative to the popular sparsity assumption to facilitate estimation in high-dimensional problems. We estimate the precision instead of the covariance matrix because the resulting negative log-likelihood function is convex and because the precision matrix can be directly applied in many statistical procedures.

Starting with a fixed-rank estimator, we have shown how it can be used to provide a more general estimator by maximizing a penalized likelihood criterion. The theoretical conditions for a valid penalty function have been studied in general, and a specific example, which is related to the Akaike information criterion, has been discussed and tested. Under these conditions, we have derived the convergence rates of the estimation error in the Frobenius norm. Numerically, we have proposed a blockwise coordinate descent algorithm that optimizes our objective function by iteratively updating the low-rank component and the diagonal component, and provided both simulated and real data examples showing that our method could have some advantages over a number of alternative estimators.

An immediate extension of our method is that it can be adapted easily to solve the latent

variable graphical model selection problem. As mentioned in Section 3.1, Chandrasekaran et al. (2012) decomposed the observed marginal precision matrix into a sparse and a low-rank component. They used the ℓ_1-norm as a penalty to encourage sparsity and the nuclear- or trace-norm as a penalty to encourage low-rank-ness. If the rank can be fixed *a priori* to be r, then we can extend our method easily to solve this problem, by removing the constraint $D \in \mathbf{D}^p$ and adding an ℓ_1-penalty $\|D\|_1$ to the objective function in (3.3) instead. If the rank r cannot be fixed, then our rank-penalized method in Section 3.4 can be extended analogously. To solve the modified optimization problem, we only need to modify Algorithm 1 slightly by adding an ℓ_1-penalty on D in Step 3 to solve for a sparse rather than diagonal component while the low-rank component is fixed.

Chapter 4

High-dimensional Covariance Matrix Estimation by a Joint Diagonal and Low-rank decomposition

4.1 Introduction

In this chapter, we seek to estimate large covariance matrices of multiple categories simultaneously. Previously, we considered matrix structures to facilitate estimation of a single covariance matrix; the key was to reduce the number of unknown parameters and improve the estimation accuracy by encouraging these structures. Now, we show that exploiting a common matrix component across categories further reduces the number of unknown parameters and allows samples from every category to contribute to the estimation of all categories.

To keep this chapter self-contained, we will briefly reiterate the relevant part of the literature review on estimating a single large covariance matrix. Then we will proceed with a review of joint estimation methods.

4.1.1 Estimation of a high-dimensional covariance matrix

Researchers have studied numerous matrix structures to facilitate the estimation of a large covariance matrix. One simple yet useful approach is to ignore the correlations and retain only the diagonal elements of S, the sample covariance matrix; a linear discriminant rule that applies such diagonal matrix is referred to as a naive Bayes classifier or an independence rule (Fan and Fan, 2008). Sparsity is a common and well-established assumption, in which the covariance matrix or its inverse is believed to have only a few non-zero off-diagonal elements. A sparse covariance matrix indicates a small number of correlated covariates, and such an estimator can be obtained by either thresholding (Cai and Liu, 2011a; Shao et al., 2011) or lasso-type regularization (Xue et al., 2012; Rothman, 2012). For Gaussian data, zeros in the inverse covariance matrix, or precision matrix, means conditional independence; imposing ℓ_1 penalization on the precision matrix can encourage such a sparse structure (Friedman et al., 2008; Rothman et al., 2008; Cai et al., 2011).

The factor model is an alternative to the sparse structure, it assumes that the overall variance can be explained by a few latent factors and some error terms. Assuming observable factors, Fan et al. (2008) proposed a covariance matrix estimator, in which the loading matrix was estimated with regression, and the covariance matrix of the error terms was estimated with a diagonal matrix. Fan et al. (2011) generalized this method by estimating the latter with a sparse instead of diagonal matrix. Chandrasekaran et al. (2012) proposed another framework, which is closely related to the factor model yet interpreted from a different perspective. They assumed that the observable variables and the latent factors are jointly Gaussian, the conditional precision matrix of the observables given the latent factors is sparse, and the number of latent factors is small. These conditions give rise to a decomposition of the marginal precision matrix of the observables into a sparse matrix and a low-rank matrix. Then, they formed a precision matrix estimator using the observed data by encouraging such a decomposition. They applied the ℓ_1 norm and the nuclear norm to recover the sparse matrix and the low-rank matrix respectively. Our decomposition of the covariance/precision matrix into a diagonal matrix and a low-rank matrix in Chapter 3 could also be interpreted with a factor model; this method features direct penalization on the rank. As it is fundamental to this work, we will discuss the detail later on.

For a more comprehensive review on estimating a single large covariance matrix, see Cai et al. (2016b). We proceed with the joint estimation of high-dimensional covariance matrices.

4.1.2 Joint estimation of high-dimensional covariance matrices

In many applications, when we work with multiple categories of data, it is reasonable to assume that different categories have category-specific characteristics while also share some common features. For example, consider a webpage dataset, which includes a category of student and a category of faculty. On the one hand, term frequencies of a student webpage could be uniquely related to academic information and job seeking, while term frequencies of a faculty webpage might be related to research interests and professional activities, on the other hand, these two categories could both be related to teaching and studying. Another example is gene expression data. Across categories such as normal and cancerous or various subtypes of the same disease, distributions of gene expression levels could differ through some pathways while be similarly related to others. In these cases, jointly estimating multiple covariance matrices could outperform either estimating the same covariance matrix for all categories or estimating multiple covariance matrices completely independently.

Researchers have proposed some methods with the merit of joint estimation. The focus has been on the existence of both shared and non-shared links in graphical models. Guo et al. (2011) reparameterized each off-diagonal element of a precision matrix as the product of a common parameter and a category-specific parameter; then, they imposed lasso-type penalties on the common parameters to encourage universal zero entries and on the category-specific parameters to encourage zeros for associated categories. Danaher et al. (2014) considered the fused graphical lasso (FGL) and the group graphical lasso (GGL). Both methods apply the conventional ℓ_1 penalty to all precision matrices so that they have sparse patterns. Regarding shared matrix structures, the FGL penalizes differences between precision matrices and encourages not only similar network structures but also similar edge values; the GGL applies a group lasso penalty to elements in the same position of different precision matrices and simultaneously encourages them to be zero. Cai et al.

(2016a) proposed to minimize the maximum of the ℓ_1 norms of the precision matrices, subject to a constraint that encourages a common sparse pattern. Without applying the likelihood, this method does not require independence among the random vectors across categories.

4.1.3 Outline and summary of this chapter

In this chapter, we estimate high-dimensional covariance/precision matrices of multiple categories, by considering an innovative method of joint estimation. In the estimation, each covariance/precision matrix is encouraged to decompose into a diagonal matrix, a low-rank matrix, both shared across categories, and a category-specific low-rank matrix. This decomposition can be interpreted under the framework of factor models. Just as graphical models can share network structures, when data are believed to be affected by latent factors, it is a reasonable assumption that the effects of some factors are common across categories while those of the other factors are specific to one of these categories.

In Section 4.2, we firstly summarize the decomposition assumption for a single covariance matrix, and then discuss the proposed joint decomposition assumption in detail. In Section 4.3, we consider pre-selected matrix ranks and study properties of associated estimators. In Section 4.4, an AIC-type penalty is imposed to encourage the decomposition and automatically select the matrix ranks. Some consistency properties of the estimators are developed under conditions on the population covariance matrices and the penalty function. An algorithm for obtaining the estimators is introduced in Section 4.5. In Section 4.6, we experiment a variety of matrix setups and show nice performances of the proposed estimators. In Section 4.7, through real data analysis, we demonstrate how the latent factors can be identified with the estimated low-rank matrices and how quadratic discriminant analysis can apply the proposed estimators. All proofs are relegated to the appendices.

4.1.4 Notations

Before proceeding with the methodology, we introduce a few notations. We let $\mathbf{R}^{p_1 \times p_2}$ denote the set of $p_1 \times p_2$ matrices and $\mathbf{R}^p$ denote the set of p-vectors. We use $\mathbf{S}^p$ denote the set of symmetric matrices in $\mathbf{R}^{p \times p}$, $\mathbf{S}^p_+$ denote the set of positive semi-definite matrices in $\mathbf{S}^p$, and $\mathbf{S}^p_{++}$ denote the set of strictly positive definite matrices in $\mathbf{S}^p$. Whenever necessary, another superscript r is added to indicate a subset of matrices with rank $\leq r$, e.g., $\mathbf{S}^{p,r}$, $\mathbf{S}^{p,r}_+$ and $\mathbf{S}^{p,r}_{++}$. In a similar manner, $\mathbf{D}^p$, $\mathbf{D}^p_+$, and $\mathbf{D}^p_{++}$ denote sets of diagonal matrices with real, non-negative and positive diagonal elements, respectively.

For $A \in \mathbf{S}^p$, we let $\mathrm{tr}(A)$ denote its trace, $|A|$ denote its determinant, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote its largest and smallest eigenvalues respectively. For matrix norms, we let $\|A\|_F = \{\mathrm{tr}(A^T A)\}^{1/2}$ denote its Frobenius norm, $\|A\|_* = \mathrm{tr}\{(A^T A)^{1/2}\}$ denote its nuclear norm, and $\|A\|_{op} = \{\lambda_{\max}(AA^T)\}^{1/2}$ denote its operator norm. For a vector a, we use $\|a\|_2$ to denote its ℓ_2 norm.

We use I_p to denote the $p \times p$ identity matrix. The function $\mathrm{diag}(\cdot)$, depending on the type of the input, either converts a vector to a diagonal matrix by setting its diagonal elements to be the input vector, or converts a matrix to a vector by extracting the diagonal elements.

4.2 Problem set-up and model assumption

4.2.1 The "diagonal + low-rank" decomposition

To begin with, we briefly summarize the method in Chapter 3 and lay a foundation for the upcoming new method.

Previously, we studied high-dimensional covariance matrix estimation by limiting our discussion to data drawn from one distribution, e.g., $N(0, \Sigma_0^\dagger)$, and assuming $\Sigma_0^\dagger = D_{\Sigma_0}^\dagger + L_{\Sigma_0}^\dagger$, in which $D_{\Sigma_0}^\dagger \in \mathbf{D}^p_{++}$ and $L_{\Sigma_0}^\dagger \in \mathbf{S}_+^{p,r_0^\dagger}$. This assumption can be related to the factor model, so that $r_0^\dagger$ is the number of latent factors, $L_{\Sigma_0}^\dagger$ is the variance explained by the latent

factors, and $D^{\dagger}_{\Sigma_0}$ is the variance explained by error terms. The idea was to encourage such a decomposition and reduce the number of unknown parameters while estimating $\Sigma^{\dagger}_0$.

The decomposition of $\Sigma^{\dagger}_0$ is equivalent to the decomposition of its inverse, i.e., $\Theta^{\dagger}_0 = D^{\dagger}_0 - L^{\dagger}_0$, where $D^{\dagger}_0 \in \mathbf{D}^p_{++}$ and $L^{\dagger}_0 \in \mathbf{S}^{p,r^{\dagger}_0}_{+}$. Therefore, we aimed at $\Theta^{\dagger}_0$, the precision matrix, and obtained its estimator $\widehat{\Theta}^{\dagger}$ by solving

$$\begin{aligned} \min_{\Theta} \quad & \operatorname{tr}(\Theta S) - \log|\Theta| + \tau\{\operatorname{rank}(L)\}, \\ \text{subject to} \quad & -L + D = \Theta,\ \Theta \in \mathbf{S}^p_+,\ L \in \mathbf{S}^p_+,\ D \in \mathbf{D}^p, \end{aligned} \tag{4.1}$$

where the objective function is the negative log-likelihood of the normal distribution plus a monotonically increasing penalty function of the rank, $\tau(\cdot)$. An example of $\tau(\cdot)$, that ensures nice properties of $\widehat{\Theta}^{\dagger}$, is a scaled Akaike information criterion (AIC) penalty.

In the current work, we consider an extension of the "diagonal + low-rank" condition, so that the low-rank component is further explored in the context of multiple categories.

4.2.2 The "joint diagonal + low-rank" decomposition

Consider a heterogeneous dataset $X = (X^{(1)}, \ldots, X^{(K)})$, in which $X^{(k)} = (x^{(k)}_1, \ldots, x^{(k)}_{n_k})$ contains n_k samples identically drawn from the p-variate Gaussian distribution $N(0, \Sigma^{(k)}_0)$. We also assume that all random samples in X are independently drawn. The observations in each of the K categories are assumed to be centered without loss of generality. Let $n = \sum_{k=1}^{K} n_k$.

In the situation of multiple categories, instead of assuming the previous "diagonal + low-rank" decomposition for each category, we propose a "joint diagonal + low-rank" decomposition,

$$\Sigma^{(k)}_0 = D_{\Sigma_0} + L_{\Sigma_0} + L^{(k)}_{\Sigma_0} \quad (k = 1, \ldots, K), \tag{4.2}$$

in which $D_{\Sigma_0} \in \mathbf{D}^p_{++}$, L_{Σ_0} and $L^{(k)}_{\Sigma_0}$'s are positive semi-definite, $\operatorname{rank}(L_{\Sigma_0}) = r_0$ and $\operatorname{rank}(L^{(k)}_{\Sigma_0}) = (v_{0k} - r_0)$. That $r_0 \leq v_{0k}$ is implicit. For the purpose of brevity, a vector will be used to represent v_{0k}'s whenever applicable: $v_0 = (v_{01}, \ldots, v_{0K})$.

We can see that D_{Σ_0} and L_{Σ_0} are shared across categories, while $L^{(k)}_{\Sigma_0}$'s are category-specific. Furthermore, v_{0k}'s represent to what degree a covariance matrix of a certain category conforms to the "diagonal + low-rank" structure, and r_0 decides to what extent a joint matrix structure might be exploited.

To eliminate the ambiguity of v_0 and r_0, we formalize their definitions. We first let $v_{0k} = \text{rank}(L^{(k)}_*)$, in which

$$\begin{aligned} L^{(k)}_* &= \arg\min_L \ \text{rank}(L) \\ \text{subject to} \quad & D + L = \Sigma^{(k)}_0, \ D \in \mathbf{D}^p_{++}, \ L \in \mathbf{S}^p_+, \end{aligned} \tag{4.3}$$

and then we let $r_0 = \text{rank}(L_{\Sigma_0})$, in which

$$\begin{aligned} L_{\Sigma_0} &= \arg\max_L \ \text{rank}(L) \\ \text{subject to} \quad & L + L^{(k)} = L^{(k)}_*, \\ & \text{rank}(L) + \text{rank}(L^{(k)}) = v_{0k}, \\ & L \in \mathbf{S}^p_+, \ L^{(k)} \in \mathbf{S}^p_+. \end{aligned} \tag{4.4}$$

Definition (4.3) means, v_{0k}'s are the smallest attainable ranks through separate diagonal and low-rank matrix decompositions. Definition (4.4) means, after v_{0k}'s and $L^{(k)}_*$'s are defined, r_0 is the largest rank of the common low-rank component that might be isolated from $L^{(k)}_*$'s. When defining r_0, we force the summation of $\text{rank}(L)$ and $\text{rank}(L^{(k)})$ to equal v_{0k}, so that the definition aligns with our algorithm. For the purpose of efficiency, in the algorithm (Section 4.5), we will first find and fix v_{0k}'s; then, we will seek r_0 by searching through $0 \le r \le \min_k v_{0k}$ — for each r, L and $L^{(k)}$ are estimated with ranks restricted to be r and $v_{0k} - r$ respectively.

Although the decomposition (4.2) is trivially possible for any set of positive definite matrices if there is no restriction on v_0 or r_0, we will concentrate on the situation when the decomposition is most useful for matrix estimation. That is when v_{0k}'s are small and r_0 is relatively large — small v_{0k}'s and relatively large r_0 lead to a small number of unknown parameters to be estimated .

The decomposition (4.2) could be understood under the factor model framework. If the random vectors can be explained by latent factors and error terms as follows:

$$x_i^{(k)} = \Gamma f_i^{(k)} + \Gamma^{(k)} g_i^{(k)} + \epsilon_i^{(k)},\ k = 1, \ldots, K,\ i = 1, \ldots, n_k, \tag{4.5}$$

in which $\Gamma \in \mathbf{R}^{p \times r_0}$, $\Gamma^{(k)} \in \mathbf{R}^{p \times (v_{0k} - r_0)}$, $f_i^{(k)} \sim N(0, I_{r_0})$, $g_i^{(k)} \sim N(0, I_{v_{0k} - r_0})$, $\epsilon_i^{(k)} \sim N(0, D)$, $D \in \mathbf{D}_{++}^p$, and $f_i^{(k)}$, $g_i^{(k)}$ and $\epsilon_i^{(k)}$ are independent, (4.2) is immediately true. The factor model (4.5) suggests that all random vectors are affected by f factors through the common loading Γ, the random vectors in category k are affected by g factors through a category-specific loading $\Gamma^{(k)}$, and all error terms distribute identically. In spite of the close connection with factor models, our main concern is still estimating $\Sigma_0^{(k)}$'s; loading matrix estimation and factor interpretation are considered by-products and will be demonstrated only in real data analysis (Section 4.7).

In our previous work, when we imposed the "diagonal + low-rank" structure on a covariance matrix, we compared it to the sparse matrix structure. We pointed out that (i) the "diagonal+low-rank" structure, just like the sparse structure, reduces the number of parameters to be estimated; (ii) the rank, or the number of non-zero eigenvalues, of the low-rank component is analogous to the number of non-zero off-diagonal elements of the covariance/precision matrix under the sparse assumption — both indicate to what extent the assumed structures can simplify the matrices. Similarly, the "joint diagonal + low-rank" could be compared to joint sparse structures. The category-wise comparison is the same as comparing the "diagonal+low-rank" structure to the sparse structure. To see the comparison from the perspective of "joint", let us use the fused graphical lasso (Danaher et al., 2014) as an example. The "joint diagonal + low-rank" decomposition encourages a shared low-rank component and further reduces the overall number of unknown parameters, just like the fused graphical lasso, which encourages elements in precision matrices to be identical across categories and simplifies the overall matrix structure.

4.3 Precision matrix estimation with fixed ranks

Now we shift our focus to the structure of the precision matrices under the aforementioned decomposition assumption, and then propose a joint estimation method on top of that.

Let $\Theta_0^{(k)} = (\Sigma_0^{(k)})^{-1}$ be the precision matrix of category k. Just as decomposing the covariance matrix and its inverse are equivalent in the "diagonal+low-rank" case, the equivalence in the "joint diagonal+low-rank" case can be established (Henderson and Searle, 1981):

$$\begin{aligned}
\Theta_0^{(k)} &= (D_{\Sigma_0} + L_{\Sigma_0} + L_{\Sigma_0}^{(k)})^{-1} \\
&= -(D_{\Sigma_0} + L_{\Sigma_0})^{-1} \left\{ I_p + L_{\Sigma_0}^{(k)} (D_{\Sigma_0} + L_{\Sigma_0})^{-1} \right\}^{-1} L_{\Sigma_0}^{(k)} (D_{\Sigma_0} + L_{\Sigma_0})^{-1} + (D_{\Sigma_0} + L_{\Sigma_0})^{-1} \\
&= -(D_{\Sigma_0} + L_{\Sigma_0})^{-1} \left\{ I_p + L_{\Sigma_0}^{(k)} (D_{\Sigma_0} + L_{\Sigma_0})^{-1} \right\}^{-1} L_{\Sigma_0}^{(k)} (D_{\Sigma_0} + L_{\Sigma_0})^{-1} \\
&\quad -D_{\Sigma_0}^{-1} \left(I_p + L_{\Sigma_0} D_{\Sigma_0}^{-1} \right)^{-1} L_{\Sigma_0} D_{\Sigma_0}^{-1} + D_{\Sigma_0}^{-1} \\
&\triangleq -L_0^{(k)} - L_0 + D_0, \qquad (4.6)
\end{aligned}$$

in which $D_0 \in \mathbf{D}_{++}^p$, L_0 and $L_0^{(k)}$'s are positive semi-definite, $\text{rank}(L_0) = r_0$, and $\text{rank}(L_0^{(k)}) = (v_{0k} - r_0)$. We henceforth work with the precision matrices, since they are also the natural optimization variables in the likelihood maximization. We write the population precision matrices as $\Theta_0 = (\Theta_0^{(1)}, \ldots, \Theta_0^{(K)})$. This "list of matrices" notation is used to simplify the discussion of multiple categories.

Apart from the "joint diagonal + low-rank" structure, we also make the common conditions of "bounded eigenvalues" and "comparable sample sizes". Therefore, the conditions about the population covariance/precision matrices are as follows:

Condition 4.1. *There exist constants $c_1, c_2 > 0$ such that $c_1 \leq \lambda_{min}(\Sigma_0^{(k)}) \leq \lambda_{\max}(\Sigma_0^{(k)}) \leq c_2$, or equivalently, $c_2^{-1} \leq \lambda_{min}(\Theta_0^{(k)}) \leq \lambda_{\max}(\Theta_0^{(k)}) \leq c_1^{-1}$, uniformly with respect to p.*

Condition 4.2. *All sample sizes are of the same order, i.e., $n_k \asymp n_{k'}$, $k, k' = 1, \ldots, K$.*

Condition 4.3. *For some $v_{0k} = o(p)$ and $r_0 \leq v_{0k}$, $\Sigma_0^{(k)}$ can be decomposed as $\Sigma_0^{(k)} = D_{\Sigma_0} + L_{\Sigma_0} + L_{\Sigma_0}^{(k)}$, where $D_{\Sigma_0} \in \mathbf{D}_{++}^p$, $L_{\Sigma_0} \in \mathbf{S}_+^p$, $L_{\Sigma_0}^{(k)} \in \mathbf{S}_+^p$, $\text{rank}(L_{\Sigma_0}) = r_0$ and $\text{rank}(L_{\Sigma_0}^{(k)}) = (v_{0k} - r_0)$; or equivalently, $\Theta_0^{(k)}$ can be decomposed as $\Theta_0^{(k)} = D_0 - L_0 - L_0^{(k)}$, where $D_0 \in \mathbf{D}_{++}^p$, $L_0 \in \mathbf{S}_+^p$, $L_0^{(k)} \in \mathbf{S}_+^p$, $\text{rank}(L_0) = r_0$ and $\text{rank}(L_0^{(k)}) = (v_{0k} - r_0)$.*

Being ready to estimate Θ_0, we firstly consider the simple situation when the ranks of L_0 and $L_0^{(k)}$ are pre-selected as r and $(v_k - r)$ respectively. In this case, the precision

matrix estimators can be obtained by solving

$$\begin{aligned} \min_{\Theta} \quad & \sum_{k=1}^{K} n_k \left\{ \mathrm{tr}(\Theta^{(k)} S^{(k)}) - \log |\Theta^{(k)}| \right\}, \\ \text{subject to} \quad & \Theta^{(k)} = D - L - L^{(k)}, \\ & \Theta^{(k)} \in \mathbf{S}_+^p, \ D \in \mathbf{D}^p, \\ & L \in \mathbf{S}_+^{p,r}, \ L^{(k)} \in \mathbf{S}_+^{p,(v_k - r)}, \end{aligned} \tag{4.7}$$

where $S^{(k)}$ is the sample covariance matrix of category k, and the objective function is the negative log-likelihood of multiple independent normal distributions. In the following discussion, we use $\Theta = (\Theta^{(1)}, \ldots, \Theta^{(K)})$ to denote the optimization variable and $\widehat{\Theta}_{r,v} = (\widehat{\Theta}_{r,v}^{(1)}, \ldots, \widehat{\Theta}_{r,v}^{(K)})$ to denote the fixed-rank solution. In the subscript of $\widehat{\Theta}_{r,v}$, $v = (v_1, \ldots, v_K)$ represents a vector of ranks.

To facilitate the forthcoming discussion of the estimation accuracy, we set up a few concepts. We let

$$\begin{aligned} \mathbf{F}_{r,v} \ = \ & \{\Theta = (\Theta^{(1)}, \ldots, \Theta^{(K)}) \mid \Theta^{(k)} = D - L - L^{(k)}, \\ & D \in \mathbf{D}_{++}^p, \ L \in \mathbf{S}_+^{p,r}, \ L^{(k)} \in \mathbf{S}_+^{p,(v_k - r)} \text{ for all } k\} \end{aligned}$$

denote the search space of (4.7), and define the distance between Θ_0 and $\mathbf{F}_{r,v}$ as

$$d_{r,v} = \min_{\Theta \in \mathbf{F}_{r,v}} \sum_{k=1}^{K} \|\Theta^{(k)} - \Theta_0^{(k)}\|_F.$$

We also let $\Theta_{r,v} = (\Theta_{r,v}^{(1)}, \ldots, \Theta_{r,v}^{(K)}) \in \mathbf{F}_{r,v}$ be the element closest to Θ_0 in the search space, i.e., $\sum_{k=1}^{K} \|\Theta_{r,v}^{(k)} - \Theta_0^{(k)}\|_F = d_{r,v}$.

To gain some intuition about $d_{r,v}$, we consider two cases. These two cases do not cover the whole picture, but the remaining cases can be discussed in the same manner.

When $v_k \geq v_{0k}$ (for all $k = 1, \ldots, K$) and $r \leq r_0$, we have $\Theta_0 \in \mathbf{F}_{r,v}$ and $d_{r,v} = 0$. To see this, we eigen-decompose L_0 as $L_0 = U \,\mathrm{diag}(\lambda_1, \ldots, \lambda_{r_0}, 0, \ldots, 0) U^T$, and simply let

$$\begin{aligned} D \ &= \ D_0, \\ L \ &= \ U \,\mathrm{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0) U^T, \\ L^{(k)} \ &= \ U \,\mathrm{diag}(0, \ldots, 0, \lambda_{r+1}, \ldots, \lambda_{r_0}, 0, \ldots, 0) U^T + L_0^{(k)}; \end{aligned}$$

then, $\Theta_0^{(k)} = D - L - L^{(k)}$ and all constraints in $\mathbf{F}_{r,v}$ are satisfied.

On the other hand, when $v_k < v_{0k}$ (for all $k = 1, \ldots, K$) and $r > r_0$, Θ_0 is not in the search space anymore, and we have $d_{r,v} > 0$. To find an upper bound for $d_{r,v}$, we construct a list of matrices that belongs to $\mathbf{F}_{r,v}$. In addition to the previous eigen-decomposition of L_0, we eigen-decompose $L_0^{(k)}$ as $L_0^{(k)} = U^{(k)} \operatorname{diag}(\lambda_1^{(k)}, \ldots, \lambda_{v_{0k}-r_0}^{(k)}, 0, \ldots, 0)(U^{(k)})^T$, in which the eigenvalues are in descending order. Then, let

$$\begin{aligned} D &= D_0, \\ L &= L_0, \\ L^{(k)} &= U^{(k)} \operatorname{diag}(\lambda_1^{(k)}, \ldots, \lambda_{v_k-r}^{(k)}, 0, \ldots, 0)(U^{(k)})^T. \end{aligned}$$

If $\Theta^{(k)} = D - L - L^{(k)}$, we have $\Theta \in \mathbf{F}_{r,v}$. Thus, the distance has an upper bound

$$\begin{aligned} d_{r,v} &\leq \sum_{k=1}^{K} \|\Theta^{(k)} - \Theta_0^{(k)}\|_F \\ &= \sum_{k=1}^{K} \|(D - D_0) - (L - L_0) - (L^{(k)} - L_0^{(k)})\|_F \\ &= \sum_{k=1}^{K} \|(\lambda_{v_k-r+1}^{(k)}, \ldots, \lambda_{v_{0k}-r_0}^{(k)})^T\|_2. \end{aligned} \tag{4.8}$$

On the right-hand-side of (4.8) are the smallest $\{(v_{0k} - v_k) + (r - r_0)\}$ (i.e., the total number of "misspecified" ranks for category k) eigenvalues of $L_0^{(k)}$; if they are small, we can anticipate small $d_{r,v}$.

Now, we establish consistency properties of $\widehat{\Theta}_{r,v}$, under the condition that $d_{r,v}$ is relatively small.

Theorem 4.1. *Suppose conditions 4.1, 4.2 and 4.3 hold and the ranks r and v are prespecified so that $d_{r,v} = O\{\max(a_{n,p,v}, b_{n,p})\}$, then the solution of (4.7), $\widehat{\Theta}_{r,v}$, has the property:*

$$\sum_{k=1}^{K} \|\widehat{\Theta}_{r,v}^{(k)} - \Theta_0^{(k)}\|_F = O_p\left\{\max(a_{n,p,v}, b_{n,p})\right\},$$

in which

$$a_{n,p,v} = (\max_k v_k)^{1/2}(p/n)^{1/2}, \quad b_{n,p} = \{(p \log p)/n\}^{1/2}.$$

Corollary 4.1. *Suppose conditions 4.1, 4.2 and 4.3 hold and the ranks r and v are pre-specified so that $d_{r,v} = O\{\max(a_{n,p,v_0}, b_{n,p})\}$ and $\max_k v_k = O\{\max(\max_k v_{0k}, \log p)\}$, then the solution of (4.7), $\widehat{\Theta}_{r,v}$, has the property:*

$$\sum_{k=1}^{K} \|\widehat{\Theta}_{r,v}^{(k)} - \Theta_0^{(k)}\|_F = O_p\left\{\max(a_{n,p,v_0}, b_{n,p})\right\},$$

in which

$$a_{n,p,v_0} = (\max_k v_{0k})^{1/2}(p/n)^{1/2}, \quad b_{n,p} = \{(p\log p)/n\}^{1/2}.$$

Remark 4.1. *Theorem 4.1 and and Corollary 4.1 contain three situations: (i) $v_k \geq v_{0k}$ $(k = 1, \ldots, K)$ and $r \leq r_0$; (ii) $v_k < v_{0k}$ $(k = 1, \ldots, K)$ and $r > r_0$, but the distance from Θ_0 to $\mathbf{F}_{r,v}$ is reasonably small; (iii) the remaining combinations of v and r, and $d_{r,v}$ is small.*

In situation (i), the conditions on $d_{r,v}$ in both Theorem 4.1 and Corollary 4.1 hold trivially. Theorem 4.1 suggests, when the convergence rate is determined by v_k's, it gets worse as $\max_k v_k$ gets larger. This aligns with the intuition because, as long as $\Theta_0 \in \mathbf{F}_{r,v}$ already, larger v_k's introduce extra unnecessary parameters and lead to larger estimation error. Corollary 4.1 suggests, even if $\max_k v_k > \max_k v_{0k}$, as long as it is not too large in terms of the order, the convergence rate is as if v_k's are chosen correctly.

Situation (ii) contains two interesting facts. Firstly, Let us take the true ranks (i.e., r_0 and v_0) as the benchmark case, under which we have the convergence rate $\max(a_{n,p,v_0}, b_{n,p})$. Corollary 4.1 states that when $d_{r,v}$ does not exceed the estimation error of the benchmark case, the error caused by choosing inaccurate ranks is dominated and $\widehat{\Theta}_{r,v}$ is asymptotically no worse than $\widehat{\Theta}_{r_0,v_0}$. Secondly, according to Theorem 4.1, presuming smaller v_k $(v_k < v_{0k})$ might be beneficial — when v_k's are small, so is $a_{n,p,v}$. To be more specific, when $v_k < v_{0k}$, the estimation error could be lower than that of the benchmark case as long as $d_{r,v}$ does not exceed this error. The intuition is, if the advantage of estimating fewer parameters overweighs the disadvantage of Θ_0 not being in $\mathbf{F}_{r,v}$, we may as well just use smaller v_k's. It is worth noticing that this intuition also applies to $r > r_0$ even if it is not shown through the convergence rate. The upper bound in (4.8) suggests that the scenario being discussed could happen when there are some rapidly degenerating eigenvalues in the population low-rank components.

Situation (iii) is a mix of (i) and (ii); for every instance in situation (iii), v and r satisfy some inequalities in (i) and some in (ii). This situation could suffer from either extra parameters or $\Theta_0 \notin \mathbf{F}_{r,v}$. The dominant rank in v, which determines the convergence rate, is not necessarily an overstated one (i.e., $v_k > v_{0k}$) if there is any; for example, even $v_k \geq v_{0k}$ and $v_{k'} < v_{0k'}$ for $k' \neq k$, it is possible that $v_{k'} > v_k$. Therefore, the discussion of (i) applies when the dominant rank is one of the overstated ranks, and that of (ii) applies when an understated rank dominates.

Although the estimators do enjoy good properties when we are able to well-specify the ranks, sometimes there is no prior information for us to do so. Ill-specified ranks could lead to large error, caused by either too many unnecessary parameters or $\mathbf{F}_{r,v}$ being too far from Θ_0. For the latter, if $d_{r,v}$ diverges, so must the estimation error as $\sum_{k=1}^{K} \|\widehat{\Theta}_{r,v}^{(k)} - \Theta_0^{(k)}\|_F \geq d_{r,v}$ by the definition of $d_{r,v}$. To avoid these unpleasant situations, we consider applying a rank penalty in the next section.

4.4 Precision matrix estimation with rank penalty

Let us consider the AIC penalty for fixed p,

$$\tau(r, v) = \sum_{k=1}^{K} \{2p(v_k - r) - (v_k - r)(v_k - r - 1)\} + \{2p(r+1) - r(r-1)\},$$

which is the number of unknown parameters in (4.7). Some simple calculus shows that it increases with v_k and decreases with r when $K > 1$ and $p \geq (\sum_{k=1}^{K} v_k)/(K-1)$; the former is guaranteed while dealing with multiple categories, and the latter is true given $v_{0k} = o(p)$ and we do not consider large v_k in practice. This suggests that the AIC penalty aligns with our goal of encouraging small v_k's and relatively large r.

We modify the AIC penalty by considering an additional tuning parameter λ and

propose the following penalized optimization problem,

$$\begin{aligned} \min_{\Theta} \quad & \sum_{k=1}^{K} n_k \left\{ \mathrm{tr}(\Theta^{(k)} S^{(k)}) - \log |\Theta^{(k)}| \right\} + \lambda \tau(r, v), \\ \text{subject to} \quad & \Theta^{(k)} = D - L - L^{(k)}, \\ & \Theta^{(k)} \in \mathbf{S}_+^p, \ D \in \mathbf{D}^p, \\ & L \in \mathbf{S}_+^p, \ L^{(k)} \in \mathbf{S}_+^p. \end{aligned} \tag{4.9}$$

Let $\widehat{\Theta} = (\widehat{\Theta}^{(1)}, \ldots, \widehat{\Theta}^{(K)})$ be the solution. If the ranks are fixed, so is $\tau(r, v)$, and (4.9) reduces to (4.7); therefore, $\widehat{\Theta}$ must be in

$$\left\{ \widehat{\Theta}_{r,v} \mid v_k = 1, \ldots, p, \ r = 1, \ldots, \min_k v_k \right\},$$

which is the set of solutions of (4.7) under various fixed ranks.

Now we think about what makes a good penalty and show that the proposed penalty qualifies. Recall that we aim to prevent too large $\max_k v_k$ and $d_{r,v}$; to formalize these scenarios to be avoided, we define the corresponding sets of ranks as

$$\begin{aligned} \mathbf{A}_1 &= \left\{ (r, v) \mid (\max_k v_k) / \max(\max_k v_{0k}, \log p) \to \infty \right\}, \\ \mathbf{A}_2 &= \left\{ (r, v) \mid d_{r,v} / \max(a_{n,p,v_0}, b_{n,p}) \to \infty \right\}. \end{aligned}$$

Corollary 4.1 states that, when $(r, v) \in \mathbf{A}_1^c \cap \mathbf{A}_2^c$, the convergence rate is as if r and v are set to be the true ranks. Thus, if $\tau(r, v)$ and λ together can adjust the objective function so that an element of $\mathbf{A}_1 \cup \mathbf{A}_2$ is never chosen over (r_0, v_0), we will be able to guarantee that $\widehat{\Theta}$ has the same convergence rate as $\widehat{\Theta}_{r_0,v_0}$. As a matter of fact, this is the case for the proposed penalty, if we have the following additional condition about λ,

Condition 4.4. *$\lambda \to \infty$ and $\lambda = o(\delta_{n,p})$, where*

$$\delta_{n,p} = \min_{(r,v) \in \mathbf{A}_2 \cap \mathbf{A}_1^c} (d_{r,v})^2 / a_{n,p,v_0}^2.$$

Theorem 4.2. *Suppose conditions 4.1 – 4.4 hold, then the solution of (4.9), $\widehat{\Theta}$, has the property:*

$$\sum_{k=1}^{K} \| \widehat{\Theta}^{(k)} - \Theta_0^{(k)} \|_F = O_p \left\{ \max(a_{n,p,v_0}, b_{n,p}) \right\},$$

in which

$$a_{n,p,v_0} = (\max_k v_{0k})^{1/2}(p/n)^{1/2}, \quad b_{n,p} = \{(p\log p)/n\}^{1/2}.$$

Remark 4.2. *Condition 4.4 says λ has to be sufficiently large to approach the infinity while not too large to exceed the rate of $\delta_{n,p}$. To understand this, we consider $\mathbf{A}_1$ and $\mathbf{A}_2 \cap \mathbf{A}_1^c$ separately. On the one hand, to exclude $\mathbf{A}_1$, we want large $\lambda\tau(r,v)$ to penalize the overstated ranks (i.e.,$v_k > v_{0k}$); on the other hand, we want to avoid overly large $\lambda\tau(r,v)$, which might lead to $\mathbf{A}_2 \cap \mathbf{A}_1^c$, where the convergence is ruined by either understated v_k or overstated r (too few parameters and $\Theta_0 \notin \mathbf{F}_{r,v}$).*

4.5 Algorithm

Now we consider how to solve the optimization problem (4.9). We have mentioned in Section 4.4 that $\widehat{\Theta}$ can only be one of $\{\widehat{\Theta}_{r,v} \mid v_k = 0,\ldots,p, r = 0,\ldots,\min_k v_k\}$; thus, a straightforward solution is to obtain every $\widehat{\Theta}_{r,v}$ and identify the one that minimizes the objective function in (4.9). However, as we discussed in Section 4.2.2, the decomposition method is most useful for estimating large covariance/precision matrices when v_{0k}'s are small; therefore, we propose to consider only a subset $\{\widehat{\Theta}_{r,v} \mid v_k \in \mathbb{Z}^{(k)}, r = 0,\ldots,\min_k v_k\}$, in which $\mathbb{Z}^{(k)} \subset \{0,\ldots,p\}$ includes small ranks and is chosen by the user in practice. Considering subsets is also more computationally efficient.

Furthermore, we first determine v_k's by applying the "diagonal+low-rank" to each category separately, and then fix v and obtain $\widehat{\Theta}_{r,v}$ for various r. In this way, we avoid considering the combinations of possible values of v_k's as well as the combinations of possible values of r with unfavorable values of v, thus dramatically reduce the computational cost. To be more specific, we first solve (4.1) for each category, and the ranks $v_k \in \mathbb{Z}^{(k)}$ associated with the solutions are selected. Then we solve (4.7) for each r and obtain a series of fixed-rank estimators $\{\widehat{\Theta}_{r,v} \mid r \le \min_k v_k\}$. At last, over this set of estimators, we seek the minimizer of the objective function in (4.9) and take it as the final solution.

The algorithm for solving (4.1) can be found in Chapter 3 and is omitted here. To solve (4.7), we apply a blockwise coordinate descent method, in which D, L and $L^{(k)}$'s are

iteratively updated till the convergence of the objective function. To update D with fixed L and $L^{(k)}$'s, we apply RMSProp, a modified version of gradient descent with adaptive learning rate (Tieleman and Hinton, 2012). To update L with fixed D and $L^{(k)}$'s, we write $L = RR^T$ ($R \in \mathbf{R}^{p,r}$) and minimize the objective function with respect to R with the same descent method. To update $L^{(k)}$'s with fixed D and L, we notice that each $L^{(k)}$ affects the objective function only through the likelihood of its own category; thus, $L^{(k)}$'s can be obtained independently from each other. As obtaining optimal $L^{(k)}$'s boils down to K single category problems, we will borrow the step of updating the low-rank component in the "diagonal + low-rank" method.

Our experience with the numerical experiment shows that a proper initialization for (4.7) can make the computation more efficient. For $r = 0$, we initialize D to be the analytical minimizer of the objective function when L and $L^{(k)}$'s are set to be 0. For each $r > 0$, we initialize D to be $\widehat{D}_{r-1,v}$, where $\widehat{D}_{r-1,v}$ is the diagonal component of $\widehat{\Theta}_{r-1,v}$, and then we initialize R (and $L = RR^T$) to be the analytical minimizer of the objective function when $L^{(k)}$'s are set to be 0.

See Algorithm 2 for the details. For the purpose of presentation, let

$$f(D, L, L^{(1)}, \ldots, L^{(K)}) = \sum_{k=1}^{K} n_k \left[\mathrm{tr}\{(D - L - L^{(k)})S^{(k)}\} - \log |D - L - L^{(k)}| \right]$$

be the objective function, and $S = n^{-1} \sum_{k=1}^{K} (n_k S^{(k)})$ be the pooled sample covariance. The analytical solution of D in line 3 is straightforward by basic calculus. The analytical solutions of L and $L^{(k)}$'s in line 7 and 12 are direct generalizations of Lemma B.4 in Appendix B.2; these solutions degenerate to 0 if the corresponding ranks (r or $(v_k - r)$'s) are 0. The numerical update of L in line 23 degenerates to 0 if $r = 0$.

4.6 Numerical experiment

In this section, for the purpose of comparison, we investigate the estimation accuracy of sample covariance matrices $S^{(k)}$'s, diagonal matrices, which keep the diagonal elements of $S^{(k)}$'s, the "diagonal + low-rank", which is separately applied to every category, the

Algorithm 2: Blockwise coordinate descent algorithm for solving (4.7) for fixed (r, v).

$f_{old} = \infty$;
if $r = 0$ **then**
  $D = \mathrm{diag}(1/\mathrm{diag}(S))$;
else
  $D = \widehat{D}_{r-1,v}$;
end
$R = D^{1/2}QW^{1/2}$, $L = RR^T$, in which
  $W = \mathrm{diag}\{1 - 1/\max(w_1, 1), \ldots, 1 - 1/\max(w_r, 1)\}$, $Q = (q_1, \ldots, q_r)$;
  $w_1, \ldots, w_r$ are the r largest eigenvalues of $D^{1/2}SD^{1/2}$;
  $q_1, \ldots, q_r$ are the associated eigenvectors;
while do
  $L^{(k)} = (D - L)^{1/2}Q^{(k)}W^{(k)}(Q^{(k)})^T(D - L)^{1/2}$, in which
    $W^{(k)} = \mathrm{diag}\{1 - 1/\max(w_1^{(k)}, 1), \ldots, 1 - 1/\max(w_{v_k - r}^{(k)}, 1)\}$;
    $Q^{(k)} = (q_1^{(k)}, \ldots, q_{v_k - r}^{(k)})$;
    $w_1^{(k)}, \ldots, w_{v_k - r}^{(k)}$ are the $v_k - r$ largest eigenvalues of $(D - L)^{1/2}S^{(k)}(D - L)^{1/2}$;
    $q_1^{(k)}, \ldots, q_{v_k - r}^{(k)}$ are the associated eigenvectors;
  $f_{new} = f(D, L, L^{(1)}, \ldots, L^{(K)})$;
  if $|f_{new} - f_{old}| < tol_1$ **then**
    return D, L, $L^{(k)}$'s;
  end
  $f_{old} = f_{new}$;
  update D with Algorithm 3, $g(D) = f(D, L, L^{(1)}, \ldots, L^{(K)})$;
  update R (and L) with Algorithm 3, $g(R) = f(D, RR^T, L^{(1)}, \ldots, L^{(K)})$;
end

Algorithm 3: RMSProp for line 22 and 23 in Algorithm 2. The initialization $x^{(0)}$ is the value (of D or R) before the update; m has the same shape as x, and operations in line 3 and 4 are elementwise; $\nabla g(x)$ is the gradient; α is the learning rate, a scalar.

1 Initialization: $x = x^{(0)}$, $m = 0$, $g_{old} = +\infty$;
2 **while do**
3 $\quad m = 0.9m + 0.1\{\nabla g(x)\}^2$;
4 $\quad x = x - \alpha \nabla g(x) \left\{(m)^{1/2} + 10^{-8}\right\}^{-1}$;
5 $\quad g_{new} = g(x)$;
6 $\quad$ **if** $|g_{new} - g_{old}| < tol_2$ **then**
7 $\quad\quad$ **return** x;
8 $\quad$ **end**
9 $\quad g_{old} = g_{new}$;
10 **end**

"joint diagonal+low-rank" and the fused graphical lasso (Danaher et al., 2014). The fused graphical lasso is implemented by the `R` package `JGL`.

We set $K = 3$ and experiment $p = 50$, $p = 100$ and $p = 200$. The sample sizes are $n_k = 100,\ k = 1, 2, 3$. For each category, data are independently and identically generated from a multivariate normal distribution with zero mean.

In the following examples, we employ various matrices structures to demonstrate the performances. To set up some notations, $1_{p\times p}$ is a $p \times p$ matrix with all elements being one; $0_{p_1\times p_2}$ is a $p_1 \times p_2$ matrix with all elements being zero; (b_1, b_2) represents a matrix formed by stacking b_1 and b_2 horizontally (given b_1, b_2 are matrices with the same number of rows); $U(\cdot, \cdot)$ represents a Uniform distribution with the input lower bound and upper bound, $U(\cdot)$ represents a Uniform distribution on the input interval, and $Ber(\cdot)$ denotes a Bernoulli distribution with the input success probability; $k = 1, 2, 3$ in all examples; whenever values are randomly drawn, values with different indices are drawn independently from each other.

Example 1: The matrices are constructed based on the compound symmetry structure,

$\Sigma_0^{(k)} = 0.5I_p + 0.21_{p\times p} + 0.2a^{(k)}(a^{(k)})^T$, where $a^{(k)} \in \mathbf{R}^p$ and $a_j^{(k)} \sim U(0,1)$.

Example 2: The matrices are block diagonal and have different number of blocks, $\Sigma_0^{(k)} = 0.5I_p + 0.2aa^T + 0.2a^{(k)}(a^{(k)})^T$, $a = (b_1, b_2)$, $a^{(1)} = 0$, $a^{(2)} = (b_3, b_4)$, $a^{(3)} = (b_5, b_6)$, $b_i = (0_{q\times(i-1)q}, 1_{q\times q}, 0_{q\times(p-iq)})^T$ and $q = \lfloor p/(2K) \rfloor$.

Example 3: The matrices have the "joint low-rank+diagonal" structure and are sparse, and the diagonal component is randomly drawn. Let $\Sigma_0^{(k)} = \mathrm{diag}(d) + 0.2aa^T + 0.2a^{(k)}(a^{(k)})^T$, where $d \in \mathbf{R}^p$, $a \in \mathbf{R}^{p\times 4}$, $a^{(k)} \in \mathbf{R}^{p\times 2}$, $d_j \sim U(0.2, 0.6)$, $a_{j_1,j_2} \sim Ber(0.4)$ and $a_{j_1,j_2}^{(k)} \sim Ber(0.4)$.

Example 4: The matrices have a perturbed "joint low-rank +diagonal" structure. Let $B^{(k)} = 0.5I_p + 0.2aa^T + 0.2a^{(k)}(a^{(k)})^T + P^{(k)}$, where $a \in \mathbf{R}^{p\times 2}$, $a^{(k)} \in \mathbf{R}^p$, $a_{j_1,j_2} \sim Ber(0.4)$, $a_j^{(k)} \sim Ber(0.4)$. The perturbation matrix $P^{(k)} = 0.01\{P_0^{(k)} + (P_0^{(k)})^T\}/2$, where $P_0^{(k)} \in \mathbf{R}^{p\times p}$ and each element in $P_0^{(k)}$ is randomly drawn from $U(0,1)$ with probability 0.2 and set to zero otherwise. Let $\Sigma_0^{(k)} = B^{(k)} + \{|\min(\lambda_{\min}(B^{(k)}), 0)| + 0.05\}I_p$ so that it's positive definite.

Example 5: The precision matrices are sparse and some elements are shared across categories. Let $B_0^{(k)} = 0.2a + 0.2a^{(k)}$, where $a, a^{(k)} \in \mathbf{R}^{p\times p}$, $a_{j_1,j_2} \sim Ber(0.2)$, $a_{j_1,j_2}^{(k)} \sim Ber(0.2)$, and $B^{(k)} = 0.5I_p + \{B_0^{(k)} + (B_0^{(k)})^T\}/2$. Let $\Sigma_0^{(k)} = [B^{(k)} + \{|\min(\lambda_{\min}(B^{(k)}), 0)| + 0.05\}I_p]^{-1}$.

Example 6: The matrices correspond to K networks — this is a similar structure as experimented by Danaher et al. (2014). Let $A \in \mathbf{R}^{p\times p}$ be the adjacency matrix of a network that contains 10 equally sized unconnected subnetworks, each with a power law degree distribution. Let $(B_0)_{j_1,j_2} = A_{j_1,j_2}a_{j_1,j_2}$, where $a_{j_1,j_2} \sim U\{(-0.4,-0.1) \cup (0.4,0.1)\}$, $B_1 = I_p + (B_0 + B_0^T)/2$, $B_2 = [B_1 + \{|\min(\lambda_{\min}(B_1), 0)| + 0.05\}I_p]^{-1}$, $B_{j_1,j_2} = d_{j_1,j_2}(B_2)_{j_1,j_2}\{(B_2)_{j_1,j_1}(B_2)_{j_2,j_2}\}^{-1/2}$, where d_{j_1,j_2} is 0.6 if $j_1 \neq j_2$ and 1 otherwise. Finally, $\Sigma_0^{(1)} = B$, $\Sigma_0^{(2)}$ is the same as $\Sigma_0^{(1)}$ except one subnetwork block is the identity, and $\Sigma_0^{(3)}$ is the same as $\Sigma_0^{(2)}$ except an additional subnetwork block is the identity.

To implement our methods, we set $\mathbb{Z}^{(k)} = \{0, 1, 2, 4, 6, 8\}$ for $k = 1, 2, 3$. The true ranks

of Example 1 – Example 3 are included in $\mathbb{Z}^{(k)}$ whereas those of the other three examples are not.

Tuning parameters for DL, JDL and FGL are all selected via minimizing the negative log-likelihoods of separately generated validation sets (Rothman et al., 2008); the validation sample size is also 100 for each category. The tuning parameter sets for DL and JDL are $\{0.2, 0.4, \ldots, 1.2\}$ and $\{0.8, 1.0, \ldots, 1.8\}$ respectively. On a side note, unlike the size of $\mathbb{Z}^{(k)}$, the sizes of the tuning parameter sets do not affect our computational cost. FGL has two tuning parameters — λ_1 for sparsity and λ_2 for shared network and edge values. We let both the tuning parameter sets of λ_1 and λ_2 contain 5 evenly spaced real numbers over $(0.01, 0.1)$ for Example 1 – Example 5 and $(0.1, 0.3)$ for Example 6, and the best pair is selected after validation on the grid.

4.6.1 Estimation accuracy

The estimation accuracy is evaluated by Kullback-Leibler loss:

$$L_{KL}(\Theta, \Theta_0) = \sum_{k=1}^{K} \left\{ \mathrm{tr}(\Sigma_0^{(k)} \Theta^{(k)}) - \log |\Sigma_0^{(k)} \Theta^{(k)}| - p \right\}. \tag{4.10}$$

The loss quantifies how far the proposed estimators are from the population values; it is minimized at $\Theta = \Theta_0$ and has a minimum value of zero. Output from DL, JDL and FGL can be directly plugged into (4.10), while the sample covariance matrices and diagonal matrices have to be inverted first. As the sample covariance matrices become singular for $p = 100$ and $p = 200$, corresponding losses are marked as "NA".

See Table 4.1 for the results. The sample covariance matrix is the worst among all, and the diagonal matrix outperforms the sample covariance matrix but is still worse than the other three. For Example 1 and 3, where the "joint diagonal + low-rank" structure is satisfied, DL and JDL produce nice results and enjoy lower KL losses than FGL. For Example 4, where the "joint diagonal + low-rank" structure is randomly perturbed, the nice performances of DL and JDL show their robustness. Example 2 explores the block diagonal structure and satisfies both the "joint diagonal + low-rank" structure and the joint sparse graphical model; we can see that DL, JDL and FGL all perform well, and JDL

and FGL outperform DL since they both exploit the joint matrix structure. For Example 5 and 6, which are built to favor the joint sparse graphical model, FGL produces better results as expected, but our methods also show some power in these two cases.

Table 4.1: Average (standard error) of Kullback-Leibler loss over 100 replications.

		Sample Cov	Diagonal	DL	JDL	FGL
Example 1	$p = 50$	112.07(0.57)	53.43(0.02)	4.22(0.03)	2.64(0.02)	5.37(0.02)
	$p = 100$	NA	115.2(0.05)	8.61(0.05)	4.88(0.06)	11.8(0.04)
	$p = 200$	NA	241.1(0.08)	16.6(0.10)	8.81(0.07)	24.9(0.07)
Example 2	$p = 50$	112.29(0.56)	14.13(0.02)	6.63(0.05)	3.62(0.04)	4.07(0.04)
	$p = 100$	NA	36.97(0.02)	13.2(0.07)	7.09(0.04)	9.62(0.08)
	$p = 200$	NA	90.82(0.04)	26.0(0.10)	14.1(0.08)	20.6(0.11)
Example 3	$p = 50$	111.79(0.57)	70.130(0.02)	10.62(0.06)	6.74(0.05)	12.2(0.05)
	$p = 100$	NA	175.51(0.04)	21.45(0.11)	13.02(0.1)	27.37(0.1)
	$p = 200$	NA	409.48(0.07)	43.14(0.18)	26.4(0.15)	57.3(0.17)
Example 4	$p = 50$	112.03(0.56)	40.60(0.02)	7.400(0.05)	3.930(0.04)	7.330(0.04)
	$p = 100$	NA	86.58(0.03)	14.58(0.08)	7.590(0.06)	15.67(0.08)
	$p = 200$	NA	189.8(0.06)	28.45(0.12)	15.46(0.09)	33.93(0.15)
Example 5	$p = 50$	111.23(0.63)	64.830(0.02)	27.00(0.09)	25.96(0.09)	18.58(0.07)
	$p = 100$	NA	131.73(0.03)	70.84(0.12)	68.18(0.11)	52.22(0.09)
	$p = 200$	NA	273.85(0.04)	170.7(0.20)	165.4(0.19)	137.6(0.16)
Example 6	$p = 50$	111.75(0.55)	3.080(0.02)	3.080(0.02)	2.01(0.01)	1.56(0.02)
	$p = 100$	NA	6.490(0.03)	6.490(0.03)	4.38(0.02)	3.58(0.03)
	$p = 200$	NA	13.06(0.04)	13.06(0.04)	8.80(0.02)	7.26(0.03)

In the end, we draw the most essential conclusion of this section that JDL consistently outperform DL when there is some shared matrix structure across the categories, suggesting that exploiting such structure indeed helps improve the estimation accuracy.

4.6.2 Rank recovery

In this section, we provide the empirical evidence of the rank recovery performance — how JDL identifies the ranks of L and $L^{(k)}$'s, by comparing the 10 largest eigenvalues of the population low-rank components with those of the estimated ones.

We firstly find out the population low-rank components. For Example 1 – Example 3, with information about D_{Σ_0}, L_{Σ_0} and $L^{(k)}_{\Sigma_0}$ in the example setup, D_0, L_0 and $L^{(k)}_0$ can be derived by (4.6). Example 4 only approximately satisfies the joint diagonal and low-rank decomposition; we let $D_{\Sigma_0} = 0.5\,\mathrm{diag}(d)$, $L_{\Sigma_0} = 0.2aa^T$ and $L^{(k)}_{\Sigma_0} = 0.2a^{(k)}(a^{(k)})^T$, and the corresponding D_0, L_0 and $L^{(k)}_0$ can be derived by (4.6). Example 5 and Example 6 are omitted in this section for the lack of population values. The estimated low-rank components are directly produced by the JDL algorithm. As the rank recovery performances are similar for various dimensions, we show the plots for $p = 100$.

See Figure 4.1 for the result. In each plot, the number of non-zero eigenvalues correspond to the rank of the matrix. We conclude that on average JDL does yield successful recovery of the ranks of L_0 and $L^{(k)}_0$'s.

From the plots of Example 4, it can be seen that the overall recovery is not as accurate as those of the other three examples. This is a consequence of $v_{0k} \notin \mathbb{Z}^{(k)}$ instead of the perturbation. To show this, we let $\mathbb{Z}^{(k)}$ contain all integers from 0 to 8 and reproduce the plots (Figure 4.2). This suggests that (i) including more ranks in $\mathbb{Z}_k$ might improve the performance at the cost of computational expense; (ii) JDL performs well even if the population covariance matrix does not conform precisely to the decomposition assumption.

4.7 Real data analysis

We apply the JDL to `WebKB`, a dataset that contains webpages of computer science departments of a few universities. The webpages were collected in 1997 and manually classified into 7 categories: `student`, `faculty`, `project`, `course`, `staff`, `department`, and `other`. The dataset has been pre-proccessed by Cardoso-Cachopo (2007) and can be downloaded from `http://ana.cachopo.org/datasets-for-single-label-text-categorization`.

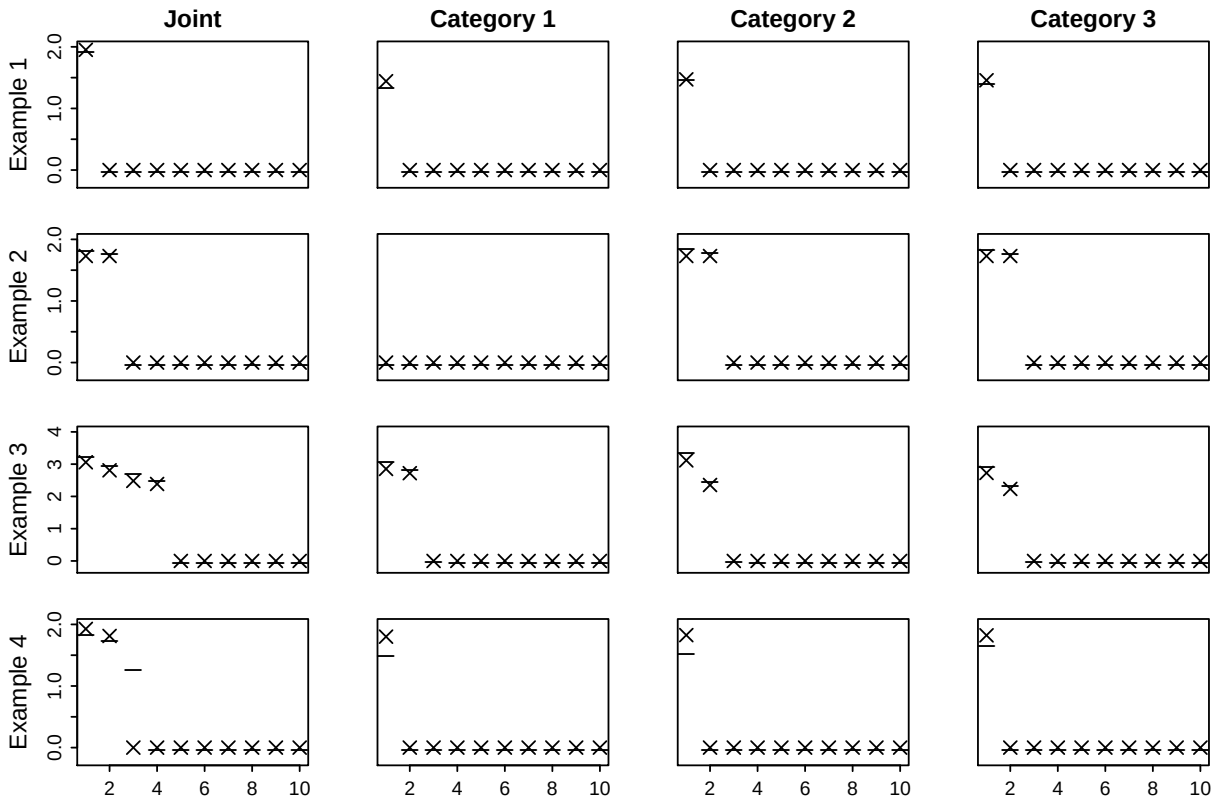

Figure 4.1: Comparison of the 10 largest eigenvalues of the population low-rank components ("×") and those of the estimated ones ("−"). For the estimated eigenvalues, the bars represent the averages over 100 replications. The leftmost column corresponds to the joint low-rank component L and the other three correspond to $L^{(k)}$, $k = 1, 2, 3$.

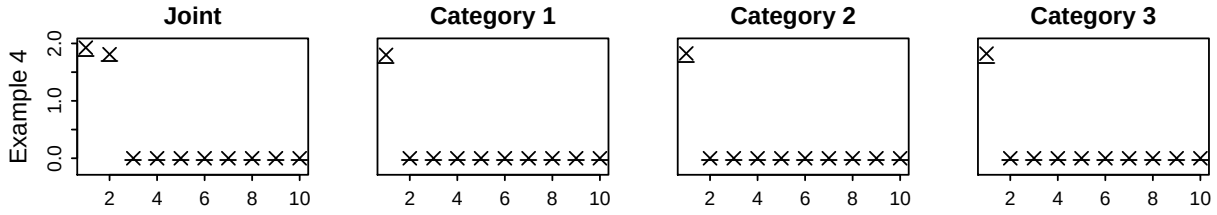

Figure 4.2: Comparison of the 10 largest eigenvalues of the population low-rank components ("×") and those of the estimated ones ("−") for Example 4 when $v_{0k} \in \mathbb{Z}^{(k)}$.

Pre-processing procedures include (i) discarding the categories with too small sample size (`staff` and `department`) or very diverse pages (`other`); (ii) randomly choosing 2/3 of the samples to be training data and letting the remaining 1/3 to be testing data (Table 4.2); (iii) standard text processing such as removing short words and stemming.

Table 4.2: Sample sizes of `WebKB`

Class	Training data	Testing data	Total
`student`	1097	544	1641
`faculty`	750	374	1124
`project`	336	168	504
`course`	620	310	930

With `R` package `tm`, we use the standard "term frequency – inverse document frequency" (tf-idf) to weight a document-term matrix. Let $tf_{i,j}$ be the count of occurrences of term j in document i, we use the weighting function $0.5 + 0.5\, tf_{i,j}/\max_j(tf_{i,j})$ for term frequencies and $\log_2(N/df_j)$ for inverse document frequencies, where N denotes the total number of documents and df_j is the number of documents in which term j occurs. The final weight is the product of the term frequency and the inverse document frequency. During weighting, the training set and the testing set are combined, and no label information is used.

The feature selection is done by applying document frequency thresholding (Largeron et al., 2011) to the training data. We keep $p = 200$ out of 7203 terms with the highest

document frequencies. This method assumes that a term that occurs in only a small number of documents is not an effective feature for categorization.

To perform the "diagonal + low-rank" method for each category, we let $\mathbb{Z}^{(k)}$, $k = 1, \ldots, 4$, contain every integer from 0 to 10 and the tuning parameter sets contain 6 evenly spaced real numbers over $(2.4, 3.4)$ for the category `project` and $(2.7, 3.7)$ for the other categories. In the joint diagonal and low-rank decomposition step, similarly, the tuning parameter set contains 6 evenly spaced real numbers over $(4.5, 5.5)$. We employ heavy penalties to enhance the interpretability of the latent factors. It will also be seen later that this regularization improves the performance of quadratic discriminant analysis (QDA).

We select the tuning parameters by a 3-fold cross-validation. The training set is partitioned into 3 subsamples with equal sizes. In each round, one of these subsamples is used as the validation data. The cross-validation scores are the log-likelihoods of normal distributions, evaluated on validation data and averaged over the rounds. The tuning parameter associated with the highest score is selected.

Recall that we mentioned in Section 4.2.2 that the "joint diagonal + low-rank" assumption can be understood as factor models. The effects of r_0 factors are common, and the effects of $v_{0k} - r_0$ factors are specific to category k; now we can take a look at these factors in the context of the webpage data.

The selected ranks are $v = (5, 5, 2, 3)$ and $r = 2$. The estimated components of the precision matrices are produced by JDL and then converted to $\widehat{L}_{\Sigma_0}$ and $\widehat{L}^{(k)}_{\Sigma_0}$ by a similar expansion to (4.6). We first write $\widehat{L}_{\Sigma_0} = RR'$ ($R \in \mathbf{R}^{p \times r}$) and obtain R by multiplying the square roots of the largest r eigenvalues of $\widehat{L}_{\Sigma_0}$ to their associated eigenvectors. Then, we apply a `VARIMAX` rotation to R, which maximizes the variance of squared loadings, so that each factor tends to have less but "larger" (in terms of the magnitude) non-zero loadings. At last, we identify, for each factor, the covariates associated with "large" loadings and try to understand what the factor represents and affects. Likewise, $\widehat{L}^{(k)}_{\Sigma_0}$ can be analyzed.

The identified covariates are plotted in Figure 4.3. The shown words are among the top 10 (for 3D plot) or 15 (for 2D plot) covariates of at least one of the factors in the same subplot. Category `project` does not have factors with category-specific effects and is omitted.

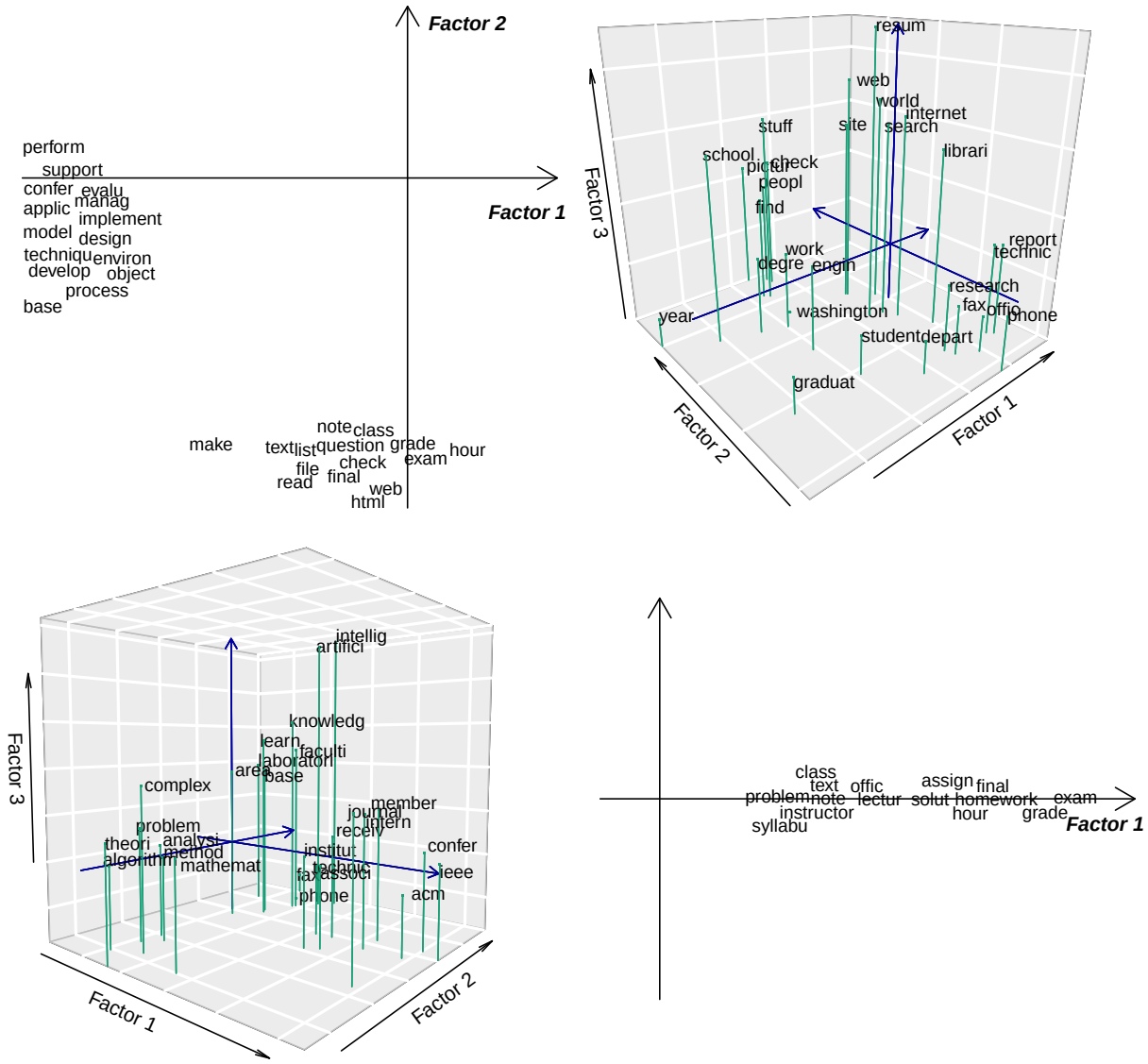

Figure 4.3: Loadings of covariates after the `VARIMAX` rotation. The subplots are, from left to right, top to bottom, for factors with common effects, factors that affect category `student`, `faculty` and `course`. The last subplot only contains one factor.

We can see that, for factors that affect all categories alike, one is related to technical support and the other is related to teaching and studying. The other three categories all have their characteristic words. To be more specific, `student` has a factor of academic information (e.g., `year`, `graduate`) and a factor of job seeking (e.g., `resume`); `faculty` has some words related to research interest (e.g., `artific`, `intellig`, `laboratori`) and others to academic activity (e.g., `journal`, `associ`, `confer`, `member`); `course` only has one extra factor, and as expected it represents course information and materials (e.g., `lectur`, `homework`, `syllabu`).

The final task is to see how the regularization affects the performance of QDA. To this end, we consider a score for each category based on the Bayes rule,

$$Q_k(x) = 2 \log \pi^{(k)} + \log |\Theta^{(k)}| - (x - \mu^{(k)})^T \Theta^{(k)} (x - \mu^{(k)}),$$

in which x is an unlabeled vector, $\pi^{(k)}$ is the category prior and $\mu^{(k)}$ is the category mean. Based on the training data, we estimate $\pi^{(k)}$ with n_k/n and $\mu^{(k)}$ with category sample mean. The observation x will be classified to the category with the largest score.

We intend to experiment the same covariance/precision matrix estimators as in Section 4.6. However, as there are 2 covariates with zero variance in category `project`, we have to use the Moore-Penrose generalized inverse for the sample covariance matrix and adjust these 2 variances to be the smallest non-zero variance for the method "Diagonal" as well as the initialization of DL and JDL. In addition, as the `R` package `JGL` does not cope with this situation, we decide to omit the joint graphical lasso.

Random forest (RF) is included as a suggestion of the accuracy that can be achieved for this dataset; it is trained after the same pre-processing and feature selection procedures. All classifiers are applied to the test data and evaluated by classification accuracy. See Table 4.3 for the results. We conclude that (i) the separate rank regularization (DL) does help enhance the performance and (ii) exploiting the shared matrix structure (JDL) makes further improvement.

Table 4.3: Classification accuracy of QDA rules (based on various covariance/precision matrix estimators) and random forest.

	Sample Cov	Diagonal	DL	JDL	RF
Accuracy	75.28%	80.30%	82.31%	86.17%	88.18%

4.8 Conclusion

We proposed a method that jointly estimates high-dimensional covariance/precision matrices of multiple categories. The method decomposes each covariance/precision matrix into a shared diagonal matrix, a shared low-rank matrix and a category-specific low-rank matrix.

Starting with fixed-rank estimators, we emphasized the importance of accurate pre-selected ranks and pointed out the difficult of specifying them. Then, we considered an AIC-type penalty that encourages the proposed decomposition and automatically selects the ranks. We established that, under certain technical conditions, the estimators obtained via imposing the penalty have the same consistency property as fixed-rank estimators with correct pre-selected ranks.

An algorithm, which iteratively updates the diagonal matrix and the low-rank matrices, has been developed; and several techniques to reduce the computational cost have been discussed. We used simulations to empirically assess the estimation accuracy of our method and were able to see the advantage of exploiting the shared matrix structure. In real data analysis, the estimators were applied to factor model analysis and quadratic discriminant analysis; we found that interpretable factors could be identified, and the regularization of ranks did improve the classification accuracy.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we studied a series of matrix structures to facilitate high-dimensional covariance matrix estimation. Firstly, we investigated the compound symmetry structure in the context of quadratic discriminant analysis. Then, we generalized it and considered the more flexible "diagonal + low-rank" structure. In the end, we studied the "joint diagonal+low-rank" structure in order to simultaneously estimate multiple covariance matrices while exploiting some common structure among them.

Based on the compound symmetry structure, we developed a set of QDA rules. The ppQDA rule forms a substitute for each covariance matrix by pooling both diagonal elements and off-diagonal elements of the sample covariance matrix. The pQDA rule ignores the correlations and is a special case. The Se-ppQDA rule and the Se-pQDA rule generalize the ppQDA rule and the pQDA rule respectively in order to handle nonnormal data. Theoretical properties of some of these rules and empirical performances of all of them were provided. We showed that, in spite of the simplicity, these rules enjoy low misclassification probability as long as the population covariance matrices moderately satisfy the assumed matrix structure. In practice, the optimal rule could vary from case to case. We suggest that users investigate the data, pre-process the data, and decide a suitable rule.

The compound symmetry structure could be written as the summation of a scaled identity matrix and a rank-1 matrix. To pursue and generalize this decomposition, we proposed a covariance matrix estimator, the DL estimator, that could be written as the summation of a diagonal matrix and a low-rank matrix. An important step of obtaining the DL estimator is to find an appropriate rank, which could exploit the decomposition to its fullest potential. We started with considering a pre-selected rank and showed that consistency of the DL estimator could be established when the rank is "approximately" correct. However, specifying a relatively accurate rank is not trivial in general, and an inaccurate rank could lead to inconsistency. To tackle this problem, we considered a penalty, which was directly imposed on the rank. It could be proven that, with a proper choice of the penalty function, the obtained estimator converges as if the correct rank was provided. We compared the DL with the graphical lasso via simulations, and concluded that, although both methods outperform the sample covariance matrix and the diagonal matrix, that keeps the diagonal elements of the sample covariance matrix, their relative performances depend on whether the "diagonal + low-rank" or the sparsity assumption is closer to the reality.

To extend the DL estimator from estimating a single covariance matrix to estimating multiple ones, we proposed the "joint diagonal + low-rank" structure. This particular structure allows the matrices to share some common components. To be more specific, each covariance matrix has the "diagonal + low-rank" structure when viewed independently, but the low-rank component can be further separated into a category-specific low-rank matrix and a shared low-rank matrix. Exploiting such a matrix structure, we developed an estimator — the JDL estimator. As for the DL estimator, we explored estimations with either pre-selected ranks or a rank penalty, and established consistency properties when the pre-selected ranks are "approximately" correct or the penalty is properly defined. In the implementation of the JDL algorithm, there is a shared rank and K (the number of categories) category-specific ranks to be taken care of; to avoid considering complicated combinations of the ranks and the consequent high computational cost, we suggested users to apply the penalized DL to each category to decide the category-specific ranks first and then apply the penalized JDL to decide the shared rank and obtain the estimators. The simulations demonstrated the advantage of the JDL over independently applying the DL

to every category. We showed, via real data analysis, that the JDL estimator could be used to identify factors as well as to facilitate discriminant analysis.

Through this thesis, we explored a series of covariance matrix estimators and established their nice properties both theoretically and empirically. However, these properties were developed only under certain conditions. High-dimensional covariance matrix estimation is a challenging task, and it is implausible to give a universally optimal solution. All estimators are subject to some underlying assumptions about the structure of the population covariance matrix; therefore, for a specific data set, it is crucial to make a decision about which estimators to use and how to pre-process the data to use the estimators to their fullest potential.

5.2 Future Work

One possible extension is to consider relaxing the normality assumptions in Chapter 3 and Chapter 4. To do so, we would almost certainly need to make explicit assumptions about the tail behavior of the data distribution, which might change the convergence rates of the resulting estimators. Although our objective functions are based on the normal likelihood, they work by pushing the covariance matrix estimators towards the sample covariance matrices on one hand and encouraging the assumed "diagonal+low-rank" or "joint diagonal +low-rank" on the other. As a result, the estimation accuracy depends on how well the sample covariance matrices can approximate their population counterparts, which is affected by the tail behavior of the data distribution.

Since the DL and JDL estimators are generalizations of the compound symmetry structure, which we considered in the context of discriminant analysis. It is natural to ask whether we can construct new discriminant rules by substituting DL or JDL estimators for the compound symmetry estimators in the ppQDA rule. A small experiment has been done on this in the real data analysis in Chapter 4; however, many properties still remain to be investigated. Intuitively, such new rules would work under more relaxed assumptions about the population covariance matrices than the ppQDA rule. Furthermore, we might be able to compare the new rules with the Bayes decision rule through the already established

bounds of estimation errors of the DL or JDL estimators — the QDA rules should converge to the Bayes decision rule if the estimators converge to their population counterparts.

In order to obtain the DL estimator and the JDL estimator, we applied algorithms that alternately update variables in the objective functions. This could lead to local minimizers instead of global ones. A convenient solution is to initialize from multiple starting points to increase the chance of finding global minimizers. We did not recommend this, because our deterministic initialization ("warm starts") already produced nice results in numerical experiments, and it did not seem worthwhile to increase the computational cost. However, how the local minimizers compare with the global ones remains an open question. In the future, we might be interested in developing theoretical properties of the local minimizers or establishing connections between the local minimizers and the global minimizers.

Another potential extension of JDL is to consider dependence across the categories. From a "factor model" point of view, the current framework assumes that the random vectors depend on latent factors, while all the factors are independent of each other, in spite of the common effects introduced by common loading matrices. However, when certain random vectors of various categories depend on the same factors, dependence must be considered. For example, stocks of different sectors (categories) might be related to the same set of economic indicators; in this case, all stock returns at a certain time point are related to the indicators at that time point and must not be independent of each other. To take the dependence into account, it will be necessary to alter the objective function, since building the overall likelihood by multiplying likelihoods of every observation requires independence. To this end, we might consider a likelihood based on the joint distribution of all categories or another function that also pushes the estimators towards the sample covariance matrices.

In the end, motivated by the compound symmetry structure, "diagonal + low-rank" and "joint diagonal + low-rank", many other matrix structures might be considered. The "factor model" interpretation suggests some reasonable options. For instance, we can replace the diagonal component with a sparse one, so that the implied independence of the error terms is relaxed. We can also encourage the low-rank matrix or the associated loading matrix to be sparse; a sparse low-rank matrix indicates many uncorrelated covariates, and a sparse loading matrix implies that each covariate depends on a smaller number of latent

factors. In the joint estimation case, if the low-rank matrices are assumed to be sparse, the model indicates not only shared factor effects but also shared network links. That being said, it is not trivial to choose proper penalties to encourage these structures, and developing efficient algorithms is also a challenge.

References

Akaike, H. (1987). Factor analysis and aic. *Psychometrika*, 52(3):317–332.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley.

Aoshima, M. and Yata, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics*, 66(5):983–1010.

Aoshima, M. and Yata, K. (2015). High-dimensional quadratic classifiers in non-sparse settings. *arXiv preprint arXiv:1503.04549*.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 6(2):311–329.

Banerjee, O., Ghaoui, L. E., and dAspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

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

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

Butte, A. (2002). The use and analysis of microarray data. *Nature Reviews Drug Discovery*, 1(12):951–960.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2016a). Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26(2):445–464.

Cai, T. T. and Liu, W. (2011a). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684.

Cai, T. T. and Liu, W. (2011b). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577.

Cai, T. T., Liu, W., and Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cai, T. T., Liu, W., and Xia, Y. (2013a). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *Journal of the American Statistical Association*, 108(501):265–277.

Cai, T. T., Ma, Z., and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, 161(3-4):781–815.

Cai, T. T., Ren, Z., and Zhou, H. H. (2013b). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probability Theory and Related Fields*, 156(1-2):101–143.

Cai, T. T., Ren, Z., and Zhou, H. H. (2016b). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59.

Cardoso-Cachopo, A. (2007). Improving Methods for Single-label Text Categorization. PhD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935 – 1967.

Chen, S. and Qin, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397.

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(457):77–87.

El Karoui, N. (2010). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *The Annals of Statistics*, 38(6):3487–3566.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36(6):2605.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771.

Fan, J., Ke, Z. T., Liu, H., and Xia, L. (2015). QUADRO: A supervised dimension reduction method via Rayleigh quotient optimization. *The Annals of Statistics*, 43(4):1498–1534.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39(6):3320–3356.

Fan, J., Liao, Y., and Mincheva, M. (2013a). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Fan, Y., Jin, J., and Yao, Z. (2013b). Optimal classification in sparse Gaussian graphic model. *The Annals of Statistics*, 41(5):2537–2571.

Friedman, J. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Hao, N., Dong, B., and Fan, J. (2015). Sparsifying the fisher linear discriminant by rotation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):827–851.

Hashorva, E. and Hüsler, J. (2003). On multivariate Gaussian tails. *Annals of the Institute of Statistical Mathematics*, 55(3):507–522.

Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60.

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

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278.

Largeron, C., Moulin, C., and Géry, M. (2011). Entropy based feature selection for text categorization. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 924–928. ACM.

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

Li, J. and Chen, S. (2012). Two sample tests for high-dimensional covariance matrices. *The Annals of Statistics*, 40(2):908–940.

Li, Q. and Shao, J. (2015). Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, 25(2):457–473.

Lin, Y. and Jeon, Y. (2003). Discriminant analysis through a semiparametric model. *Biometrika*, 90(2):379–392.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328.

Lofberg, J. (2004). YALMIP: A toolbox for modeling and optimization in MATLAB. In *Proceedings of the 2004 IEEE International Symposium on Computer Aided Control Systems Design*, pages 284–289. IEEE.

Magnus, J. R. (1985). On differentiating eigenvalues and eigenvectors. *Econometric Theory*, 1(02):179–191.

Mai, Q. (2013). A review of discriminant analysis in high dimensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(3):190–197.

Mai, Q. and Zou, H. (2015). Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, 135:175–188.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42.

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

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

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Ockenhouse, C. F., Hu, W., Kester, K. E., Cummings, J. F., Stewart, A., Heppner, D. G., Jedlicka, A. E., Scott, A. L., Wolfe, N. D., Vahey, M., and Burke, D. S. (2006). Common and divergent immune response signaling pathways discovered in peripheral blood mononuclear cell gene expression patterns in presymptomatic and clinically apparent malaria. *Infection and Immunity*, 74(10):5561–5573.

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501.

Rocha, G. V., Zhao, P., and Yu, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE). *arXiv preprint arXiv:0807.3734*.

Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740.

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

Savage, I. R. (1962). Mill's ratio for multivariate normal distributions. *Journal of Research of the National Bureau of Standards B*, 66B(3):93–96.

Schott, J. R. (2005). *Matrix Analysis for Statistics*. John Wiley & Sons.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265.

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493.

Sun, T. and Zhang, C. (2013). Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418.

Taeb, A. and Chandrasekaran, V. (2017). Interpreting latent variables in factor models via convex optimization. *Mathematical Programming*. doi:10.1007/s10107-017-1187-7.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2).

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

Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772.

Xue, L., Ma, S., and Zou, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *Journal of the American Statistical Association*, 107(500):1480–1491.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286.

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

APPENDICES

Appendix A

Proofs of Chapter 2

A.1 Proofs of Theorems 2.1 and 2.2

The following lemma shows that the doubly pooled covariance matrix used in the ppQDA function is positive definite, which is due to all its eigenvalues being positive.

Lemma A.1. *Let $\Sigma = (\sigma_{ij})$ be a $p \times p$ covariance matrix, a and r be the average of diagonal and off-diagonal entries of Σ, respectively. Then for $p > 2$, $a - r > 0$, $a + (p-1)r > 0$, and $A = (a_{ij})$ is positive definite, where $a_{ij} = a$ if $i = j$, otherwise $a_{ij} = r$, for $i, j = 1, \cdots, p$.*

Proof. Notice that the matrix A has p eigenvalues which are $a + (p-1)r, a - r, \cdots, a - r$. To finish the proof, we only need to show that $a - r > 0$ and $a + (p-1)r > 0$.

For $1 \leq i < j \leq p$, let $\mathbf{e}_{ij}$ be a p-dimensional column vector whose i-th element is 1, j-th element is -1, and all other elements are 0. As $\Sigma = (\sigma_{ij})$ is a $p \times p$ covariance matrix, then

$$\mathbf{e}_{ij}'\Sigma\mathbf{e}_{ij} = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij} > 0 \text{ and } \sum_{1 \leq i < j \leq p} \mathbf{e}_{ij}'\Sigma\mathbf{e}_{ij} = p(p-1)(a-r) > 0.$$

Therefore, $a - r > 0$ if $p > 2$.

Let $\mathbf{1}_p$ be a p-dimensional column vector of 1's, then

$$\mathbf{1}_p'\Sigma\mathbf{1}_p = p[a + (p-1)r] > 0,$$

and $a + (p-1)r > 0$. This finishes the proof. □

The following lemma shows that the ppQDA function with true parameters enjoys the property of asymptotically perfect classification. We accomplish this by showing that the probability of misclassifying $\mathbf{x}$ from class 1 to class 2 tends to 0 as the ppQDA function is negative when the dimension p is sufficiently large. The probability of misclassifying $\mathbf{x}$ from class 2 to class 1 tending to 0 can also be proved in a similar fashion.

Lemma A.2. *Let Q be the ppQDA function with true parameters. Under conditions 2.1 and 2.4,*

$$\lim_{p\to\infty} R_p = \lim_{p\to\infty} \mathbb{P}(Q > 0|\mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(Q \leq 0|\mathbf{x} \in \mathcal{C}_2) = 0.$$

Proof. We only focus on the probability of misclassifying $\mathbf{x}$ from class 1 to class 2, i.e. $\mathbb{P}\left(Q > 0|\mathbf{x} \in \mathcal{C}_1\right)$. For $i = 1, 2$, let $A_i = T\Lambda_i T'$ be the eigen decomposition of A_i, where

$$\Lambda_i = diag\Big(a_i - r_i, \cdots, a_i - r_i, a_i + (p-1)r_i\Big),$$

$T = (\mathbf{t}_1, \ldots, \mathbf{t}_p)$ and $\mathbf{t}_p = (1/\sqrt{p}) \cdot \mathbf{1}_p$. Define $\alpha_j = \mathbf{t}_j'(\mathbf{x} - \boldsymbol{\mu}_1)$ and $\beta_j = \mathbf{t}_j'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, for $j = 1, \cdots, p$. The quadratic classification function with true parameters can be expressed

as

$$
\begin{aligned}
Q &= \ln\Big(|A_1|/|A_2|\Big) + (\mathbf{x}-\boldsymbol{\mu}_1)'A_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \\
&\quad -(\mathbf{x}-\boldsymbol{\mu}_1+\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'A_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_1+\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\
&= \ln\Big(|A_1|/|A_2|\Big) + (\mathbf{x}-\boldsymbol{\mu}_1)'T\Lambda_1^{-1}T'(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_1)'T\Lambda_2^{-1}T'(\mathbf{x}-\boldsymbol{\mu}_1) \\
&\quad -2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'T\Lambda_2^{-1}T'(\mathbf{x}-\boldsymbol{\mu}_1) - (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'T\Lambda_2^{-1}T'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\
&= \ln\Big(|A_1|/|A_2|\Big) + \Big[1/(a_1-r_1) - 1/(a_2-r_2)\Big]\sum_{j=1}^{p-1}\alpha_j^2 \\
&\quad +\alpha_p^2\Big\{1/[a_1+(p-1)r_1] - 1/[a_2+(p-1)r_2]\Big\} - 2\sum_{j=1}^{p-1}\beta_j\alpha_j/(a_2-r_2) \\
&\quad -2\beta_p\alpha_p/[a_2+(p-1)r_2] - \sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2) - \beta_p^2/[a_2+(p-1)r_2] \\
&= \ln\Big(|A_1|/|A_2|\Big) + \Big[1/(a_1-r_1) - 1/(a_2-r_2)\Big]\sum_{j=1}^{p-1}\alpha_j^2 \\
&\quad +\alpha_p^2/[a_1+(p-1)r_1] - 2\sum_{j=1}^{p-1}\beta_j\alpha_j/(a_2-r_2) \\
&\quad -\sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2) - (\alpha_p+\beta_p)^2/[a_2+(p-1)r_2]. \qquad \text{(A.1)}
\end{aligned}
$$

Next we consider $\sum_{j=1}^{p-1}\alpha_j^2$, α_p^2 and $\sum_{j=1}^{p-1}\beta_j\alpha_j$ in (A.1) separately, followed by discussing all other terms in (A.1). First of all,

$$
\begin{aligned}
\sum_{j=1}^{p-1}\alpha_j^2 &= (\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{t}_1,\ldots,\mathbf{t}_{p-1})(\mathbf{t}_1,\ldots,\mathbf{t}_{p-1})'(\mathbf{x}-\boldsymbol{\mu}_1) \\
&= (\mathbf{x}-\boldsymbol{\mu}_1)'(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_1),
\end{aligned}
$$

such that

$$
\begin{aligned}
\mathbb{E}\left(\sum_{j=1}^{p-1}\alpha_j^2\right) &= tr\left[\left(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\right)\Sigma_1\right] \\
&= (p-1)(a_1-r_1).
\end{aligned}
$$

In addition,

$$\begin{aligned} Var\left(\sum_{j=1}^{p-1}\alpha_j^2\right) &= 2tr\left[\left(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\right)\Sigma_1\left(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\right)\Sigma_1\right] \\ &= 2\left[tr(\Sigma_1^2)-\frac{2}{p}Su(\Sigma_1^2)+\frac{1}{p^2}Su^2(\Sigma_1)\right] \\ &= 2(p-1)(a_1-r_1)^2+o(p^2). \end{aligned}$$

The last equality is due to Condition 2.4-3 and Condition 2.4-4. Notice that Condition 2.4-3 is equivalent to

$$tr(\Sigma_i^2)-(p-1)(a_i-r_i)^2=Su^2(\Sigma_i)/p^2+o(p^2)$$

and Condition 2.4-4 is equivalent to

$$Su(\Sigma_i^2)=Su^2(\Sigma_i)/p+o(p^2),$$

for $i=1,2$. Hence,

$$\sum_{j=1}^{p-1}\alpha_j^2 = \ (p-1)(a_1-r_1)+o_p(p). \tag{A.2}$$

Secondly, given that $\alpha_p\sim N(0,Su(\Sigma_1)/p)$, then

$$\left[a_1+(p-1)r_1\right]^{-1}\alpha_p^2\sim\chi_1^2. \tag{A.3}$$

Thirdly, notice that $\sum_{j=1}^{p-1}\beta_j\alpha_j$ can be expressed as

$$\sum_{j=1}^{p-1}\beta_j\alpha_j =(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_1),$$

with $\mathbb{E}\left(\sum_{j=1}^{p-1}\beta_j\alpha_j\right)=0$ and

$$Var\left(\sum_{j=1}^{p-1}\beta_j\alpha_j\right) =(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p')\Sigma_1(I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2).$$

Let $\lambda_{\max}(\Sigma_1 - A_1)$ be the largest eigenvalue of $\Sigma_1 - A_1$. According to Condition 2.4-3, $tr\left[(\Sigma_1 - A_1)^2\right] = o(p^2)$, then $\lambda^2_{\max}(\Sigma_1 - A_1) = o(p^2)$ and $\lambda_{\max}(\Sigma_1 - A_1) = o(p)$. As a result,

$$\begin{aligned}
& Var\left(\sum_{j=1}^{p-1}\beta_j\alpha_j\right) \\
= \; & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\Sigma_1 - A_1 + A_1)(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
\leq \; & (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')A_1(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
& +\lambda_{\max}(\Sigma_1 - A_1)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(I_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\
= \; & (a_1 - r_1)\sum_{j=1}^{p-1}\beta_j^2 + o(p)\sum_{j=1}^{p-1}\beta_j^2 \\
= \; & o(p)\sum_{j=1}^{p-1}\beta_j^2.
\end{aligned}$$

Therefore,

$$\sum_{j=1}^{p-1}\beta_j\alpha_j = o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right). \tag{A.4}$$

Plugging (A.2), (A.3), (A.4) into (A.1), we have

$$\begin{aligned} Q &= \ln\Big(|A_1|/|A_2|\Big) + \Big[1/(a_1-r_1) - 1/(a_2-r_2)\Big]\Big[(p-1)(a_1-r_1) + o_p(p)\Big] \\ &\quad + O_p(1) - \Big[2/(a_2-r_2)\Big] o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) \\ &\quad - \sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2) - (\alpha_p+\beta_p)^2/\Big[a_2+(p-1)r_2\Big] \\ &= (p-1)\left\{1-(a_1-r_1)/(a_2-r_2) + \ln\left[(a_1-r_1)/(a_2-r_2)\right]\right\} \\ &\quad + \ln\Big\{\left[a_1+(p-1)r_1\right]/\left[a_2+(p-1)r_2\right]\Big\} + o_p(p) + O_p(1) + o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) \\ &\quad - \sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2) - (\alpha_p+\beta_p)^2/\left[a_2+(p-1)r_2\right]. \end{aligned} \tag{A.5}$$

According to conditions 2.1 and 2.3, $|1-(a_1-r_1)/(a_2-r_2)| > 2\delta_0/c$ and for $p\to\infty$,

$$(p-1)\left[1-(a_1-r_1)/(a_2-r_2) + \ln\Big((a_1-r_1)/(a_2-r_2)\Big)\right] \to -\infty \tag{A.6}$$

at the order of p. If $\sum_{j=1}^{p-1}\beta_j^2 = O(p)$, then $o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)$ is dominated by (A.6). On the other hand, if $\sum_{j=1}^{p-1}\beta_j^2$ has the order of $p^{1+\epsilon}$ for some $\epsilon>0$, then $o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)$ is dominated by $\sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2)$. All the other terms in (A.5) are either negative or dominated by (A.6). Thus, we conclude that $Q<0$ when p is sufficiently large, and the probability of misclassifying $\mathbf{x}$ from class 1 to class 2,

$$\mathbb{P}\left(Q>0|\mathbf{x}\in\mathcal{C}_1\right)\to 0, \text{ as } p\to\infty.$$

It can be proved in a similar fashion that the probability of misclassifying $\mathbf{x}$ from class 2 to class 1 also converges to 0. This finishes the proof. □

Remark A.1. *Now we discuss how Condition 2.4-2 can be relaxed. To achieve asymptotically perfect classification, we want Q in (A.5) to be negative for large p, for which (A.6)*

is critical but guaranteed by Condition 2.4-2. Alternatively, if $(\mu_{\mathbf{1}} - \mu_{\mathbf{2}})$ *is not so close to the origin such that* $\sum_{j=1}^{p-1} \beta_j^2/(a_2 - r_2)$ *can dominate the other terms in (A.6), then* Q *can still be negative for large* p *with Condition 2.4-2 being relaxed.*

In summary, the Condition 2.4-2 on covariance matrices is sufficient for ppQDA to achieve the property of asymptotically perfect classification. However, such property could also be attributed to distinct location parameters with Condition 2.4-2 being relaxed.

The following lemma shows that pQDA with true parameters also enjoys the property of asymptotically perfect classification. The proof is similar to that of the previous lemma but much simpler due to its simpler structure of the pQDA function than that of the ppQDA function.

Lemma A.3. *Let* Q_0 *be the pQDA function with true parameters. Under conditions 2.1 and 2.5,*

$$\lim_{p\to\infty} R_{0,p} = \lim_{p\to\infty} \mathbb{P}(Q_0 > 0|\mathbf{x} \in \mathcal{C}_1) + \mathbb{P}(Q_0 \leq 0|\mathbf{x} \in \mathcal{C}_2) = 0.$$

Proof. Similar to the proof of Lemma A.2, the quadratic classification function with true parameters can be expressed as

$$\begin{aligned} Q_0 &= p\ln(a_1/a_2) + (1/a_1 - 1/a_2)(\mathbf{x} - \boldsymbol{\mu}_1)'(\mathbf{x} - \boldsymbol{\mu}_1) \\ &\quad -2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\mathbf{x} - \boldsymbol{\mu}_1)/a_2 - ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2/a_2. \end{aligned} \tag{A.7}$$

We can show that

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu}_1)'(\mathbf{x} - \boldsymbol{\mu}_1) &= tr(\Sigma_1) + O_p\left[\sqrt{tr(\Sigma_1^2)}\right] \\ &= pa_1 + O_p(\sqrt{p}) \end{aligned} \tag{A.8}$$

and

$$\begin{aligned} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(\mathbf{x} - \boldsymbol{\mu}_1) &= O_p\left[\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}\right] \\ &= O_p\Big(||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||\Big). \end{aligned} \tag{A.9}$$

The final equality in (A.8) and (A.9) is due to Condition 2.5-1.

Plugging (A.8) and (A.9) into (A.7), we have

$$\begin{aligned} Q_0 = \; & p\Big[1 - a_1/a_2 + \ln\left(a_1/a_2\right)\Big] + O_p(\sqrt{p}) \\ & + O_p\Big(||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||\Big) - ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2/a_2. \end{aligned} \tag{A.10}$$

Under (B.2), it can be shown that $Q_0 < 0$ when p is sufficiently large, i.e.,

$$\mathbb{P}\left(Q_0 > 0 | \mathbf{x} \in \mathcal{C}_1\right) \to 0.$$

Similarly, we can prove tht $\mathbb{P}\left(Q_0 \leq 0 | \mathbf{x} \in \mathcal{C}_2\right) \to 0$. This finishes the proof. □

Remark A.2. *Bounded eigenvalues of* Σ_1 *assure that* $\sqrt{tr(\Sigma_i^2)} = O(\sqrt{p})$ *in (A.8).*

The following lemma presents the estimation accuracy of various estimators, and will be repeatedly used in our proof of the asymptotically perfect classification property for the proposed ppQDA function.

Lemma A.4. *Let* $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}, \Sigma)$*, where the* $p \times p$ *covariance matrix* Σ *is symmetric and positive definite. Define* $a = tr(\Sigma)/p$ *and* $r = [Su(\Sigma) - tr(\Sigma)]/[p(p-1)]$*, i.e., the average of diagonal and off-diagonal entries of* Σ*, respectively. Let* $\widehat{\boldsymbol{\mu}}$ *and* $\widehat{\Sigma}$ *denote the sample mean and sample covariance matrix, i.e.,* $\widehat{\boldsymbol{\mu}} = \sum_{k=1}^n \mathbf{y}_k/n$ *and* $\widehat{\Sigma} = \sum_{k=1}^n (\mathbf{y}_k - \widehat{\boldsymbol{\mu}})(\mathbf{y}_k - \widehat{\boldsymbol{\mu}})'/(n-1)$*. Let* $\widehat{a} = tr(\widehat{\Sigma})/p$ *and* $\widehat{r} = [Su(\widehat{\Sigma}) - tr(\widehat{\Sigma})]/[p(p-1)]$*. Given* $a - r > \delta > 0$ *for some* $\delta > 0$ *and Condition 2.1, we have*

$$tr(\widehat{\Sigma}) = tr(\Sigma) + O_p\left(\sqrt{tr(\Sigma^2)/n}\right), \tag{A.11}$$

$$Su(\widehat{\Sigma}) = Su(\Sigma) + O_p\left(\sqrt{Su^2(\Sigma)/n}\right), \tag{A.12}$$

$$\begin{aligned} \widehat{a} - \widehat{r} &= a - r + O_p\left(p^{-1}\sqrt{tr(\Sigma^2)/n} + p^{-2}\sqrt{Su^2(\Sigma)/n}\right) \\ &= a - r + O_p\left(n^{-1/2}\right), \end{aligned} \tag{A.13}$$

$$\widehat{a} + (p-1)\widehat{r} = a + (p-1)r + O_p\left(p^{-1}\sqrt{Su^2(\Sigma)/n}\right), \tag{A.14}$$

$$\begin{aligned} (\widehat{a} - \widehat{r})^{-1} &= (a-r)^{-1} + O_p\left(p^{-1}\sqrt{tr(\Sigma^2)/n} + p^{-2}\sqrt{Su^2(\Sigma)/n}\right) \\ &= (a-r)^{-1} + O_p\left(n^{-1/2}\right), \end{aligned} \tag{A.15}$$

$$\left[\widehat{a} + (p-1)\widehat{r}\right]^{-1} = [a + (p-1)r]^{-1} + O_p\left\{n^{-1/2}[a + (p-1)r]^{-1}\right\}. \tag{A.16}$$

Proof. To prove (A.11), it can be shown that

$$tr(\widehat{\Sigma}) = \frac{1}{n-1}\left[\sum_{k=1}^{n}(\mathbf{y}_k - \boldsymbol{\mu})'(\mathbf{y}_k - \boldsymbol{\mu}) - n(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right],$$

in which

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{n}(\mathbf{y}_k - \boldsymbol{\mu})'(\mathbf{y}_k - \boldsymbol{\mu})\right] &= ntr(\Sigma) \\
\mathbb{E}\left[(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right] &= tr(\Sigma)/n \\
Var\left[\sum_{k=1}^{n}(\mathbf{y}_k - \boldsymbol{\mu})'(\mathbf{y}_k - \boldsymbol{\mu})\right] &= 2ntr(\Sigma^2) \\
Var\left[(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right] &= 2tr(\Sigma^2)/n^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
tr(\widehat{\Sigma}) &= \frac{1}{n-1}\left\{ntr(\Sigma) + O_p\left[\sqrt{ntr(\Sigma^2)}\right] - tr(\Sigma) + O_p\left[\sqrt{tr(\Sigma^2)}\right]\right\} \\
&= tr(\Sigma) + O_p\left[\sqrt{tr(\Sigma^2)/n}\right].
\end{aligned}$$

To prove (A.12), it can be shown that

$$Su(\widehat{\Sigma}) = \frac{1}{n-1}\sum_{k=1}^{n}\mathbf{1_p}'(\mathbf{y}_k - \widehat{\boldsymbol{\mu}})(\mathbf{y}_k - \widehat{\boldsymbol{\mu}})'\mathbf{1_p},$$

for which $\mathbb{E}\left[Su(\widehat{\Sigma})\right] = Su(\Sigma)$ and $Var\left[Su(\widehat{\Sigma})\right] = 2Su^2(\Sigma)/(n-1)$. Thus,

$$Su(\widehat{\Sigma}) = Su(\Sigma) + O_p\left[\sqrt{Su^2(\Sigma)/n}\right].$$

According to (A.11) and (A.12), (A.13) and (A.14) follow directly. In addition,

$$\widehat{a} - a = O_p\left[p^{-1}\sqrt{tr(\Sigma^2)/n}\right]$$

and

$$\widehat{r} - r = O_p\left(p^{-2}\left[\sqrt{Su^2(\Sigma)/n} + \sqrt{tr(\Sigma^2)/n}\right]\right).$$

Due to Condition 2.1, we have $\widehat{a}-a=o_p(1)$ and $\widehat{r}-r=o_p(1)$. Therefore, the consistency of $\widehat{a}$ and $\widehat{r}$ is proved.

To prove (A.15), by Taylor expansion,

$$\begin{aligned}(\widehat{a}-\widehat{r})^{-1} &= (a-r)^{-1}+(a-r)^{-2}O_p\left[p^{-1}\sqrt{tr(\Sigma^2)/n}+p^{-2}\sqrt{Su^2(\Sigma)/n}\right]\\ &= (a-r)^{-1}+O_p\left[p^{-1}\sqrt{tr(\Sigma^2)/n}+p^{-2}\sqrt{Su^2(\Sigma)/n}\right].\end{aligned}$$

To prove (A.16), define $D=\{[\widehat{a}+(p-1)\widehat{r}]-[a+(p-1)r]\}\left[a+(p-1)r\right]^{-1}$. According to (A.14), it can be shown that $D=O_p(n^{-1/2})$. By Taylor expansion,

$$\begin{aligned}[\widehat{a}+(p-1)\widehat{r}]^{-1} &= [a+(p-1)r]^{-1}+[a+(p-1)r]^{-1}\sum_{l=1}^{\infty}(-1)^l D^l\\ &= [a+(p-1)r]^{-1}+O_p\left\{n^{-1/2}[a+(p-1)r]^{-1}\right\}.\end{aligned}$$

This finishes the proof. □

Proof of Theorem 2.1. In Lemma A.2, we show that $\mathbb{P}\left(Q>0|\mathbf{x}\in\mathcal{C}_1\right)\to 0$ where Q is the ppQDA function with true parameters, though true parameters are unknown in practice. Next, we prove the asymptotically perfect classification property for the proposed ppQDA function (with the estimators of unknown parameters plugged in), i.e.,

$$\widehat{Q}=\ln\left(|\widehat{A}_1|/|\widehat{A}_2|\right)+(\mathbf{x}-\widehat{\boldsymbol{\mu}}_1)'\widehat{A}_1^{-1}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_1)-(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2)'\widehat{A}_2^{-1}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2).$$

Once again, we focus on the probability of misclassifying $\mathbf{x}$ from class 1 to class 2, i.e., $\mathbb{P}\left(\widehat{Q}>0|\mathbf{x}\in\mathcal{C}_1\right)$. The main strategy is to show that $\widehat{Q}-Q$ can be dominated by Q, which leads to $\mathbb{P}\left(\widehat{Q}>0|\mathbf{x}\in\mathcal{C}_1\right)=\mathbb{P}\left(\widehat{Q}-Q+Q>0|\mathbf{x}\in\mathcal{C}_1\right)\to 0$ when p is sufficiently large. We start by examining those three terms in $\widehat{Q}$ separately.

First of all, we focus on $\ln\left(|\widehat{A}_1|/|\widehat{A}_2|\right)$ in $\widehat{Q}$.

$$\begin{aligned}\ln\left(|\widehat{A}_1|/|\widehat{A}_2|\right) &= (p-1)\left[\ln\left(\widehat{a}_1-\widehat{r}_1\right)-\ln\left(\widehat{a}_2-\widehat{r}_2\right)\right]\\ &\quad+\ln\left[\widehat{a}_1+(p-1)\widehat{r}_1\right]-\ln\left[\widehat{a}_2+(p-1)\widehat{r}_2\right],\end{aligned}$$

where according to Taylor expansion, (A.13) and (A.14), for $i = 1, 2$,

$$\ln(\widehat{a}_i - \widehat{r}_i) = \ln(a_i - r_i) + (a_i - r_i)^{-1} O_p\left[p^{-1}\sqrt{tr(\Sigma_i^2)/n_i} + p^{-2}\sqrt{Su^2(\Sigma_i)/n_i}\right]$$

and

$$\begin{aligned}\ln[\widehat{a}_i + (p-1)\widehat{r}_i] &= \ln[a_i + (p-1)r_i] \\ &\quad + [a_i + (p-1)r_i]^{-1} O_p\left[p^{-1}\sqrt{Su^2(\Sigma_i)/n_i}\right].\end{aligned}$$

Therefore,

$$\begin{aligned}\ln(\widehat{a}_1 - \widehat{r}_1) - \ln(\widehat{a}_2 - \widehat{r}_2) &= \ln(a_1 - r_1) - \ln(a_2 - r_2) + O_p(n^{-1/2}) \\ \ln[\widehat{a}_1 + (p-1)\widehat{r}_1] - \ln[\widehat{a}_2 + (p-1)\widehat{r}_2] &= \ln[a_1 + (p-1)r_1] - \ln[a_2 + (p-1)r_2] \\ &\quad + O_p(n^{-1/2}).\end{aligned}$$

In summary,

$$\ln\left(|\widehat{A}_1|/|\widehat{A}_2|\right) = \ln(|A_1|/|A_2|) + O_p(pn^{-1/2}). \tag{A.17}$$

Secondly, we focus on $(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)'\widehat{A}_1^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)$ in $\widehat{Q}$.

$$\begin{aligned}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)'\widehat{A}_1^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1) &= (\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)'T\widehat{\Lambda}_1^{-1}T'(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1) \\ &= (\widehat{a}_1 - \widehat{r}_1)^{-1}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)'(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1) + \Big\{[\widehat{a}_1 + (p-1)\widehat{r}_1]^{-1} \\ &\quad -(\widehat{a}_1 - \widehat{r}_1)^{-1}\Big\}(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)'\left(\frac{1}{p}\mathbf{1}_p\mathbf{1}_p'\right)(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1). \\ &\equiv (\widehat{a}_1 - \widehat{r}_1)^{-1} \cdot \mathrm{I} \\ &\quad + p^{-1}\Big\{[\widehat{a}_1 + (p-1)\widehat{r}_1]^{-1} - (\widehat{a}_1 - \widehat{r}_1)^{-1}\Big\} \cdot \mathrm{II}.\end{aligned} \tag{A.18}$$

As $\widehat{\boldsymbol{\mu}}_i$ is the sample mean, let $\widehat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i + \widehat{\boldsymbol{\epsilon}}_i$ for $i = 1, 2$, then $\widehat{\boldsymbol{\epsilon}}_i \sim N(\mathbf{0}, \Sigma_i/n_i)$. We consider I and II in (A.18) separately, where

$$\begin{aligned}\mathrm{I} &= (\mathbf{x} - \widehat{\boldsymbol{\mu}}_1)'(\mathbf{x} - \widehat{\boldsymbol{\mu}}_1) \\ &= (\mathbf{x} - \boldsymbol{\mu}_1)'(\mathbf{x} - \boldsymbol{\mu}_1) - 2(\mathbf{x} - \boldsymbol{\mu}_1)'\widehat{\boldsymbol{\epsilon}}_1 + \widehat{\boldsymbol{\epsilon}}_1'\widehat{\boldsymbol{\epsilon}}_1,\end{aligned}$$

in which

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{x}-\boldsymbol{\mu}_1) &= tr(\Sigma_1)+O_p\left[\sqrt{tr(\Sigma_1^2)}\right]\\ (\mathbf{x}-\boldsymbol{\mu}_1)'\widehat{\boldsymbol{\epsilon}}_1 &= O_p\left[\sqrt{tr(\Sigma_1^2)/n_1}\right]\\ \widehat{\boldsymbol{\epsilon}}_1'\widehat{\boldsymbol{\epsilon}}_1 &= tr(\Sigma_1)/n_1+O_p\left[\sqrt{tr(\Sigma_1^2)}/n_1\right]\end{aligned}$$

Hence,

$$\mathrm{I}=(\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{x}-\boldsymbol{\mu}_1)+O_p\left[\sqrt{tr(\Sigma_1^2)/n_1}\right]+tr(\Sigma_1)/n_1.$$

In addition,

$$\begin{aligned}\mathrm{II}=&(\mathbf{x}-\widehat{\boldsymbol{\mu}}_1)'\left(\mathbf{1}_p\mathbf{1}_p'\right)(\mathbf{x}-\widehat{\boldsymbol{\mu}}_1)\\ =&(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1)-2(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'\widehat{\boldsymbol{\epsilon}}_1+\widehat{\boldsymbol{\epsilon}}_1'\mathbf{1}_p\mathbf{1}_p'\widehat{\boldsymbol{\epsilon}}_1,\end{aligned}$$

in which

$$\begin{aligned}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1) &= O_p\left[Su(\Sigma_1)\right]\\ (\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'\widehat{\boldsymbol{\epsilon}}_1 &= O_p\left[\sqrt{Su^2(\Sigma_1)/n_1}\right]\\ \widehat{\boldsymbol{\epsilon}}_1'\mathbf{1}_p\mathbf{1}_p'\widehat{\boldsymbol{\epsilon}}_1 &= O_p\left[\sqrt{Su^2(\Sigma_1)/n_1^2}\right].\end{aligned}$$

Hence,

$$\mathrm{II}=(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1)+O_p(\sqrt{Su^2(\Sigma_1)/n_1}).$$

According to I, II, and Lemma A.4 ((A.15) and (A.16) specifically), (A.18) becomes

$$\begin{aligned}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_1)'\widehat{A}_1^{-1}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_1) &= \left[(a_1-r_1)^{-1}+O_p(n^{-1/2})\right]\cdot \mathrm{I} + p^{-1}\Big\{\left[a_1+(p-1)r_1\right]^{-1} \\ &\quad +O_p\left[n^{-1/2}[a_1+(p-1)r_1]^{-1}\right]-(a_1-r_1)^{-1}+O_p\left(n^{-1/2}\right)\Big\}\cdot \mathrm{II} \\ &= (a_1-r_1)^{-1}\Big[(\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{x}-\boldsymbol{\mu}_1) \\ &\quad -p^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1)\Big] \\ &\quad +p^{-1}[a_1+(p-1)r_1]^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)'\mathbf{1}_p\mathbf{1}_p'(\mathbf{x}-\boldsymbol{\mu}_1) \\ &\quad +O_p(pn_1^{-1/2}) \\ &= (\mathbf{x}-\boldsymbol{\mu}_1)'A_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+O_p(pn_1^{-1/2}). \qquad (A.19)\end{aligned}$$

Thirdly, we focus on $(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2)'\widehat{A}_2^{-1}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2)$ in $\widehat{Q}$.

$$\begin{aligned}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2)'\widehat{A}_2^{-1}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2) &= (\widehat{a}_2-\widehat{r}_2)^{-1}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2) \\ &\quad +[\widehat{a}_2+(p-1)\widehat{r}_2]^{-1}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2)'\left(p^{-1}\mathbf{1}_p\mathbf{1}_p'\right)(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2) \\ &\equiv (\widehat{a}_2-\widehat{r}_2)^{-1}\cdot \mathrm{III}+[\widehat{a}_2+(p-1)\widehat{r}_2]^{-1}\cdot \mathrm{IV}. \qquad (A.20)\end{aligned}$$

We consider III and IV separately. First of all,

$$\begin{aligned}\mathrm{III} &= (\mathbf{x}-\widehat{\boldsymbol{\mu}}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2) \\ &= (\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_2)-2(\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')\widehat{\boldsymbol{\epsilon}}_2 \\ &\quad +\widehat{\boldsymbol{\epsilon}}_2'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')\widehat{\boldsymbol{\epsilon}}_2 \\ &\equiv \mathrm{III}_1-2\cdot\mathrm{III}_2+\mathrm{III}_3,\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}(\mathrm{III}_1) =& \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_2)\right] \\ =& tr\left[(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_1\right]+(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\ =& (p-1)(a_1-r_1)+\sum_{j=1}^{p-1}\beta_j^2.\end{aligned}$$

With the techniques in the derivation of (A.2) and (A.4), we have

$$\begin{aligned} Var(\mathrm{III}_1) &= 2tr\left[(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_1(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_1\right] \\ &\quad +4(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_1(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &\leq 2\left[tr(\Sigma_1^2) - 2p^{-1}Su(\Sigma_1^2) + p^{-2}Su^2(\Sigma_1)\right] + o\left(p\sum_{j=1}^{p-1}\beta_j^2\right) \\ &= o(p^2) + o\left(p\sum_{j=1}^{p-1}\beta_j^2\right). \end{aligned}$$

Hence,

$$\mathrm{III}_1 = (p-1)(a_1 - r_1) + \sum_{j=1}^{p-1}\beta_j^2 + o_p(p) + o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right).$$

In addition,

$$\mathbb{E}(\mathrm{III}_2) = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\widehat{\boldsymbol{\epsilon}}_2\right] = 0.$$

By the techniques in the derivation of (A.4) and Condition 2.1, we have

$$\begin{aligned} Var(\mathrm{III}_2) &= Var\left[(\mathbf{x} - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\widehat{\boldsymbol{\epsilon}}_2\right] \\ &= n_2^{-1}tr\Big\{\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' + \Sigma_1\right](I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Big\} \\ &= n_2^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &\quad +n_2^{-1}tr\left[\Sigma_1(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\right] \\ &\leq o\left(n^{-1}p\sum_{j=1}^{p-1}\beta_j^2\right) + n_2^{-1}\big[tr(\Sigma_1\Sigma_2) - p^{-1}Su(\Sigma_1\Sigma_2) - p^{-1}Su(\Sigma_2\Sigma_1) \\ &\quad +p^{-2}Su(\Sigma_1)Su(\Sigma_2)\big] \\ &= o\left(n^{-1}p\sum_{j=1}^{p-1}\beta_j^2\right) + O(p^2/n), \end{aligned}$$

Hence,

$$\mathrm{III}_2 = o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) + O_p(pn^{-1/2})$$

Last but not least,

$$\mathbb{E}(\mathrm{III}_3) = \mathbb{E}\left[\widehat{\boldsymbol{\epsilon}}_2'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\widehat{\boldsymbol{\epsilon}}_2\right] = n_2^{-1}(p-1)(a_2 - r_2).$$

By Condition 2.1,

$$\begin{aligned} Var(\mathrm{III}_3) &= Var\left[\widehat{\boldsymbol{\epsilon}}_2'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\widehat{\boldsymbol{\epsilon}}_2\right] \\ &= 2n_2^{-2}tr\left[(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')\Sigma_2\right] \\ &= 2n_2^{-2}\left[tr(\Sigma_2^2) - 2p^{-1}Su(\Sigma_2^2) + p^{-2}Su^2(\Sigma_2)\right] \\ &= O\left(p^2/n^2\right). \end{aligned}$$

Hence,

$$\mathrm{III}_3 = n_2^{-1}(p-1)(a_2 - r_2) + O_p\left(p/n\right)$$

Combining III_1, III_2 and III_3, we have

$$\begin{aligned} \mathrm{III} &= (\mathbf{x} - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x} - \boldsymbol{\mu}_2) - 2\cdot\mathrm{III}_2 + \mathrm{III}_3 \\ &= (\mathbf{x} - \boldsymbol{\mu}_2)'(I_p - p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x} - \boldsymbol{\mu}_2) + o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) + O_p(pn^{-1/2}) \\ &\quad + n_2^{-1}(p-1)(a_2 - r_2). \end{aligned}$$

Secondly, we focus on IV,

$$\begin{aligned} \mathrm{IV} &= (\mathbf{x} - \widehat{\boldsymbol{\mu}}_2)'\left(p^{-1}\mathbf{1}_p\mathbf{1}_p'\right)(\mathbf{x} - \widehat{\boldsymbol{\mu}}_2) \\ &= \left[p^{-1/2}\mathbf{1}_p'(\mathbf{x} - \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 - \widehat{\boldsymbol{\epsilon}}_2)\right]^2 \\ &= (\alpha_p + \beta_p - \mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2)^2, \end{aligned}$$

in which $(\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2)^2 = O_p\big\{[a_2 + (p-1)r_2]/n_2\big\}$, so that

$$\mathrm{IV} = (\alpha_p + \beta_p)^2 - 2(\alpha_p + \beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2 + O_p\big\{[a_2 + (p-1)r_2]/n_2\big\}.$$

Plugging III and IV in (A.20), we have

$$
\begin{aligned}
& (\mathbf{x}-\widehat{\boldsymbol{\mu}}_2)'\widehat{A}_2^{-1}(\mathbf{x}-\widehat{\boldsymbol{\mu}}_2) \\
= \; & (\widehat{a}_2-\widehat{r}_2)^{-1}\cdot \mathrm{III} + [\widehat{a}_2+(p-1)\widehat{r}_2]^{-1}\cdot \mathrm{IV} \\
= \; & \Big[(a_2-r_2)^{-1}+O_p(n^{-1/2})\Big]\Bigg[(\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_2) \\
& +o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right)+O_p(pn^{-1/2})+n_2^{-1}(p-1)(a_2-r_2)\Bigg] \\
& +\Bigg\{\,[a_2+(p-1)r_2]^{-1}+O_p\Big(n^{-1/2}[a_2+(p-1)r_2]^{-1}\Big)\Bigg\}\Bigg[(\alpha_p+\beta_p)^2 \\
& -2(\alpha_p+\beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2+O_p\Big([a_2+(p-1)r_2]/n_2\Big)\Bigg] \\
= \; & (\mathbf{x}-\boldsymbol{\mu}_2)'A_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)+O_p\left(n^{-1/2}\sum_{j=1}^{p-1}\beta_j^2\right) \\
& +O_p\left(pn^{-1/2}\right)+o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) \\
& +O_p\left\{n^{-1/2}\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)^2\right\} \\
& +O_p\left\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2\right\}. \qquad \text{(A.21)}
\end{aligned}
$$

where

$$
\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu}_2)'A_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \;=\; & (a_2-r_2)^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)'(I_p-p^{-1}\mathbf{1}_p\mathbf{1}_p')(\mathbf{x}-\boldsymbol{\mu}_2) \\
& +\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)^2.
\end{aligned}
$$

Based on (A.17), (A.19),and (A.21), we have

$$
\begin{aligned}
\widehat{Q}-Q \;=\; & O_p\left(n^{-1/2}\sum_{j=1}^{p-1}\beta_j^2\right)+O_p\left(pn^{-1/2}\right)+o_p\left(n^{-1/2}\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) \\
& +O_p\left\{n^{-1/2}\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)^2\right\} \\
& +O_p\left\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2\right\}. \qquad \text{(A.22)}
\end{aligned}
$$

Recall (A.5), in which

$$\begin{aligned} Q &= (p-1)\left\{1-(a_1-r_1)/(a_2-r_2)+\ln\left[(a_1-r_1)/(a_2-r_2)\right]\right\} \\ &\quad +\ln\Big\{\left[a_1+(p-1)r_1\right]/\left[a_2+(p-1)r_2\right]\Big\}+o_p(p)+O_p(1)+o_p\left(\sqrt{p\sum_{j=1}^{p-1}\beta_j^2}\right) \\ &\quad -\sum_{j=1}^{p-1}\beta_j^2/(a_2-r_2)-(\alpha_p+\beta_p)^2/\left[a_2+(p-1)r_2\right] \end{aligned} \tag{A.23}$$

Comparing (A.22) with (A.23), to show that $\widehat{Q}-Q$ is dominated by Q, we need to consider the last term in (A.22) only, i.e., $O_p\left\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2\right\}$. Notice that all other terms in (A.22) are dominated by the leading negative terms in (A.23). It can be shown that

$$\begin{aligned} \mathbb{E}\Big\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2\Big\} &= 0, \\ Var\Big\{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2\Big\} &= \left[a_2+(p-1)r_2\right]^{-2} \\ &\quad \cdot\Big\{\left[Su(\Sigma_1)/p+\beta_p^2\right]\left[Su(\Sigma_2)/(pn_2)\right]\Big\}. \end{aligned}$$

That is, given that $Su(\Sigma_i)=pa_i+p(p-1)r_i$ for $i=1,2$, we have

$$\begin{aligned} \left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2 &= O_p\left\{\sqrt{[\beta_p^2+a_1+(p-1)r_1][a_2+(p-1)r_2]^{-1}/n_2}\right\} \\ &= O_p\left\{n^{-1/2}p^{1/2}|\beta_p|\right\}+O_p\left(pn^{-1/2}\right). \end{aligned}$$

The second equality is by conditions 2.1 and 2.4-1. If $|\beta_p|=O(\sqrt{p})$, the above reduces to $O_p(pn^{-1/2})$ and is dominated by the leading negative terms in (A.23). Otherwise, if $|\beta_p|$ has the order of $p^{1/2+\epsilon}$, for some $\epsilon>0$, then

$$\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2 = o_p\left\{n^{-1/2}[a_2+(p-1)r_2]^{-1}(\alpha_p+\beta_p)^2\right\}, \tag{A.24}$$

where the right-hand side already appears in (A.22) and is dominated by the leading negative terms in (A.23). To show (A.24), notice that

$$\frac{\left[a_2+(p-1)r_2\right]^{-1}(\alpha_p+\beta_p)\mathbf{t}_p'\widehat{\boldsymbol{\epsilon}}_2}{n^{-1/2}[a_2+(p-1)r_2]^{-1}(\alpha_p+\beta_p)^2}=\frac{O_p\left[n^{-1/2}\sqrt{a_2+(p-1)r_2}\right]}{n^{-1/2}\left\{\beta_p+O_p\left[\sqrt{a_1+(p-1)r_1}\right]\right\}},$$

which tends to 0 when p is sufficiently large.

This finishes the proof. □

Proof of Theorem 2.2. The proof is similar to the proof of Theorem 2.1 and is omitted. □

A.2 Proof of Theorem 2.3

Next, we prove the asymptotically perfect classification property of $\widehat{Q}_{\widehat{h},0}$, the proposed Se-pQDA rule, which involves estimated parameters and estimated transformation functions. We begin by dealing with $Q_{\widehat{h},0}$, the Se-pQDA rule with true parameters but estimated transformation functions; and proceed to prove that the error introduced by the estimated transformation functions does not affect the convergence of the misclassification probability of $Q_{h,0}$, the Se-pQDA rule with true parameters and true transformation functions; we then return to consider $\widehat{Q}_{\widehat{h},0}$.

Without loss of generality, we use class 1 training data to estimate the transformation functions. Hence, for $\mathbf{x} \in \mathcal{C}_1$, we have $h_j(x_j) \sim N(0,1),\ j = 1, \cdots, p$, and $\boldsymbol{\mu}_1 = \mathbb{E}[\mathbf{h}(\mathbf{x})] = \mathbf{0}$. With a slight abuse of notation, the estimated and true marginal CDF's of class 1 are denoted by $\widehat{F}_j(\cdot)$ and $F_j(\cdot)$ respectively.

Notice that the pQDA rule with true parameters assigns $\mathbf{x}$ to class 1 if $Q_0 \leq 0$ and to class 2 otherwise, where

$$\begin{aligned} Q_0 &= p \ln(a_1/a_2) + a_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)'(\mathbf{x}-\boldsymbol{\mu}_1) - a_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)'(\mathbf{x}-\boldsymbol{\mu}_2) \\ &= p \ln(a_1/a_2) + a_1^{-1}\sum_{j=1}^{p}(x_j-\mu_{1j})^2 - a_2^{-1}\sum_{j=1}^{p}(x_j-\mu_{2j})^2 \\ &= \left(a_1^{-1}-a_2^{-1}\right)\sum_{j=1}^{p}(x_j-\eta_j)^2 + C, \end{aligned}$$

in which $\boldsymbol{\eta} = (a_1^{-1}-a_2^{-1})^{-1}(a_1^{-1}\boldsymbol{\mu}_1 - a_2^{-1}\boldsymbol{\mu}_2)$ and

$$C = p\ln(a_1/a_2) + a_1^{-1}\boldsymbol{\mu}_1'\boldsymbol{\mu}_1 - a_2^{-1}\boldsymbol{\mu}_2'\boldsymbol{\mu}_2 - (a_1^{-1}-a_2^{-1})\sum_{j=1}^{p}\eta_j^2.$$

For the Se-pQDA rule, we essentially apply the pQDA rule on the transformed data. If we plug in the true transformation functions and true parameters, the Se-pQDA function $Q_{h,0}$ becomes Q_0 for the transformed data, where

$$Q_{h,0} = \left(a_1^{-1} - a_2^{-1}\right) \sum_{j=1}^{p} \left[h_j(x_j) - \eta_j\right]^2 + C,$$

If we plug in the estimated transformation functions but true parameters, the Se-pQDA function becomes

$$Q_{\widehat{h},0} = \left(a_1^{-1} - a_2^{-1}\right) \sum_{j=1}^{p} \left[\widehat{h}_j(x_j) - \eta_j\right]^2 + C.$$

The corresponding misclassification probability can be expressed as

$$\mathbb{P}\left(Q_{\widehat{h},0} > 0 | \mathbf{x} \in \mathcal{C}_1\right).$$

We have shown that the pQDA function Q_0 (or $Q_{h,0}$ for transformed data) enjoys the property of asymptotically perfect classification. To show that $Q_{\widehat{h},0}$ enjoys the same property, we are to compare $\sum_{j=1}^{p} \left[\widehat{h}_j(x_j) - \eta_j\right]^2$ in $Q_{\widehat{h},0}$ with $\sum_{j=1}^{p} \left[h_j(x_j) - \eta_j\right]^2$ in $Q_{h,0}$.

The following inequalities regarding the normal distribution are repeatedly used in our proof.

Proposition A.1. *Let $\phi(t)$ and $\Phi(t)$ be the pdf and cdf of $N(0,1)$, then we have*

(a) for $t \geq 1$,

$$\frac{\phi(t)}{2t} \leq 1 - \Phi(t) \leq \frac{\phi(t)}{t};$$

(b) for $t \geq 0.99$,

$$\Phi^{-1}(t) \leq \sqrt{2 \ln\left(\frac{1}{1-t}\right)};$$

The following lemma shows that $\left|\widehat{h}_j(x_j) - \eta_j\right|^2$ is close to $\left|h_j(x_j) - \eta_j\right|^2$ for $h_j(x_j) \in A_n$.

Lemma A.5. *For some $0 < \gamma_1 < 1$, let $A_n = \left[-\sqrt{\gamma_1 \ln n}, \sqrt{\gamma_1 \ln n}\right]$. When n is sufficiently large, for any $\epsilon > 0$, we have for $j = 1, \cdots, p$,*

$$\begin{aligned} & \mathbb{P}\left\{ \sup_{h_j(x_j) \in A_n} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - \left[h_j(x_j) - \eta_j\right]^2 \right| > \epsilon \right\} \\ \leq \; & 2\exp\left\{ -n^{1-\gamma_1} \left[C_1 \pi^2 \gamma_1 \ln n \ln\left(4n^{\frac{\gamma_1}{2}} \sqrt{2\pi\gamma_1 \ln n}\right)\right]^{-1} \epsilon^2 \right\} \\ & + 2\exp\left[-n^{1-\gamma_1} (C_2 \pi \gamma_1 \ln n)^{-1}\right], \end{aligned}$$

where C_1 and C_2 are some positive constants.

Proof. By mean value theorem,

$$\left[\widehat{h}_j(x_j) - \eta_j\right]^2 - \left[h_j(x_j) - \eta_j\right]^2 = \; 2\left[\Phi^{-1}(\xi) - \eta_j\right] \left(\Phi^{-1}\right)'(\xi) \left[\widehat{F}_j(x_j) - F_j(x_j)\right],$$

for some $\xi \in \left[\min\left(\widehat{F}_j(x_j), F_j(x_j)\right), \max\left(\widehat{F}_j(x_j), F_j(x_j)\right)\right]$.

To show that $\left|\widehat{h}_j(x_j) - \eta_j\right|^2$ is close to $\left|h_j(x_j) - \eta_j\right|^2$ for $h_j(x_j) \in A_n$, first of all, we bound $\left|\left(\Phi^{-1}\right)'(\xi)\right|$. By considering the range of $F_j(x_j)$ and $\widehat{F}_j(x_j)$ for $h_j(x_j) \in A_n$, Mai and Zou (2015) show that, with probability no less than $1 - 2\exp\left[-n^{1-\gamma_1}/(16\pi\gamma_1 \ln n)\right]$,

$$n^{-\gamma_1/2} / \left[4(2\pi\gamma_1 \ln n)^{1/2}\right] \leq \xi \leq 1 - n^{-\gamma_1/2} / \left[4(2\pi\gamma_1 \ln n)^{1/2}\right]. \tag{A.25}$$

In conjunction with Proposition A.1, it can be shown that

$$\left|\left(\Phi^{-1}\right)'(\xi)\right| = \left\{\phi\left[\Phi^{-1}(\xi)\right]\right\}^{-1} \leq 8\pi n^{\gamma_1/2} \sqrt{\gamma_1 \ln n}.$$

Next, we bound $|\Phi^{-1}(\xi) - \eta_j|$. Due to (A.25) and Proposition A.1, with probability no less than $1 - 2\exp\left[-n^{1-\gamma_1}(16\pi\gamma_1 \ln n)^{-1}\right]$,

$$\begin{aligned} \left|\Phi^{-1}(\xi) - \eta_j\right| &\leq \left|\Phi^{-1}(\xi)\right| + |\eta_j| \\ &\leq \sqrt{2\ln\left(4n^{\gamma_1/2}\sqrt{2\pi\gamma_1 \ln n}\right)} + |\eta_j|. \end{aligned}$$

As $|\eta_j|$'s do not diverge with n, we bound the following product, when n is sufficiently large,

$$\begin{aligned} 2\left|\Phi^{-1}(\xi)-\eta_j\right|\left|\left(\Phi^{-1}\right)'(\xi)\right| &\leq 32\sqrt{\ln\left(4n^{\gamma_1/2}\sqrt{2\pi\gamma_1\ln n}\right)}\left(\pi n^{\gamma_1/2}\sqrt{\gamma_1\ln n}\right) \\ &\equiv M_n^*. \end{aligned}$$

Therefore,

$$\begin{aligned} & \mathbb{P}\left\{\sup_{h_j(x_j)\in A_n}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2-\left[h_j(x_j)-\eta_j\right]^2\right|>\epsilon\right\} \\ \leq\ & \mathbb{P}\left[M_n^*\sup_{h_j(x_j)\in A_n}\left|\widehat{F}_j(x_j)-F_j(x_j)\right|>\epsilon\right] \\ & +2\exp\left[-n^{1-\gamma_1}(16\pi\gamma_1\ln n)^{-1}\right]. \end{aligned} \tag{A.26}$$

The probability involving M_n^* on the right hand side,

$$\begin{aligned} & \mathbb{P}\left[M_n^*\sup_{h_j(x_j)\in A_n}\left|\widehat{F}_j(x_j)-F_j(x_j)\right|>\epsilon\right] \\ \leq\ & \mathbb{P}\left[M_n^*\sup_{h_j(x_j)\in A_n}\left|\widehat{F}_j(x_j)-\widetilde{F}_j(x_j)\right|>\epsilon/2\right] \\ & +\mathbb{P}\left[M_n^*\sup_{h_j(x_j)\in A_n}\left|F_j(x_j)-\widetilde{F}_j(x_j)\right|>\epsilon/2\right]. \end{aligned} \tag{A.27}$$

As $\sup_{h_j(x_j)\in A_n}\left|\widehat{F}_j(x_j)-\widetilde{F}_j(x_j)\right|\leq 1/n^2$ by definition and $M_n^*/n^2\to 0$, the first probability on the right hand side of (A.27) is 0 when n is sufficiently large. The second probability,

$$\begin{aligned} & \mathbb{P}\left[M_n^*\sup_{h_j(x)\in A_n}\left|F_j(x)-\widetilde{F}_j(x)\right|>\epsilon/2\right] \\ \leq\ & 2\exp\left\{-2n\left[\epsilon/(2M_n^*)\right]^2\right\} \\ \leq\ & 2\exp\left\{-n^{1-\gamma_1}\epsilon^2\left[C_1\pi^2\gamma_1\ln n\ln\left(4n^{\gamma_1/2}\sqrt{2\pi\gamma_1\ln n}\right)\right]^{-1}\right\}, \end{aligned} \tag{A.28}$$

where C_1 is a positive constant and the first inequality is from Dvoretzky-Kiefer-Wolfowitz (DKW) inequality.

Combining (A.26), (A.27) and (A.28), we finish the proof. □

Lemma A.5 shows that $\left|\widehat{h}_j(x_j) - \eta_j\right|^2$ is close to $|h_j(x_j) - \eta_j|^2$ for $h_j(x_j) \in A_n$. Next we focus on A_n^c, which will be partitioned into three regions. For some positive constants $0 < \gamma_1 < 1$, $\gamma_2 > 0$ and $\gamma_3 > 0$, we define:

$$\begin{aligned} B_n &= [-\gamma_2 \ln n, -\sqrt{\gamma_1 \ln n}) \cup (\sqrt{\gamma_1 \ln n}, \gamma_2 \ln n]; \\ C_n &= [-n^{\gamma_3}, -\gamma_2 \ln n) \cup (\gamma_2 \ln n, n^{\gamma_3}]; \\ D_n &= (-\infty, -n^{\gamma_3}) \cup (n^{\gamma_3}, +\infty). \end{aligned}$$

Although the regions are similar to those in Mai and Zou (2015), we consider how many components of a new obsevation fall into each region to establish the accuracy of the QDA rule that depends on the estimated transformation ($Q_{\widehat{h},0}$), whereas they considered how many samples (of a particular dimension) fall into each region to establish the accuracy of estimated parameters. This major difference is discussed in detail later.

Lemma A.6. *Let $\rho_{j_1 j_2}$ be the correlation between $h_{j_1}(x_{j_1})$ and $h_{j_2}(x_{j_2})$, for $j_1, j_2 = 1, 2, \ldots, p$, and $\rho = \max\{0, \max_{j_1 \neq j_2}(\rho_{j_1 j_2})\}$. Let α_1 and α_2 be positive constants such that $\alpha_1 > 1 - \gamma_1/[2(\rho+1)]$. Define $\#B_n = \#\{j : h_j(x_j) \in B_n\}$, i.e., the number of marginal random variables $h_j(x_j)$'s that fall into B_n, and C_n, D_n analogously. For sufficiently large n, we have*

$$\sup_{h_j(x_j) \in B_n} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - [h_j(x_j) - \eta_j]^2 \right| \leq \left(2\sqrt{\ln n} + c_6\right)^2 + (\gamma_2 \ln n + c_6)^2; \tag{A.29}$$

$$\sup_{h_j(x_j) \in C_n} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - [h_j(x_j) - \eta_j]^2 \right| \leq \left(2\sqrt{\ln n} + c_6\right)^2 + (n^{\gamma_3} + c_6)^2; \tag{A.30}$$

$$\mathbb{P}(\#B_n > pn^{\alpha_1-1}) = O\left\{n^{2\left[1-\alpha_1-\frac{\gamma_1}{2(1+\rho)}\right]}\left[(\ln n)\left(1-n^{1-\alpha_1-\gamma_1/2}\right)^2\right]^{-1}\right\}; \tag{A.31}$$

$$\mathbb{P}(\#C_n > pn^{\alpha_2-1}) = O\left\{\frac{p^{-1}(\gamma_2 \ln n)\exp\left[-\frac{(\gamma_2 \ln n)^2}{2}\right] + \exp\left[-\frac{(\gamma_2 \ln n)^2}{\rho+1}\right]}{n^{2\alpha_2-2}\left(\gamma_2 \ln n\right)^2\left[1-n^{1-\alpha_2}\exp\left(-\frac{(\gamma_2 \ln n)^2}{2}\right)/\left(\gamma_2 \ln n\right)\right]^2}\right\}; \tag{A.32}$$

$$\mathbb{P}(\#D_n > p/n) = O\left\{\frac{p^{-1}n^{2-\gamma_3}\exp\left(-\frac{n^{2\gamma_3}}{2}\right) + n^{2-2\gamma_3}\exp\left(-\frac{n^{2\gamma_3}}{1+\rho}\right)}{\left[1-n^{1-\gamma_3}\exp\left(-\frac{n^{2\gamma_3}}{2}\right)\right]^2}\right\}. \tag{A.33}$$

Proof. Inequalities (A.29) and (A.30) are because the range of $\widehat{h}_j(x_j)$ is decided by its definition and Proposition A.1 and the range of $h_j(x_j)$ is decided by the definitions of B_n and C_n. To be more specific about $\widehat{h}_j(x_j)$,

$$\left|\widehat{h}_j(x_j)\right| \leq \Phi^{-1}(1-1/n^2) \leq 2\sqrt{\ln n}.$$

Now we prove (A.31). Let $w_j = 1_{\{h_j(x_j)\in B_n\}}$ be the indicator of whether $h_j(x_j)$ is in B_n. Then the probability of $h_j(x_j)$ falling into B_n is

$$p_j = \mathbb{P}\left[h_j(x_j) \in B_n\right] = \mathbb{E}(w_j).$$

Similarly, the probability of both $h_{j_1}(x_{j_1})$ and $h_{j_2}(x_{j_2})$ falling into B_n is defined as

$$p_{j_1 j_2} = \mathbb{P}\left[h_{j_1}(x_{j_1}) \in B_n, h_{j_2}(x_{j_2}) \in B_n\right] = \mathbb{E}(w_{j_1} w_{j_2}).$$

To examine the order of $\mathbb{P}(\#B_n > pn^{\alpha_1-1})$, we now focus on p_j and $p_{j_1 j_2}$ which are both useful for bounding $\mathbb{P}(\#B_n > pn^{\alpha_1-1})$ as shown later.

For p_j, because of normality, the definition of B_n and Proposition A.1, when n is sufficiently large,

$$p_j \leq 2\left[1-\Phi\left(\sqrt{\gamma_1 \ln n}\right)\right] \leq \sqrt{2}n^{-\gamma_1/2}/\sqrt{\pi\gamma_1 \ln n} \leq n^{-\gamma_1/2}.$$

For $p_{j_1 j_2}$, consider the following bivariate normal random vector

$$\begin{bmatrix} h_{j_1}(x_{j_1}) \\ h_{j_2}(x_{j_2}) \end{bmatrix} \sim N \left[\mathbf{0}, \begin{pmatrix} 1 & \rho_{j_1 j_2} \\ \rho_{j_1 j_2} & 1 \end{pmatrix} \right].$$

Then,

$$\begin{aligned} p_{j_1 j_2} &\leq 4\, \mathbb{P}\left[h_{j_1}(x_{j_1}) > \sqrt{\gamma_1 \ln n},\ h_{j_2}(x_{j_2}) > \sqrt{\gamma_1 \ln n} \right] \\ &\leq (1-\rho_{j_1 j_2})^{-2} (1-\rho_{j_1 j_2}^2)^{3/2} (\gamma_1 \ln n)^{-1} \exp\left(-\frac{\gamma_1 \ln n}{1+\rho_{j_1 j_2}} \right) \\ &\leq (1-\rho)^{-2} (\gamma_1 \ln n)^{-1} \exp\left(-\frac{\gamma_1 \ln n}{1+\rho} \right), \end{aligned}$$

where the second inequality is due to the bound of Mill's ratio for multivariate normal distribution (Savage, 1962; Hashorva and Hüsler, 2003). Thus,

$$\begin{aligned} \mathbb{P}(\#B_n > pn^{\alpha_1 - 1}) &= \mathbb{P}\left(\sum_{j=1}^{p} w_j > pn^{\alpha_1 - 1} \right) \\ &\leq \mathbb{P}\left(\sum_{j=1}^{p} w_j - \sum_{j=1}^{p} p_j > pn^{\alpha_1 - 1} - pn^{-\gamma_1/2} \right) \\ &\leq \mathbb{E}\left[\left(\sum_{j=1}^{p} w_j - \sum_{j=1}^{p} p_j \right)^2 \right] (pn^{\alpha_1 - 1} - pn^{-\frac{\gamma_1}{2}})^{-2}. \end{aligned}$$

Now we focus on the expectation on the right hand side of the previous inequality,

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{j=1}^{p} w_j - \sum_{j=1}^{p} p_j\right)^2\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{p} w_j\right)^2 + \left(\sum_{j=1}^{p} p_j\right)^2 - 2\left(\sum_{j=1}^{p} w_j\right)\left(\sum_{j=1}^{p} p_j\right)\right] \\
&= \left[\sum_{j=1}^{p} p_j + 2\sum_{j_1<j_2} p_{j_1 j_2} - \left(\sum_{j=1}^{p} p_j\right)^2\right] \\
&\leq \left[\sqrt{2}pn^{-\gamma_1/2}\left(\sqrt{\pi\gamma_1 \ln n}\right)^{-1}\right. \\
&\quad \left. + p(p-1)(1-\rho)^{-2}(\gamma_1 \ln n)^{-1}\exp\left(-\frac{\gamma_1 \ln n}{1+\rho}\right)\right] \\
&= O\left[pn^{-\gamma_1/2}\left(\sqrt{\pi\gamma_1 \ln n}\right)^{-1} + p^2(\ln n)^{-1}n^{-\gamma_1/(1+\rho)}\right] \\
&= O\left[p^2(\ln n)^{-1}n^{-\gamma_1/(1+\rho)}\right].
\end{aligned}$$

The last equality is because the ratio between the first and second item in the right hand side of the second last equality tends to 0, i.e.,

$$\frac{pn^{-\gamma_1/2}\left(\sqrt{\ln n}\right)^{-1}}{p^2(\ln n)^{-1}n^{-\gamma_1/(1+\rho)}} = p^{-1}(\ln n)^{1/2}n^{\frac{\gamma_1(1-\rho)}{2(1+\rho)}} \to 0.$$

Now we bound $\mathbb{P}(\#B_n > pn^{\alpha_1-1})$,

$$\begin{aligned}
\mathbb{P}(\#B_n > pn^{\alpha_1-1}) &\leq \mathbb{E}\left[\left(\sum_{j=1}^{p} w_j - \sum_{j=1}^{p} p_j\right)^2\right](pn^{\alpha_1-1} - pn^{-\gamma_1/2})^{-2}. \\
&= O\left[n^{-\gamma_1/(1+\rho)}(\ln n)^{-1}(n^{\alpha_1-1} - n^{-\gamma_1/2})^{-2}\right] \\
&= O\left\{n^{2\left[1-\alpha_1-\frac{\gamma_1}{2(1+\rho)}\right]}\left[(\ln n)\left(1 - n^{1-\alpha_1-\gamma_1/2}\right)^2\right]^{-1}\right\}.
\end{aligned}$$

The above right hand side is desired and tends to 0 because it is assumed that $\alpha_1 > 1 - \gamma_1/[2(1+\rho)]$. The proof of $\mathbb{P}(\#C_n > pn^{\alpha_2-1})$ and $\mathbb{P}(\#D_n > p/n)$ is similar and omitted. This finishes the proof. □

The next lemma shows that $Q_{\widehat{h},0}$, the Se-pQDA rule with estimated transformation functions but true parameters, enjoys the property of asymptotically perfect classification.

Lemma A.7. *Under Condition 2.6, if conditions 2.1, 2.3 and 2.5 hold for the transformed data, and* $p\exp\left(-n^{1-\gamma_1}/\ln^2 n\right) \to 0$, *then,*

$$\lim_{p\to\infty,n\to\infty} R_{\widehat{h},0} = \lim_{p\to\infty,n\to\infty} \mathbb{P}(Q_{\widehat{h},0} > 0|\mathbf{x}\in\mathcal{C}_1) + \mathbb{P}(Q_{\widehat{h},0} \le 0|\mathbf{x}\in\mathcal{C}_2) = 0.$$

Proof. Define $\mathcal{A}$, the collection of index j such that $h_j(x_j) \in A_n$, i.e.,

$$\mathcal{A} = \{j|h_j(x_j)\in A_n\},$$

and $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ analogously. For any $\epsilon > 0$,

$$\begin{aligned}
& \mathbb{P}\left\{p^{-1}\left|\sum_{j=1}^{p}\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \sum_{j=1}^{p}\left[h_j(x)-\eta_j\right]^2\right| > \epsilon\right\} \\
\le\ & \mathbb{P}\left\{p^{-1}\sum_{j=1}^{p}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \left[h_j(x_j)-\eta_j\right]^2\right| > \epsilon\right\} \\
\le\ & \mathbb{P}\left\{p^{-1}\#A_n \max_{j\in\mathcal{A}}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \left[h_j(x_j)-\eta_j\right]^2\right| > \epsilon/4\right\} \\
& +\mathbb{P}\left\{p^{-1}\#B_n \max_{j\in\mathcal{B}}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \left[h_j(x_j)-\eta_j\right]^2\right| > \epsilon/4\right\} \\
& +\mathbb{P}\left\{p^{-1}\#C_n \max_{j\in\mathcal{C}}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \left[h_j(x_j)-\eta_j\right]^2\right| > \epsilon/4\right\} \\
& +\mathbb{P}\left\{p^{-1}\#D_n \max_{j\in\mathcal{D}}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \left[h_j(x_j)-\eta_j\right]^2\right| > \epsilon/4\right\} \\
\le\ & \mathbb{P}\left\{\max_{j\in\mathcal{A}}\sup_{h_j(x_j)\in A_n}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \left[h_j(x_j)-\eta_j\right]^2\right| > \epsilon/4\right\} \\
& +\mathbb{P}\left\{p^{-1}\#B_n \max_{j\in\mathcal{B}}\sup_{h_j(x_j)\in B_n}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \left[h_j(x_j)-\eta_j\right]^2\right| > \epsilon/4\right\} \\
& +\mathbb{P}\left\{p^{-1}\#C_n \max_{j\in\mathcal{C}}\sup_{h_j(x_j)\in C_n}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \left[h_j(x_j)-\eta_j\right]^2\right| > \epsilon/4\right\} \\
& +\mathbb{P}\left\{p^{-1}\#D_n \max_{j\in\mathcal{D}}\left|\left[\widehat{h}_j(x_j)-\eta_j\right]^2 - \left[h_j(x_j)-\eta_j\right]^2\right| > \epsilon/4\right\}. \qquad \text{(A.34)}
\end{aligned}$$

We require $\alpha_1 < 1$ and $2\gamma_3 + \alpha_2 < 1$. By inequality (A.29) and (A.30) in Lemma A.6, if $\#B_n \leq pn^{\alpha_1 - 1}$, $\#C_n \leq pn^{\alpha_2 - 1}$ and n is sufficiently large,

$$p^{-1}\#B_n \max_{j\in\mathcal{B}} \sup_{h_j(x_j)\in B_n} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - [h_j(x_j) - \eta_j]^2 \right| \leq \epsilon/4,$$

$$p^{-1}\#C_n \max_{j\in\mathcal{C}} \sup_{h_j(x_j)\in C_n} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - [h_j(x_j) - \eta_j]^2 \right| \leq \epsilon/4;$$

therefore,

$$\begin{aligned}
& \mathbb{P}\left\{ p^{-1}\#B_n \max_{j\in\mathcal{B}} \sup_{h_j(x_j)\in B_n} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - [h_j(x_j) - \eta_j]^2 \right| > \epsilon/4 \right\} \\
\leq\ & \mathbb{P}(\#B_n > pn^{\alpha_1 - 1}), \qquad \text{(A.35)} \\
& \mathbb{P}\left\{ p^{-1}\#C_n \max_{j\in\mathcal{C}} \sup_{h_j(x_j)\in C_n} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - [h_j(x_j) - \eta_j]^2 \right| > \epsilon/4 \right\} \\
\leq\ & \mathbb{P}(\#C_n > pn^{\alpha_2 - 1}). \qquad \text{(A.36)}
\end{aligned}$$

For the probability involving D_n, when $\#D_n \leq p/n$ and n is sufficiently large,

$$\begin{aligned}
& \mathbb{P}\left\{ p^{-1}\#D_n \max_{j\in\mathcal{D}} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - [h_j(x_j) - \eta_j]^2 \right| > \epsilon/4 \right\} \\
\leq\ & \mathbb{P}\left\{ n^{-1}\left[\left(2\sqrt{\ln n} + c_7\right)^2 + \max_{j\in\mathcal{D}} \left(h_j(x_j) - \eta_j\right)^2 \right] > \epsilon/4 \right\} \\
\leq\ & \mathbb{P}\left[n^{-1/2} \max_{j\in\mathcal{D}} |h_j(x_j) - \eta_j| > \sqrt{\epsilon/8} \right] \\
\leq\ & \mathbb{P}\left[n^{-1/2} \max_{j\in\mathcal{D}} |h_j(x_j)| > \sqrt{\epsilon}/4 \right] \\
\leq\ & \sum_{j=1}^{p} \mathbb{P}\Big[|h_j(x_j)| > \sqrt{n\epsilon}/4 \Big] \\
=\ & 2p\Big[1 - \Phi\left(\sqrt{n\epsilon}/4\right) \Big] \\
\leq\ & (2p)\Big[4/(2\pi n\epsilon)^{1/2} \Big] \exp\left(-n\epsilon/32\right) \\
=\ & 4\sqrt{2}(\pi\epsilon)^{-1/2} p n^{-1/2} \exp\left(-n\epsilon/32\right). \qquad \text{(A.37)}
\end{aligned}$$

The last inequality is due to Proposition 1 for sufficiently large n. In addition, the far right hand side in (A.37) tends to 0 due to the assumption of $\ln p = o(n)$.

When n is sufficiently large, with Lemma A.5, Lemma A.6, (A.35), (A.36) and (A.37), we have

$$\begin{aligned}
& \mathbb{P}\left\{p^{-1}\left|\sum_{j=1}^{p}\left[\widehat{h}_j(x_j)-\eta_j\right]^2-\sum_{j=1}^{p}\left[h_j(x)-\eta_j\right]^2\right|>\epsilon\right\} \\
\leq\ & 2p\exp\left\{-n^{1-\gamma_1}\left[C_1\pi^2\gamma_1\ln n\ln\left(4n^{\gamma_1/2}\sqrt{2\pi\gamma_1\ln n}\right)\right]^{-1}\epsilon^2\right\} \\
& +2p\exp\left\{-n^{1-\gamma_1}(C_2\pi\gamma_1\ln n)^{-1}\right\}+\mathbb{P}(\#B_n>pn^{\alpha_1-1}) \\
& +\mathbb{P}(\#C_n>pn^{\alpha_2-1})+\mathbb{P}(\#D_n>p/n) \\
& +4\sqrt{2}(\pi\epsilon)^{-1/2}pn^{-1/2}\exp\left(-\epsilon n/32\right) \qquad \text{(A.38)} \\
\equiv\ & P'.
\end{aligned}$$

Notice that P' tends to 0 when $p\to\infty$.

For $Q_{\widehat{h},0}$, the Se-pQDA function with estimated transformation functions but true parameters, the probability of misclassifying $\mathbf{x}$ from class 1 to class 2 can be expressed as the following

$$\begin{aligned}
& P\left(Q_{\widehat{h},0}>0|\mathbf{x}\in\mathcal{C}_1\right) \\
=\ & \mathbb{P}\left\{\left(a_1^{-1}-a_2^{-1}\right)\sum_{j=1}^{p}\left[\widehat{h}_j(x_j)-\eta_j\right]^2+C>0\middle|\mathbf{x}\in\mathcal{C}_1\right\} \\
\leq\ & \mathbb{P}\left\{\left(a_1^{-1}-a_2^{-1}\right)\sum_{j=1}^{p}\left[h_j(x_j)-\eta_j\right]^2+C+p\left|a_1^{-1}-a_2^{-1}\right|\epsilon>0\middle|\mathbf{x}\in\mathcal{C}_1\right\}+P' \\
=\ & \mathbb{P}\Big[Q_{h,0}+p\left|a_1^{-1}-a_2^{-1}\right|\epsilon>0|\mathbf{x}\in\mathcal{C}_1\Big]+P'. \qquad \text{(A.39)}
\end{aligned}$$

Notice that $Q_{h,0}$, the Se-pQDA function with true transformation functions and true parameters, is equivalent to Q_0, the p-QDA rule in (A.10). We have shown that Q_0 tends to negative infinity at the order of at least p. We can choose a small $\epsilon>0$ so that $p\left|a_1^{-1}-a_2^{-1}\right|\epsilon$ is dominated by the leading negative terms in Q_0. For example, ϵ can be chosen so that $\left|a_1^{-1}-a_2^{-1}\right|\epsilon<c_5$.

Notice that $P\left(Q_{\widehat{h},0}>0|\mathbf{x}\in\mathcal{C}_1\right)$ is only one-side misclassification probability with $\widehat{h}$ being estimated from the class 1 training data. With the current $\widehat{h}$, a transformed class 2

observation obviously does not follow standard normal distribution marginally. Hence, the proof for $P\left(Q_{\widehat{h},0} \leq 0 | \mathbf{x} \in \mathcal{C}_2\right) \to 0$ needs to be modified from that of $P\left(Q_{\widehat{h},0} > 0 | \mathbf{x} \in \mathcal{C}_1\right) \to 0$. Similar to the construction of A_n, B_n, C_n and D_n when proving $P\left(Q_{\widehat{h},0} > 0 | \mathbf{x} \in \mathcal{C}_1\right) \to 0$, we construct the following regions in order to prove $P\left(Q_{\widehat{h},0} \leq 0 | \mathbf{x} \in \mathcal{C}_2\right) \to 0$.

$$\begin{aligned}
A_{nj} &= \Big[-\sigma_{2j}\sqrt{\gamma_1 \ln n} + \mu_{2j}, \sigma_{2j}\sqrt{\gamma_1 \ln n} + \mu_{2j} \Big]; \\
B_{nj} &= \Big[-\sigma_{2j}\gamma_2 \ln n + \mu_{2j}, -\sigma_{2j}\sqrt{\gamma_1 \ln n} + \mu_{2j} \Big) \\
&\quad \cup \Big(\sigma_{2j}\sqrt{\gamma_1 \ln n} + \mu_{2j}, \sigma_{2j}\gamma_2 \ln n + \mu_{2j} \Big]; \\
C_{nj} &= \Big[-\sigma_{2j} n^{\gamma_3} + \mu_{2j}, -\sigma_{2j}\gamma_2 \ln n + \mu_{2j} \Big) \\
&\quad \cup \Big(\sigma_{2j}\gamma_2 \ln n + \mu_{2j}, \sigma_{2j} n^{\gamma_3} + \mu_{2j} \Big]; \\
D_{nj} &= \Big(-\infty, -\sigma_{2j} n^{\gamma_3} + \mu_{2j} \Big) \cup \Big(\sigma_{2j} n^{\gamma_3} + \mu_{2j}, +\infty \Big). \qquad \text{(A.40)}
\end{aligned}$$

We first show that $\left|\widehat{h}_j(x_j) - \eta_j\right|^2$ is close to $|h_j(x_j) - \eta_j|^2$ for $h_j(x_j) \in A_{nj}$. Define $\gamma_1^* = \gamma_1(\sigma_{\max} + b_1)^2$, where $\sigma_{\max} = \max_{1 \leq j \leq p} \sigma_{2j}$ and b_1 is some positive constant. Let

$$A_n^* = \Big[-\sqrt{\gamma_1^* \ln n}, \sqrt{\gamma_1^* \ln n} \Big].$$

Then for sufficiently large n, $A_{nj} \subset A_n^*$ for all j, and

$$\begin{aligned}
& \mathbb{P}\left\{ \sup_{h_j(x_j) \in A_{nj}} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - \left[h_j(x_j) - \eta_j\right]^2 \right| > \epsilon \right\} \\
\leq\ & \mathbb{P}\left\{ \sup_{h_j(x_j) \in A_n^*} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - \left[h_j(x_j) - \eta_j\right]^2 \right| > \epsilon \right\}
\end{aligned}$$

Then for $0 < \gamma_1^* < 1$ and sufficiently large n,

$$\begin{aligned}
& \mathbb{P}\left\{ \sup_{h_j(x_j) \in A_{nj}} \left| \left[\widehat{h}_j(x_j) - \eta_j\right]^2 - \left[h_j(x_j) - \eta_j\right]^2 \right| > \epsilon \right\} \\
\leq\ & 2\exp\left\{ -n^{1-\gamma_1^*}\epsilon^2 \Big[C_1 \pi^2 \gamma_1^* \ln n \ln\left(4n^{\gamma_1^*/2}\sqrt{2\pi\gamma_1^* \ln n} \right) \Big]^{-1} \right\} \\
& + 2\exp\left[-n^{1-\gamma_1^*}(C_2 \pi \gamma_1^* \ln n)^{-1} \right].
\end{aligned}$$

The proof follows that of Lemma A.5 by replacing γ_1 with γ_1^*.

The proof of Lemma A.6 and Lemma A.7 alike for $B_{nj},\ C_{nj}$, and D_{nj} can be slightly modified from that of Lemma A.6 and Lemma A.7 for $B_n,\ C_n$, and D_n. Notice that scale and location change doesn't affect the order of the bounds in (A.29), (A.30) and (A.37). To bound $\#B_{nj}$ as in (A.31), notice that $h_j(x_j) \in B_{nj}$ is equivalent to

$$\sigma_{2j}^{-1}\left[h_j(x_j) - \mu_{2j}\right] \in \left[-\gamma_2 \ln n, -\sqrt{\gamma_1 \ln n}\right) \cup \left(\sqrt{\gamma_1 \ln n}, \gamma_2 \ln n\right],$$

where $\sigma_{2j}^{-1}\left[h_j(x_j) - \mu_{2j}\right] \sim N(0,1)$, so the proof follows. Bound $\#C_{nj}$ and $\#D_{nj}$ as in (A.32) and (A.33).

As for $0 < \gamma_1^* < 1$, if $(\sigma_{\max} + b_1) \leq 1$ then no extra step needs to be taken; otherwise, given other positive constants, we need to have $0 < \gamma_1(\sigma_{\max}+b_1)^2 < 1$ instead of $0 < \gamma_1 < 1$ in order to show $P\left(Q_{\widehat{h},0} \leq 0 | \mathbf{x} \in \mathcal{C}_2\right) \to 0$.

This finishes the proof. □

We now proceed to show that $\widehat{Q}_{\widehat{h},0}$, the proposed Se-pQDA rule (with estimated transformation functions and estimated parameters) also enjoys the property of asymptotically perfect classification. Its performance will be dependent upon not only the accuracy of estimated transformation functions $\widehat{h}_j(\cdot)$'s but also the accuracy of estimated parameters.

To investigate the effect of parameter estimation, we now ignore the class label for brevity. We assume that transformed data follow a multivariate normal distribution, i.e. $h(\mathbf{y}_k) \overset{i.i.d.}{\sim} N(\mu, \Sigma)$, $k = 1, \cdots, n$. Denote $\widehat{h}_j = \Phi^{-1} \circ \widehat{F}_j$, where $\widehat{F}_j$ is defined as in Section 2.3; denote, for the jth dimension, $\mu_j = \mathbb{E}[h_j(y_{jk})]$ and $\widehat{\mu}_j = (1/n)\sum_{k=1}^n \widehat{h}_j(y_{jk})$ as the true and estimated mean respectively; $\sigma_j^2 = Var[h_j(y_{jk})]$ and $\widehat{\sigma}_j^2 = (1/n)\sum_{k=1}^n \left[\widehat{h}_j(y_{jk}) - \widehat{\mu}_j\right]^2$ as the true and estimated variance respectively.

Notice that estimating $h_j's$ based on the class 1 training data ensures that after transformation the marginal distributions of class 1 data are $N(0,1)$; hence, it seems unnecessary to estimate μ_j and σ_j^2 for the transformed class 1 data. However, the estimated means and variances of the transformed class 2 data need to be examined. The following result on class 1 offers us insight on how estimated transformation functions affect the parameter estimation.

We present without proof, in the following proposition, some results from Mai and Zou (2015). Notice that, Proposition A.2 holds for every $j \in \{1, \cdots, p\}$.

Proposition A.2. *From proof of Theorem 1 in Mai and Zou (2015), for some constant C sufficiently large n and any $\epsilon > 0$,*

$$\begin{aligned} \mathbb{P}\left(|\widehat{\mu}_j - \mu_j| > \epsilon\right) &\leq \zeta_1^*(\epsilon); \\ \mathbb{P}\left(\left|\widehat{\sigma}_j^2 - \sigma_j^2\right| > \epsilon\right) &\leq \zeta_2^*(\epsilon), \end{aligned}$$

in which

$$\begin{aligned} \zeta_1^*(\epsilon) &= 2\exp(-Cn\epsilon^2) + 4\exp(-Cn^{1-\gamma_1}\epsilon^2/(\gamma_1 \ln n)) + \exp(-Cn^{2\alpha_1-1}) \\ &\quad + \exp(-Cn^{2\alpha_2-1}) + (2\pi)^{-1/2} 2\exp(-Cn^{2\gamma_3}); \\ \zeta_2^*(\epsilon) &= 2\exp(-Cn^{2\gamma_3}) + \exp(-Cn^{2\alpha_2-1}) \\ &\quad + \exp(-Cn^{2\alpha_1-1}) + 4\exp(-Cn^{1-\gamma_1}\epsilon^2/(\gamma_1^2 \ln^2 n)). \end{aligned}$$

Remark A.3. *Note that α_1, α_2, α_3, γ_1 and γ_3 are defined as in Lemma A.5 — Lemma A.7. In fact, the proof of this proposition applies similar technique. Previously, when we bound the difference between $\sum_{j=1}^p \left(\widehat{h}_j(x_j) - \eta_j\right)^2$ and $\sum_{j=1}^p \left(h_j(x_j) - \eta_j\right)^2$, we consider, across dimensions, how many components of $h(\mathbf{x})$ fall into regions A_n, B_n, C_n and D_n, respectively. Now, we bound the estimation error of mean and variance for every $j \in \{1, \cdots, p\}$; we consider, across samples, how many realizations in $\{y_{jk}, k = 1, \ldots, n\}$ fall into regions A_n, B_n, C_n and D_n, respectively.*

Remark A.4. *To summarize, the inequalities $0 < \gamma_1 < 1$, $\gamma_2 > 0$, $\gamma_3 > 0$, $\alpha_1 + \gamma_1/(2(\rho + 1)) > 1$, $\alpha_1 < 1$ and $2\gamma_3 + \alpha_2 < 1$ need to be satisfied. We can set $\gamma_1 = \theta(1 + \rho)$, $\gamma_3 = 1/6 - \theta/2$, $\alpha_1 = 1 - \theta/4$ and $\alpha_2 = 2/3$ for any $0 < \theta < 1/3$. Then,*

$$\begin{aligned} \zeta_1^*(\epsilon) &= 2\exp(-Cn\epsilon^2) + 4\exp(-Cn^{1-\theta(1+\rho)}\epsilon^2/\ln n) + \exp(-Cn^{1-\theta/2}) \\ &\quad + \exp(-Cn^{1/3}) + (2\pi)^{-1/2} 2\exp(-Cn^{1/3-\theta}); \\ \zeta_2^*(\epsilon) &= 2\exp(-Cn^{1/3-\theta}) + \exp(-Cn^{1/3}) + \exp(-Cn^{1-\theta/2}) \\ &\quad + 4\exp(-Cn^{1-\theta(1+\rho)}\epsilon^2/\ln^2 n). \end{aligned} \tag{A.41}$$

Proof of Theorem 2.3. As $h(\cdot) = \Phi^{-1} \circ F_1(\cdot)$, then $\boldsymbol{\mu}_1 = \mathbf{0}$ and $a_1 = tr(\Sigma_1)/p = 1$. Hence, $\widehat{Q}_{\widehat{h},0}$ only involves the estimates of $\boldsymbol{\mu}_2$, a_2 and $\widehat{h}_j$'s, not $\boldsymbol{\mu}_1$ and a_1. Notice that for any $\epsilon_2 > 0$,

$$\begin{aligned}\mathbb{P}\Big(\max_{1\le j\le p}|\widehat{\mu}_{2j} - \mu_{2j}| > \epsilon_2\Big) &\le \sum_{j=1}^{p}\mathbb{P}\Big(|\widehat{\mu}_{2j} - \mu_{2j}| > \epsilon_2\Big) \\ &\le p\zeta_1^*(\epsilon_2) \end{aligned} \tag{A.42}$$

and

$$\begin{aligned}\mathbb{P}\left(|\widehat{a}_2 - a_2| > \epsilon_2\right) &\le \mathbb{P}\Big(p^{-1}\sum_{j=1}^{p}\big|\widehat{\sigma}_{2j}^2 - \sigma_{2j}^2\big| > \epsilon_2\Big) \\ &\le \sum_{j=1}^{p}\mathbb{P}\left(\big|\widehat{\sigma}_{2j}^2 - \sigma_{2j}^2\big| > \epsilon_2\right) \\ &\le p\zeta_2^*(\epsilon_2). \end{aligned} \tag{A.43}$$

According to (A.41), the leading terms in the right-hand-side of (A.42) and (A.43) are both

$$p\exp(-Cn^{1/3-\theta}).$$

Thus, if $p\exp(-Cn^{1/3-\theta}) \to 0$, the right-hand-side of (A.42) and (A.43) converges to 0.

The proposed Se-pQDA function is

$$\begin{aligned}\widehat{Q}_{\widehat{h},0} &= \ln\Big(|\widehat{A}_1|/|\widehat{A}_2|\Big) + \Big[\widehat{h}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_1\Big]'\widehat{A}_1^{-1}\Big[\widehat{h}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_1\Big] - \Big[\widehat{h}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_2\Big]'\widehat{A}_2^{-1}\Big[\widehat{h}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_2\Big] \\ &= p\Big[\ln(1/\widehat{a}_2) + (1 - 1/\widehat{a}_2)\,\widehat{h}(\mathbf{x})'\widehat{h}(\mathbf{x})/p + 2\widehat{\boldsymbol{\mu}}_2'\widehat{h}(\mathbf{x})/(p\widehat{a}_2) - \widehat{\boldsymbol{\mu}}_2'\widehat{\boldsymbol{\mu}}_2/(p\widehat{a}_2)\Big].\end{aligned}$$

We now consider the above right hand side without the factor p by parts, given that $\max_{1\le j\le p}|\widehat{\mu}_{2j} - \mu_{2j}| < \epsilon_2$ and $|\widehat{a}_2 - a_2| < \epsilon_2$.

First of all,

$$\ln(1/\widehat{a}_2) \le \ln(1/a_2) + a_2^{-1}\epsilon_2 + O(\epsilon_2^2), \tag{A.44}$$

$$1 - 1/\widehat{a}_2 \le 1 - 1/a_2 + a_2^{-2}\epsilon_2 + O(\epsilon_2^2). \tag{A.45}$$

The right hand sides in (A.44) and (A.45) can be derived from Taylor expansion.

Secondly, with (A.45), we can show that

$$\begin{aligned}(1-1/\widehat{a}_2)\,\widehat{h}(\mathbf{x})'\widehat{h}(\mathbf{x})/p &\leq (1-1/a_2)\,\widehat{h}(\mathbf{x})'\widehat{h}(\mathbf{x})/p \\ &\quad + \left[a_2^{-2}\epsilon_2 + O(\epsilon_2^2)\right] 4\ln n. \end{aligned} \tag{A.46}$$

Thirdly, for any $\epsilon_3 > 0$ and sufficiently large n,

$$\begin{aligned} & \mathbb{P}\left[\left|\widehat{\boldsymbol{\mu}}_2'\widehat{h}(\mathbf{x})/(p\widehat{a}_2) - \boldsymbol{\mu}_2'\widehat{h}(\mathbf{x})/(pa_2)\right| > \epsilon_3\right] \\ \leq\ & \mathbb{P}\left[p^{-1}\left|\widehat{\boldsymbol{\mu}}_2'\widehat{h}(\mathbf{x})/\widehat{a}_2 - \widehat{\boldsymbol{\mu}}_2'\widehat{h}(\mathbf{x})/a_2\right| > \epsilon_3/2\right] \\ & + \mathbb{P}\left[p^{-1}\left|\widehat{\boldsymbol{\mu}}_2'\widehat{h}(\mathbf{x})/a_2 - \boldsymbol{\mu}_2'\widehat{h}(\mathbf{x})/a_2\right| > \epsilon_3/2\right] \\ \leq\ & \mathbb{P}\left[p^{-1}O(\epsilon_2)2\sqrt{\ln n}\sum_{j=1}^{p}\left(|\mu_{2j}| + \epsilon_2\right) > \epsilon_3/2\right] \\ & + \mathbb{P}\left(2\epsilon_2\sqrt{\ln n}/a_2 > \epsilon_3/2\right). \end{aligned} \tag{A.47}$$

Then set $\epsilon_2 = (\ln n)^{-1-\alpha}$ for some $\alpha > 0$, (A.47) tends to 0 when n is sufficiently large.

Fourthly, from (A.45),

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_2'\widehat{\boldsymbol{\mu}}_2/(\widehat{a}_2 p) &\geq (\widehat{\boldsymbol{\mu}}_2'\widehat{\boldsymbol{\mu}}_2/p)\left[1/a_2 - a_2^{-2}\epsilon_2 + O(\epsilon_2^2)\right] \\ &= \boldsymbol{\mu}_2'\boldsymbol{\mu}_2/(a_2 p) + O(\epsilon_2), \end{aligned} \tag{A.48}$$

as

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_2'\widehat{\boldsymbol{\mu}}_2/p &= \sum_{j=1}^{p}\widehat{\mu}_{2j}^2/p \\ &= \sum_{j=1}^{p}\left[\mu_{2j}^2 + 2\mu_{2j}(\widehat{\mu}_{2j} - \mu_{2j}) + (\widehat{\mu}_{2j} - \mu_{2j})^2\right]/p \\ &\geq \sum_{j=1}^{p}(\mu_{2j}^2 - |2\mu_{2j}\epsilon_2|)/p \\ &= \sum_{j=1}^{p}\mu_{2j}^2/p - 2\epsilon_2\sum_{j=1}^{p}|\mu_{2j}|/p. \end{aligned} \tag{A.49}$$

As a result of combining (A.44), (A.46), (A.47) and (A.48), the probability of misclassifying $\widehat{h}(\mathbf{x})$ from class 1 to class 2 is

$$\begin{aligned}
& P\left(\widehat{Q}_{\widehat{h},0} > 0 | \mathbf{x} \in \mathcal{C}_1\right) \\
= \; & \mathbb{P}\left[p \ln\left(1/\widehat{a}_2\right) + \left(1 - 1/\widehat{a}_2\right) \widehat{h}(\mathbf{x})' \widehat{h}(\mathbf{x}) + 2\widehat{\boldsymbol{\mu}}_2' \widehat{h}(\mathbf{x})/\widehat{a}_2 - \widehat{\boldsymbol{\mu}}_2' \widehat{\boldsymbol{\mu}}_2/\widehat{a}_2 > 0 | \mathbf{x} \in \mathcal{C}_1\right] \\
\leq \; & \mathbb{P}\left(|\widehat{a}_2 - a_2| > \epsilon_2\right) + \mathbb{P}\left(\max_{1 \leq j \leq p} |\widehat{\mu}_{2j} - \mu_{2j}| > \epsilon_2\right) \\
& + \mathbb{P}\Big[p \ln\left(1/a_2\right) + \left(1 - 1/a_2\right) \widehat{h}(\mathbf{x})' \widehat{h}(\mathbf{x}) + 2\boldsymbol{\mu}_2' \widehat{h}(\mathbf{x})/a_2 - \boldsymbol{\mu}_2' \boldsymbol{\mu}_2/a_2 \\
& + E_n > 0 | \mathbf{x} \in \mathcal{C}_1\Big] \\
\leq \; & p\zeta_1^*(\epsilon_2) + p\zeta_2^*(\epsilon_2) + \mathbb{P}\left[Q_{\widehat{h},0} + E_n > 0 | \mathbf{x} \in \mathcal{C}_1\right] \\
\leq \; & p\zeta_1^*(\epsilon_2) + p\zeta_2^*(\epsilon_2) + \mathbb{P}\left[Q_{h,0} + p\left|1 - 1/a_2\right|\epsilon + E_n > 0 | \mathbf{x} \in \mathcal{C}_1\right] + P' \qquad \text{(A.50)}
\end{aligned}$$

where $\epsilon_2 = (\ln n)^{-1-\alpha}$ for some $\alpha > 0$ and

$$E_n/p = a_2^{-1}\epsilon_2 + \left[a_2^{-2}\epsilon_2 + O(\epsilon_2^2)\right] 4 \ln n + 2\epsilon_3 + O(\epsilon_2).$$

If $p \exp(-Cn^{1/3-\theta}) \to 0$ for any $0 < \theta < 1/3$, then (A.50) goes to 0. Note that the condition in Lemma A.7 for $P' \to 0$ is satisfied because $1 - \gamma_1 = 1 - \theta/(1+\rho) > 1/3 - \theta$. We also need to choose small ϵ and ϵ_3 so that $(1 - 1/a_2)\epsilon + 2\epsilon_3$ being small in conjunction with the convergence of E_n/p ensures $\widehat{Q}_{\widehat{h},0}$ is dominated by $Q_{h,0}$ which is negative for sufficiently large p.

This proves the probability of the proposed Se-pQDA misclassifying $\widehat{h}(\mathbf{x})$ from class 1 to class 2 converges to 0. Similarly, we can prove that the other side of the misclassification probability converges to 0. This finishes the proof. □

Appendix B

Proofs of Chapter 3

B.1 Main theorems

Proof of Theorem 3.1. We use the framework of the proof for the consistency of the sparse precision matrix estimator in Rothman et al. (2008). In spite of the similar framework, our proof is essentially different from theirs in that we are to establish consistency for estimators with the "low-rank + diagonal" matrix structure.

To study the solution of the optimization problem (3.3), we firstly recall the search space,

$$\mathbf{F}_r = \{\Theta \mid L \in \mathbf{S}_+^{p,r}, D \in \mathbf{D}_{++}^p \text{ and } \Theta = -L + D\}.$$

Base on that, we define another set

$$\mathbf{E}_r = \{\Delta \mid \Delta = \Theta - \Theta_0, \Theta \in \mathbf{F}_r\},$$

which can be thought as a "centered" version of $\mathbf{F}_r$. As $r \geq r_0$ is assumed in this theorem, we straightforwardly have $\Theta_0 \in \mathbf{F}_r$ and $0 \in \mathbf{E}_r$.

Let $f(\Theta) = \mathrm{tr}(\Theta S) - \log|\Theta|$ be the value of the objective function at Θ, and $F(\Delta) = f(\Theta_0 + \Delta) - f(\Theta_0)$. Let $\widehat{\Delta}_r = \widehat{\Theta}_r - \Theta_0$, we can prove the desired result

$$\|\widehat{\Delta}_r\|_F \leq M \max(a_{n,p,r}, b_{n,p}), \tag{B.1}$$

for some constant M, by proving

$$F(\Delta) > F(0) = 0 \text{ for all } \Delta \in \mathbf{M}_{2r}, \tag{B.2}$$

in which

$$\mathbf{M}_{2r} = \mathbf{E}_{2r} \cap \{\Delta \mid \|\Delta\|_F = M \max(a_{n,p,r}, b_{n,p})\} \cap \{\Delta \mid \|\Delta\|_{op} \leq C_1\},$$

and C_1 is a constant so that $\|\widehat{\Delta}_r\|_{op} \leq C_1$ $(r = 1, \ldots, p)$. The existence of C_1 is validated by Lemma B.1.

To clarify this, we show it leads to contradiction if (B.2) is true while (B.1) is not. As $\|\widehat{\Delta}_r\|_F > M \max(a_{n,p,r}, b_{n,p})$ and $\|0\|_F < M \max(a_{n,p,r}, b_{n,p})$, there exists a real number $0 < t < 1$ so that $\|(1-t)0 + t\widehat{\Delta}_r\|_F = M \max(a_{n,p,r}, b_{n,p})$. As $\widehat{\Delta}_r \in \mathbf{E}_r$ and $0 \in \mathbf{E}_r$, we have $(1-t)0 + t\widehat{\Delta}_r \in \mathbf{E}_{2r}$. As $\|\widehat{\Delta}_r\|_{op} \leq C_1$ by Lemma B.1, we have $\|(1-t)0 + t\widehat{\Delta}_r\|_{op} \leq C_1$. Therefore, $(1-t)0 + t\widehat{\Delta}_r \in \mathbf{M}_{2r}$ and $F\{(1-t)0 + t\widehat{\Delta}_r\} > 0$ by (B.2). However, as $\widehat{\Delta}_r$ minimizes $F(\Delta)$ and $F(\widehat{\Delta}_r) \leq 0$, we also have

$$F\left\{(1-t)0 + t\widehat{\Delta}_r\right\} \leq (1-t)F(0) + tF(\widehat{\Delta}_r) \leq 0$$

by convexity of $F(\Delta)$, and this leads to contradiction.

The remaining work is to prove (B.2).

For any $\Delta \in \mathbf{M}_{2r}$, we have

$$\begin{aligned} F(\Delta) &= \operatorname{tr}\{(\Theta_0 + \Delta)S\} - \log|\Theta_0 + \Delta| - \{\operatorname{tr}(\Theta_0 S) - \log|\Theta_0|\} \\ &= \operatorname{tr}(\Delta S) - \{\log|\Theta_0 + \Delta| - \log|\Theta_0|\}. \end{aligned} \tag{B.3}$$

The bound of the second term in (B.3) is irrelevant to the assumed structure of the matrix; according to Rothman et al. (2008) and the definition of $\mathbf{M}_{2r}$.

$$\begin{aligned} \log|\Theta_0 + \Delta| - \log|\Theta_0| &\leq \operatorname{tr}(\Sigma_0\Delta) - (\|\Theta_0\|_{op} + \|\Delta\|_{op})^{-2}\|\Delta\|_F^2 \\ &\leq \operatorname{tr}(\Sigma_0\Delta) - (c_1^{-1} + C_1)^{-2}\|\Delta\|_F^2. \end{aligned} \tag{B.4}$$

We write $C_2 = (c_1^{-1} + C_1)^{-2}$. With (B.4) plugged into (B.3), we obtain

$$F(\Delta) \geq C_2\|\Delta\|_F^2 + \operatorname{tr}\{\Delta(S - \Sigma_0)\}. \tag{B.5}$$

Now we derive the bound of $\mathrm{tr}\{\Delta(S-\Sigma_0)\}$ in (B.5). We notice that any $\Delta \in \mathbf{E}_{2r}$ can be written as $\Delta = -(L-L_0)+D-D_0$, in which $-(L-L_0) \in \mathbf{S}^{p,3r}$ and $D-D_0 \in \mathbf{D}^p$. By Lemma B.2, Δ can also be decomposed as $\Delta = L_\Delta + D_\Delta$, so that $L_\Delta \in \mathbf{S}^{p,9r}$, $D_\Delta \in \mathbf{D}^p$ and $\|\Delta\|_F^2 \geq C_3\left(\|L_\Delta\|_F^2 + \|D_\Delta\|_F^2\right)$ for some constant C_3. We consider the absolute value,

$$\begin{aligned} |\,\mathrm{tr}\,\{\Delta(S-\Sigma_0)\}\,| &\leq |\,\mathrm{tr}\,\{L_\Delta(S-\Sigma_0)\}\,| + |\,\mathrm{tr}\,\{D_\Delta(S-\Sigma_0)\}\,| \\ &\leq \|L_\Delta\|_* \|S-\Sigma_0\|_{op} + \|D_\Delta\|_F \left\{\sum\nolimits_{j=1}^{p}(s_j-\sigma_{0j})^2\right\}^{1/2} \\ &\leq (9r)^{1/2}\|L_\Delta\|_F\|S-\Sigma_0\|_{op} + p^{1/2}\|D_\Delta\|_F \max_{1\leq j\leq p}|s_j-\sigma_{0j}|, \quad \text{(B.6)} \end{aligned}$$

in which s_j and σ_{0j} are the jth diagonal elements in S and Σ_0 respectively. The second inequality is because of the property of dual norm (Recht et al., 2010). The last inequality uses inequalities regarding different matrix norms (Recht et al., 2010; Rothman et al., 2008).

Under the normality assumption, with probability tending to 1, the sample covariance matrix S satisfies

$$\max_{1\leq j\leq p}|s_j-\sigma_{0j}| \leq C_4(\log p/n)^{1/2}, \quad \|S-\Sigma_0\|_{op} \leq C_4(p/n)^{1/2}, \tag{B.7}$$

for some constant C_4. The first inequality is by Lemma 1 in Rothman et al. (2008), and the second inequality is by Proposition 2.1 in Vershynin (2012).

Combine (B.6) and (B.7), we have

$$|\,\mathrm{tr}\,\{\Delta(S-\Sigma_0)\}\,| \leq C_5(\|L_\Delta\|_F + \|D_\Delta\|_F)\max(a_{n,p,r}, b_{n,p}), \tag{B.8}$$

for some constant C_5.

By (B.5), (B.8) and $\|\Delta\|_F^2 \geq C_3\left(\|L_\Delta\|_F^2 + \|D_\Delta\|_F^2\right)$,

$$\begin{aligned} F(\Delta) &\geq C_2\|\Delta\|_F^2 - C_5(\|L_\Delta\|_F + \|D_\Delta\|_F)\max(a_{n,p,r}, b_{n,p}) \\ &\geq C_2\|\Delta\|_F^2 - C_5\max(a_{n,p,r}, b_{n,p})\left\{2\left(\|L_\Delta\|_F^2 + \|D_\Delta\|_F^2\right)\right\}^{1/2} \\ &\geq C_2\|\Delta\|_F^2 - C_6\max(a_{n,p,r}, b_{n,p})\|\Delta\|_F \\ &= \|\Delta\|_F^2\left\{C_2 - C_6\max(a_{n,p,r}, b_{n,p})\|\Delta\|_F^{-1}\right\} \\ &= \|\Delta\|_F^2\left(C_2 - C_6/M\right) \\ &> 0, \end{aligned} \tag{B.9}$$

for sufficiently large constant M. Constant C_6 depends on C_3 and C_5.

This completes the proof. □

Proof of Theorem 3.2. Recall that $d_r = \min_{\Theta \in \mathbf{F}_r} \|\Theta - \Theta_0\|_F$ and Θ_r is a matrix in $\mathbf{F}_r$ so that $\|\Theta_r - \Theta_0\|_F = d_r$. As

$$\begin{aligned} \|\widehat{\Theta}_r - \Theta_0\|_F &\leq \|\widehat{\Theta}_r - \Theta_r\|_F + \|\Theta_r - \Theta_0\|_F \\ &= \|\widehat{\Theta}_r - \Theta_r\|_F + d_r \\ &= \|\widehat{\Theta}_r - \Theta_r\|_F + O\left\{\max(a_{n,p,r_0}, b_{n,p})\right\}, \end{aligned}$$

we only need to prove $\|\widehat{\Theta}_r - \Theta_r\|_F = O_p\{\max(a_{n,p,r_0}, b_{n,p})\}$.

We use similar technique as in the proof of Theorem 3.1.

Let $f(\Theta) = \mathrm{tr}(\Theta S) - \log|\Theta|$ be the value of the objective function at Θ, and $F_r(\Delta) = f(\Theta_r + \Delta) - f(\Theta_r)$. To obtain the desired result $\|\widehat{\Theta}_r - \Theta_r\|_F \leq M \max(a_{n,p,r_0}, b_{n,p})$ for some constant M, it is sufficient to prove

$$F_r(\Delta) > F_r(0) = 0 \text{ for all } \Delta \in \mathbf{M}_{2r}^r, \tag{B.10}$$

in which

$$\mathbf{M}_{2r}^r = \{\Delta \mid \Delta = \Theta - \Theta_r, \Theta \in \mathbf{F}_{2r}\} \cap \{\Delta \mid \|\Delta\|_F = M \max(a_{n,p,r_0}, b_{n,p})\} \cap \{\Delta \mid \|\Delta\|_{op} \leq C_7\}.$$

The constant C_7 is defined as follows. As

$$\|\widehat{\Theta}_r - \Theta_r\|_{op} \leq \|\widehat{\Theta}_r - \Theta_0\|_{op} + \|\Theta_r - \Theta_0\|_F \leq C_1 + d_r,$$

and $d_r \to 0$, we define $C_7 = 2C_1$ and guarantee $\|\widehat{\Theta}_r - \Theta_r\|_{op} \leq C_7$. Afterwards, the reasoning of the sufficiency of (B.10) is the same as that of the sufficiency of (B.2), and is omitted.

Now, we prove (B.10).

For any $\Delta \in \mathbf{M}_{2r}^r$, by similar argument as for (B.5) and $\|\Theta_r - \Theta_0\|_F = d_r$, with C_9 based on C_7, we have

$$\begin{aligned} F_r(\Delta) &\geq C_9\|\Delta\|_F^2 + \operatorname{tr}\{\Delta(S-\Sigma_r)\} \\ &= C_9\|\Delta\|_F^2 + \operatorname{tr}\{\Delta(S-\Sigma_0)\} + \operatorname{tr}\{\Delta(\Sigma_0-\Sigma_r)\} \\ &\geq C_9\|\Delta\|_F^2 + \operatorname{tr}\{\Delta(S-\Sigma_0)\} - \|\Delta\|_F\|\Sigma_r-\Sigma_0\|_F \\ &\geq C_9\|\Delta\|_F^2 + \operatorname{tr}\{\Delta(S-\Sigma_0)\} - C_{10}\|\Delta\|_F d_r, \end{aligned} \tag{B.11}$$

for some constant C_{10}. The second last inequality is because of Cauchy–Schwarz inequality, and the last inequality uses $\|\Sigma_r - \Sigma_0\|_F = \|\Theta_r^{-1} - \Theta_0^{-1}\|_F \leq C_{10}\|\Theta_r - \Theta_0\|_F$, which can be derived by Taylor expansion.

By similar argument as from (B.6) to (B.9), for $\Delta \in \mathbf{M}_{2r}^r$

$$|\operatorname{tr}\{\Delta(S-\Sigma_0)\}| \leq C_{11}\|\Delta\|_F \max(a_{n,p,r_0}, b_{n,p}). \tag{B.12}$$

By (B.11), (B.12) and $d_r = O(\max(a_{n,p,r_0}, b_{n,p}))$, with some constant C_{12} based on C_{10} and C_{11},

$$\begin{aligned} F_r(\Delta) &\geq C_9\|\Delta\|_F^2 - C_{12}\|\Delta\|_F \max(a_{n,p,r_0}, b_{n,p}) \\ &> 0 \end{aligned}$$

for sufficiently large M.

This completes the proof. □

Proof of Theorem 3.3. Let $f(\Theta) = \operatorname{tr}(\Theta S) - \log|\Theta|$, $\widehat{\Delta}_r = \widehat{\Theta}_r - \Theta_0$ and $F(\widehat{\Delta}_r) = f(\widehat{\Theta}_r) - f(\Theta_0)$. The objective function in (3.4) becomes $f(\widehat{\Theta}_r) + \tau(r)$ when $\operatorname{rank}(L)$ is fixed to be r.

The discussion in Section 3.4.2 shows that, the convergence rate in Theorem 3.3 is already true for $r \in \mathbf{A}_2 \cup \mathbf{A}_3 \cup \{r_0\}$. Thus, if we can prove $f(\widehat{\Theta}_r) + \tau(r) > f(\widehat{\Theta}_{r_0}) + \tau(r_0)$ for all $r \in \mathbf{A}_1 \cup \mathbf{A}_4$ so that these ranks will not be selected, the proof of the theorem will be completed.

For a particular $r \neq r_0$, $\tau(r)$ and $\tau(r_0)$ are both fixed; therefore, all we need is a lower bound of $f(\widehat{\Theta}_r) - f(\widehat{\Theta}_{r_0})$. We firstly develop a general lower bound, and then discuss $r \in \mathbf{A}_1$ and $r \in \mathbf{A}_4$ separately.

As $f(\Theta_0) \geq f(\widehat{\Theta}_{r_0})$, we have

$$f(\widehat{\Theta}_r) - f(\widehat{\Theta}_{r_0}) \geq f(\widehat{\Theta}_r) - f(\Theta_0) = F(\widehat{\Delta}_r);$$

and it is sufficient if we have a lower bound for

$$F(\widehat{\Delta}_r) = \mathrm{tr}(\widehat{\Delta}_r S) - \{\log|\Theta_0 + \widehat{\Delta}_r| - \log|\Theta_0|\}. \tag{B.13}$$

With similar argument as (B.4), we have

$$\begin{aligned} \log|\Theta_0 + \widehat{\Delta}_r| - \log|\Theta_0| &\leq \mathrm{tr}(\Sigma_0\widehat{\Delta}_r) - (\|\Theta_0\|_{op} + \|\widehat{\Delta}_r\|_{op})^{-2}\|\widehat{\Delta}\|_F^2 \\ &\leq \mathrm{tr}(\Sigma_0\widehat{\Delta}_r) - (c_1^{-1} + C_1)^{-2}\|\widehat{\Delta}_r\|_F^2. \end{aligned} \tag{B.14}$$

Just to clarify, although look alike, the bound of $\|\Delta\|_{op}$ in (B.4) is due to the definition of $\mathbf{M}_{2r}$, whereas the bound of $\|\widehat{\Delta}_r\|_{op}$ in (B.14) is because $\|\widehat{\Delta}_r\|_{op} \leq C_1$ $(r = 1, \ldots, p)$ by Lemma B.1.

Plug (B.14) into (B.13), we have

$$F(\widehat{\Delta}_r) \geq C_2\|\widehat{\Delta}_r\|_F^2 + \mathrm{tr}\{\widehat{\Delta}_r(S - \Sigma_0)\}. \tag{B.15}$$

Let $\widehat{L}_r$ and $\widehat{D}_r$ be the low-rank matrix component and diagonal matrix component of $\widehat{\Theta}_r$ respectively, we have $\widehat{\Delta}_r = -(\widehat{L}_r - L_0) + (\widehat{D}_r - D_0)$, in which $-(\widehat{L}_r - L_0) \in \mathbf{S}^{p,r+r_0}$ and $\widehat{D}_r - D_0 \in \mathbf{D}^p$. By Lemma B.2, $\widehat{\Delta}_r$ can also be written as $\widehat{\Delta}_r = L_{\widehat{\Delta}_r} + D_{\widehat{\Delta}_r}$, in which $L_{\widehat{\Delta}_r} \in \mathbf{S}^{p,3(r+r_0)}$, $D_{\widehat{\Delta}_r} \in \mathbf{D}^p$ and $\|\widehat{\Delta}_r\|_F^2 \geq C_3(\|L_{\widehat{\Delta}_r}\|_F^2 + \|D_{\widehat{\Delta}_r}\|_F^2)$.

By similar argument as (B.6) – (B.8), the second part in (B.15) can be bounded as

$$\begin{aligned} |\,\mathrm{tr}\{\widehat{\Delta}_r(S - \Sigma_0)\}| &\leq \{3(r + r_0)\}^{1/2}\,\|L_{\widehat{\Delta}_r}\|_F\|S - \Sigma_0\|_{op} + p^{1/2}\|D_{\widehat{\Delta}_r}\|_F \max_{1\leq j\leq p}|s_j - \sigma_{0j}| \\ &\leq C_{14}\|\widehat{\Delta}_r\|_F \max\{a_{n,p,(r+r_0)}, b_{n,p}\}. \end{aligned} \tag{B.16}$$

for some constant C_{14}.

Plug (B.16) into (B.15), we have

$$F(\widehat{\Delta}_r) \geq C_2\|\widehat{\Delta}_r\|_F^2 - C_{14}\|\widehat{\Delta}_r\|_F \max\{a_{n,p,(r+r_0)}, b_{n,p}\}. \tag{B.17}$$

With the general lower bound of $F(\widehat{\Delta}_r)$ obtained, we now consider $r \in \mathbf{A}_1$.

When $r \in \mathbf{A}_1$, as $r < r_0$, we replace the $a_{n,p,(r+r_0)}$ in (B.17) with a_{n,p,r_0}, and obtain

$$F(\widehat{\Delta}_r) \geq C_2\|\widehat{\Delta}_r\|_F^2 - C_{15}\|\widehat{\Delta}_r\|_F \max(a_{n,p,r_0}, b_{n,p}), \tag{B.18}$$

for some constant C_{15}. By the definition of $\mathbf{A}_1$, we can represent d_r as

$$d_r = \eta_{n,p,r_0} \max(a_{n,p,r_0}, b_{n,p})$$

for some $\eta_{n,p,r_0} \to \infty$. By the definition of the distance d_r, we have $\|\widehat{\Delta}_r\|_F \geq d_r$. With these facts, (B.18) can be simplified as

$$\begin{aligned} F(\widehat{\Delta}_r) &\geq \|\widehat{\Delta}_r\|_F^2 \left\{C_2 - C_{15}\|\widehat{\Delta}_r\|_F^{-1} \max(a_{n,p,r_0}, b_{n,p})\right\} \\ &\geq \|\widehat{\Delta}_r\|_F^2 \left(C_2 - C_{15}\eta_{n,p,r_0}^{-1}\right) \\ &\geq C_2\|\widehat{\Delta}_r\|_F^2/2 \\ &\geq C_2 d_r^2/2, \end{aligned} \tag{B.19}$$

when n and p are sufficiently large.

By (B.19) and Condition 3.3, we have

$$\begin{aligned} &\left\{f(\widehat{\Theta}_r) + \tau(r)\right\} - \left\{f(\widehat{\Theta}_{r_0}) + \tau(r_0)\right\} \\ \geq\ & C_2 d_r^2/2 + \tau(r) - \tau(r_0) \\ >\ & 0, \end{aligned} \tag{B.20}$$

when n and p are sufficiently large.

When $r \in \mathbf{A}_4$, the $a_{n,p,(r+r_0)}$ in (B.17) can be replaced with $a_{n,p,r}$, and we obtain

$$F(\widehat{\Delta}_r) \geq C_2\|\widehat{\Delta}_r\|_F^2 - C_{15}\|\widehat{\Delta}_r\|_F \max(a_{n,p,r}, b_{n,p}).$$

As $\mathbf{A}_4$ is defined so that $r/\max(r_0, \log p) \to \infty$, we have $a_{n,p,r}/b_{n,p} \to \infty$ and

$$F(\widehat{\Delta}_r) \;\geq\; C_2\|\widehat{\Delta}_r\|_F^2 - C_{15}\|\widehat{\Delta}_r\|_F a_{n,p,r}. \tag{B.21}$$

The right hand side of the inequality in (B.21) is quadratic in $\|\widehat{\Delta}_r\|_F$ and can be minimized analytically. Thus, (B.21) is bounded as

$$F(\widehat{\Delta}_r) \;\geq\; -C_{16}a_{n,p,r}^2, \tag{B.22}$$

in which C_{16} is some positive constant based on C_2 and C_{15}. ,

By (B.22) and Condition 3.4, we have

$$\begin{aligned} & \Big\{f(\widehat{\Theta}_r) + \tau(r)\Big\} - \Big\{f(\widehat{\Theta}_{r_0}) + \tau(r_0)\Big\} \\ \geq\; & -C_{16}a_{n,p,r}^2 + \tau(r) - \tau(r_0) \\ >\; & 0, \end{aligned} \tag{B.23}$$

when n and p are sufficiently large.

Results (B.20) and (B.23) together complete the proof. □

B.2 Supplementary technical details

This appendix contains some lemmas. Lemma B.1 and Lemma B.2 are repeatedly used in the proof of Theorem 3.1 – Theorem 3.3; Lemma B.3 is a useful result for the proof of Lemma B.2; Lemma B.4 is used to justify Algorithm 1.

Lemma B.1. *Let $\widehat{\Theta}_r$ be the solution of the low-rank and diagonal matrix decomposition when the rank is fixed to be r,*

$$\begin{aligned} \widehat{\Theta}_r \;&=\; \arg\min_{\Theta}\{\mathrm{tr}(\Theta S) - \log|\Theta|\}, \\ \textit{subject to} \quad & \Theta = -L + D,\ \Theta \in \mathbf{S}_+^p,\ L \in \mathbf{S}_+^{p,r},\ D \in \mathbf{D}^p, \end{aligned} \tag{B.24}$$

in which S is the sample covariance matrix, we have $\|\widehat{\Theta}_r - \Theta_0\|_{op} < C$ for some constant C, with probability tending to 1.

Proof. In the following proof, we will use the fact that, with probability tending to 1,

$$\begin{aligned} \lambda_{\max}(S^{-1}) = \lambda_{\min}^{-1}(S) &\leq \left\{\lambda_{\min}(\Sigma_0) - c(p/n)^{1/2}\right\}^{-1} \\ &\leq 2/c_1, \end{aligned} \tag{B.25}$$

for some constants c and c_1, where c_1 has been defined in Condition 3.1.

To prove this lemma, it suffices to show that

$$\lambda_{\max}(\widehat{\Theta}_r) \leq \lambda_{\max}(S^{-1}). \tag{B.26}$$

This is because

$$\begin{aligned} \|\widehat{\Theta}_r - \Theta_0\|_{op} &\leq \|\widehat{\Theta}_r - S^{-1}\|_{op} + \|S^{-1} - \Theta_0\|_{op} \\ &\leq \max\left\{\lambda_{\max}(\widehat{\Theta}_r), \lambda_{\max}(S^{-1})\right\} + \max\left\{\lambda_{\max}(S^{-1}), \lambda_{\max}(\Theta_0)\right\} \\ &\leq 4/c_1, \end{aligned}$$

The second inequality is due to the fact that $\widehat{\Theta}_r$, S^{-1} and Θ_0 are all positive definite. The last inequality is because of (B.26) and (B.25).

It remains to show (B.26). We will prove that, if $\lambda_{\max}(\Theta) > \lambda_{\max}(S^{-1})$ instead (i.e. (B.26) isn't true), then Θ must not be the solution to (B.24) because the objective function in (B.24) can always be further decreased. We conduct this proof in two steps.

Step 1: If $\lambda_{\max}(\Theta) > \lambda_{\max}(S^{-1})$, the objective function cannot reach its minimum.

Let $\Theta = D - L$ in which D and L are constrained as in (B.24). We eigen-decompose Θ as

$$\Theta = D - L = T\Lambda T^{T},$$

in which $T = (t_1, \ldots, t_p)$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$. Without loss of generality, let the eigenvalues be aligned in descending order. With basic calculus, the objective function in (B.24) can be rewritten as

$$\mathrm{tr}(\Lambda T^{T} S T) - \log|\Lambda| = \sum_{j=1}^{p} \left(\lambda_j t_j^{T} S t_j - \log\lambda_j\right), \tag{B.27}$$

for which the partial differentiation with respect to λ_1 is $t_1^T S t_1 - \lambda_1^{-1}$. Hence, due to convexity, (B.27) may reach its minimum when $\lambda_1 = (t_1^T S t_1)^{-1}$. However, $\lambda_1 > (t_1^T S t_1)^{-1}$ strictly because

$$\lambda_1 = \lambda_{\max}(\Theta) > \lambda_{\max}(S^{-1}) = \lambda_{\min}^{-1}(S) \geq (t_1^T S t_1)^{-1}.$$

Therefore, (B.27) cannot reach its minimum.

Step 2: Given that $\lambda_1 > (t_1^T S t_1)^{-1}$, the objective function can be further decreased if (not only if) we change the D (in $\Theta = D - L$) in a way that both $t_1^T S t_1$ and $\sum_{j=2}^{p} \left(\lambda_j t_j^T S t_j - \log \lambda_j\right)$ remain unchanged but λ_1 decreases.

We now show that such a change in D does exist. By employing the results of differentiating eigenvalues and eigenvectors in Magnus (1985), we have the following three results. First of all

$$d\lambda_1 = t_1^T (dD) t_1, \tag{B.28}$$

Secondly,

$$\begin{aligned} d\left(t_1^T S t_1\right) &= 2(St_1)^T dt_1 \\ &= 2(St_1)^T (\lambda_1 I_p - \Theta)^+ (dD) t_1. \end{aligned} \tag{B.29}$$

Lastly,

$$\begin{aligned} d\left\{\sum_{j=2}^{p} \left(\lambda_j t_j^T S t_j - \log \lambda_j\right)\right\} &= \sum_{j=2}^{p} (t_j^T S t_j - \lambda_j^{-1}) d\lambda_j + 2\lambda_j (St_j)^T dt_j \\ &= \sum_{j=2}^{p} (t_j^T S t_j - \lambda_j^{-1}) t_j^T (dD) t_j \\ &\quad + 2\lambda_j (St_j)^T (\lambda_j I_p - \Theta)^+ (dD) t_j, \end{aligned} \tag{B.30}$$

in which dD is a diagonal matrix representing an infinitesimal change of D and $(\cdot)^+$ is the Moore-Penrose inverse. Expressions (B.28),(B.29) and (B.30) are all linear with respect to the elements in dD and $t_1 \neq 0$ obviously. Hence, we can surely solve dD from setting (B.29) and (B.30) to be 0 and (B.28) to be negative.

In summary, we have shown that if we change D by dD, the objective function (B.27) decreases. Therefore, we have proved that Θ is not the solution to (B.24). This completes the proof. □

Lemma B.2. *If a $p \times p$ matrix M can be written as $M = L + D$, in which $L \in \mathbf{S}^{p,r}$ and $D \in \mathbf{D}^p$, then M can also be written as $M = L' + D'$, in which $L' \in \mathbf{S}^{p,3r}$, $D' \in \mathbf{D}^p$ and*

$$\|M\|_F^2 \geq C\left(\|L'\|_F^2 + \|D'\|_F^2\right),$$

for some positive constant C.

Proof. Let M_{ij} and L_{ij} be the entries in the ith row and jth column of M and L respectively; let D_j be the jth diagonal entry of D. Similarly, L'_{ij} and D'_j are defined. Define the index set $\mathbf{B} = \{j \mid L_{jj}^2 > \sum_{i \neq j} L_{ij}^2\}$.

According to Lemma B.3, the cardinality of $\mathbf{B}$ is at most $2r - 1$. We set $L'_{jj} = M_{jj}/2$ for $j \in \mathbf{B}$ and $L'_{ij} = L_{ij}$ for $i \neq j$ and $i = j \notin \mathbf{B}$; D' is set accordingly so that $M = L' + D'$. As at most $2r - 1$ diagonal entries of L' are different from those of L, $\text{rank}\,(L') < 3r$. Now we prove $\|M\|_F^2 \geq C\left(\|L'\|_F^2 + \|D'\|_F^2\right)$ for some constant C.

We notice two properties: (1) for $j \in \mathbf{B}$, $(L'_{jj})^2 = M_{jj}^2/4$; (2) for $j \notin \mathbf{B}$, $(L'_{jj})^2 \leq \sum_{i \neq j}(L'_{ij})^2 = \sum_{i \neq j} M_{ij}^2$. As a result,

$$\begin{aligned}
\|M\|_F^2 &= \sum_{j=1}^{p} M_{jj}^2 + \sum_{j=1}^{p}\sum_{i \neq j} M_{ij}^2 \\
&\geq 4\sum_{j \in \mathbf{B}}(L'_{jj})^2 + \sum_{j=1}^{p}\sum_{i \neq j}(L'_{ij})^2 \\
&\geq 1/2\left\{\sum_{j \in \mathbf{B}}(L'_{jj})^2 + 2\sum_{j=1}^{p}\sum_{i \neq j}(L'_{ij})^2\right\} \\
&\geq 1/2\left\{\sum_{j \in \mathbf{B}}(L'_{jj})^2 + \sum_{j \notin \mathbf{B}}(L'_{jj})^2 + \sum_{j=1}^{p}\sum_{i \neq j}(L'_{ij})^2\right\} \\
&= 1/2\|L'\|_F^2. \qquad \text{(B.31)}
\end{aligned}$$

The first inequality is because of property (1) and the third inequality is because of property (2).

Finally, by (B.31) and $\|D'\|_F \leq \|M\|_F + \|L'\|_F$, we have

$$\begin{aligned} \|D'\|_F^2 &\leq 2(\|M\|_F^2 + \|L'\|_F^2) \\ &\leq 6\|M\|_F^2, \end{aligned}$$

we have $\|L'\|_F^2 + \|D'\|_F^2 \leq 8\|M\|_F^2$ and $\|M\|_F^2 \geq C(\|L'\|_F^2 + \|D'\|_F^2)$ for $C = 1/8$. This completes the proof. □

Lemma B.3. *Let A be a $p \times p$ matrix with* $\text{rank}(A) = r$ *($r \leq p$) and A_{ij} be the element in the ith row and jth column, the number of column vectors in A that satisfy $A_{jj}^2 > \sum_{i \neq j} A_{ij}^2$ is at most $2r - 1$.*

Proof. Let a_j be the jth column vector in A. If it satisfies $A_{jj}^2 > \sum_{i \neq j} A_{ij}^2$, we say this column is diagonally dominant and is dominated by the jth element. Let $\mathbf{R}^p$ denote the dimension p vector space, and $\mathbf{R}^{p,r}$ denote the column space of A. Straightforwardly, $\mathbf{R}^{p,r}$ is a subspace of $\mathbf{R}^p$ that contains at most r linearly independent vectors.

Finding out the upper bound of the number of diagonally dominant column vectors in A is equivalent to considering at most how many vectors in $\mathbf{R}^{p,r}$ can be dominated by one of its entries. The equivalence requires, when we count in $\mathbf{R}^{p,r}$, if two vectors are dominated by the same entry (e.g., jth), they are counted as one vector. Now, we count in $\mathbf{R}^{p,r}$.

Without loss of generality, we assume the first r columns $(a_1, \ldots, a_r)$ in A are orthogonal to each other and are unit vectors. This is valid because for any given A, without changing the column space, we can (1) change the order of the columns by moving r linearly independent column vectors to the left and (2) orthonormalize these linearly independent vectors.

Let

$$V_{p \times r} = \Big(a_1, \ldots, a_r\Big) = \begin{pmatrix} b_1^T \\ \vdots \\ b_p^T \end{pmatrix},$$

in which $b_1, \ldots, b_p$ are $r \times 1$ vectors. Any vector in $\mathbf{R}^{p,r}$ can be written as $V_{p\times r}k$ where k is a $r \times 1$ vector; therefore, a vector dominated by the jth element can be in $\mathbf{R}^{p,r}$ if and only if there is a vector $k \neq 0$ and

$$(b_j^T k)^2 > \sum_{i\neq j}(b_i^T k)^2.$$

The inequality is equivalent to

$$k^T b_j b_j^T k > \sum_{i\neq j} k^T b_i b_i^T k,$$

and

$$k^T (V^T V - 2b_j b_j^T)k < 0.$$

The existence of k suggests $V^T V - 2b_j b_j^T$ has negative eigenvalues. As V consists of orthonormal vectors, we conclude the smallest eigenvalue of

$$V^T V - 2b_j b_j^T = I_r - 2b_j b_j^T$$

must be negative.

Let $\lambda_{\min}(\cdot)$ be the smallest eigenvalue of a matrix and u be the corresponding eigenvector of $\lambda_{\min}\left(I_r - 2b_j b_j^T\right)$. We have

$$\begin{aligned}\lambda_{\min}\left(I_r - 2b_j b_j^T\right) &= u^T\left(I_r - 2b_j b_j^T\right)u \\ &= 1 - 2(u^T b_j)^2 \\ &\geq 1 - 2\|b_j\|^2\end{aligned}$$

and consequently $\|b_j\|^2 > 1/2$. Finally, noticing $\sum_{i=1}^p \|b_i\|^2 = \sum_{j=1}^r \|a_j\|^2 = r$, we conclude there are at most $2r-1$ b_j with $\|b_j\|^2 > 1/2$. □

Lemma B.4. *When D is fixed and positive definite, the objective function*

$$\operatorname{tr}\{(D-L)S\} - \log|D-L|$$

can be minimized with respect to L analytically.

Eigen-decompose $D^{1/2}SD^{1/2}$, let $w_1, \ldots, w_p$ be the eigenvalues in descending order and $u_1, \ldots, u_p$ be the associated eigenvectors. Let $U = (u_1, \ldots, u_r)$, $V = \operatorname{diag}\{1-1/\max(w_1, 1), \ldots, 1-1/\max(w_r, 1)\}$, then $L = D^{1/2}UVU^T D^{1/2}$ is the analytic solution.

Proof. Since D and S are fixed, the target can be simplified as maximizing

$$\begin{aligned}&\operatorname{tr}(LS) + \log|D - L| \\ =&\operatorname{tr}\left\{(D^{-1/2}LD^{-1/2})D^{1/2}SD^{1/2}\right\} + \log|I_p - D^{-1/2}LD^{-1/2}| + \log|D|.\end{aligned}$$

Let the low-rank part be eigen-decomposed as $D^{-1/2}LD^{-1/2} = \widetilde{U}\widetilde{V}\widetilde{U}^T$, in which $\widetilde{U} = (\widetilde{u}_1, \ldots, \widetilde{u}_r)$ is a $p \times r$ matrix and $\widetilde{V} = \operatorname{diag}(\widetilde{v}_1, \ldots, \widetilde{v}_r)$ is a $r \times r$ diagonal matrix. Also, without of loss generality, let $\widetilde{v}_1, \ldots, \widetilde{v}_r$ be in descending order. Then, we need to maximize

$$\operatorname{tr}\left\{\widetilde{V}\widetilde{U}^T D^{1/2}SD^{1/2}\widetilde{U}\right\} + \log|I_r - \widetilde{V}|. \tag{B.32}$$

Regardless of $\widetilde{V}$, We have

$$\begin{aligned}\operatorname{tr}\left\{\widetilde{V}\widetilde{U}^T D^{1/2}SD^{1/2}\widetilde{U}\right\} &\leq \sum_{i=1}^{r}\lambda_i(\widetilde{V})\lambda_i(\widetilde{U}^T D^{1/2}SD^{1/2}\widetilde{U}) \\ &\leq \sum_{i=1}^{r}\widetilde{v}_i\lambda_i(D^{1/2}SD^{1/2}) \\ &= \sum_{i=1}^{r}\widetilde{v}_i w_i,\end{aligned}$$

where $\lambda_i(\cdot)$ is the ith largest eigenvalue of the input matrix. The first and second inequalities follow Theorem 3.34 and Theorem 3.19 in Schott (2005) respectively. The maximum can be achieved when $\widetilde{U} = U$.

When $\widetilde{U} = U$, maximizing (B.32) is equivalent to maximizing

$$\widetilde{v}_i w_i + \log(1 - \widetilde{v}_i) \quad (i = 1, \ldots, r),$$

subject to $\widetilde{v}_i \in [0, 1)$. By basic calculus, we need $\widetilde{v}_i = 1 - 1/\max(w_i, 1)$. □

Appendix C

Proofs of Chapter 4

To prove the theorems, we set up a framework to deal with the "joint diagonal + low-rank" structure. On the occasion that certain derivations can be broken down into individual categories, we refer to Appendix C, proofs of "diagonal + low-rank".

C.1 Proof of Theorem 4.1

Proof of Theorem 4.1. Firstly, by triangular inequality,

$$\begin{aligned}\sum_{k=1}^{K}\|\widehat{\Theta}_{r,v}^{(k)}-\Theta_0^{(k)}\|_F &\leq \sum_{k=1}^{K}\|\widehat{\Theta}_{r,v}^{(k)}-\Theta_{r,v}^{(k)}\|_F+\sum_{k=1}^{K}\|\Theta_{r,v}^{(k)}-\Theta_0^{(k)}\|_F \\ &= \sum_{k=1}^{K}\|\widehat{\Theta}_{r,v}^{(k)}-\Theta_{r,v}^{(k)}\|_F+d_{r,v}.\end{aligned}$$

With $d_{r,v}$ restricted by the condition of the theorem, i.e., $d_{r,v}=O\{\max(a_{n,p,v},b_{n,p})\}$, it remains to prove

$$\sum_{k=1}^{K}\|\widehat{\Theta}_{r,v}^{(k)}-\Theta_{r,v}^{(k)}\|_F = O\{\max(a_{n,p,v},b_{n,p})\}. \tag{C.1}$$

Recall the search space of (4.7),

$$\begin{aligned}\mathbf{F}_{r,v} &= \{\Theta = (\Theta^{(1)}, \ldots, \Theta^{(K)}) \mid \Theta^{(k)} = D - L - L^{(k)}, \\ &\quad D \in \mathbf{D}^p_{++},\ L \in \mathbf{S}^{p,r}_+,\ L^{(k)} \in \mathbf{S}^{p,(v_k - r)}_+ \text{ for all } k\},\end{aligned}$$

on which the solution $\widehat{\Theta}_{r,v}$ minimizes the objective function. Here, we additionally define the "centered" version of $\mathbf{F}_{r,v}$ as

$$\mathbf{E}_{r,v} = \{\Delta = (\Delta^{(1)}, \ldots, \Delta^{(K)}) \mid \Delta^{(k)} = \Theta^{(k)} - \Theta^{(k)}_{r,v},\ \Theta \in \mathbf{F}_{r,v}\}.$$

Since $\Theta_{r,v} \in \mathbf{F}_{r,v}$, we have $0 \in \mathbf{E}_{r,v}$.

Let $f^{(k)}(\Theta^{(k)}) = \operatorname{tr}(\Theta^{(k)} S^{(k)}) - \log|\Theta^{(k)}|$, $F^{(k)}(\Delta^{(k)}) = f^{(k)}(\Theta^{(k)}_{r,v} + \Delta^{(k)}) - f(\Theta^{(k)}_{r,v})$, $\widehat{\Delta}^{(k)}_{r,v} = \widehat{\Theta}^{(k)}_{r,v} - \Theta^{(k)}_{r,v}$ and $\widehat{\Delta}_{r,v} = (\widehat{\Delta}^{(1)}_{r,v}, \ldots, \widehat{\Delta}^{(K)}_{r,v})$.

We can prove the desired result, equivalent of (C.1),

$$\sum_{k=1}^{K} \|\widehat{\Delta}^{(k)}_{r,v}\|_F \leq M \max(a_{n,p,v}, b_{n,p}), \tag{C.2}$$

for some constant M, by proving

$$F(\Delta) = \sum_{k=1}^{K} n_k F^{(k)}(\Delta^{(k)}) > 0 \text{ for all } \Delta \in \mathbf{M}_{2r,2v}, \tag{C.3}$$

in which

$$\mathbf{M}_{2r,2v} = \mathbf{E}_{2r,2v} \cap \{\Delta \mid \sum_{k=1}^{K} \|\Delta^{(k)}\|_F = M \max(a_{n,p,v}, b_{n,p})\} \cap \{\Delta \mid \|\Delta^{(k)}\|_{op} \leq C_1\};$$

Constant C_1 is the upper bound of $C + d_{r,v}$ (C as in Lemma C.1) so that

$$\|\widehat{\Delta}^{(k)}_{r,v}\|_{op} \leq \|\widehat{\Theta}^{(k)}_{r,v} - \Theta^{(k)}_0\|_{op} + \|\Theta^{(k)}_0 - \Theta^{(k)}_{r,v}\|_{op} \leq C + d_{r,v} \leq C_1.$$

To show (C.3) is indeed sufficient for (C.2), we derive contradiction when (C.3) is true while (C.2) is not. If $\sum_{k=1}^{K} \|\widehat{\Delta}^{(k)}_{r,v}\|_F > M \max(a_{n,p,v}, b_{n,p})$, there exists $0 < t < 1$, so that $(1-t)0 + t\widehat{\Delta}_{r,v} \in \mathbf{E}_{2r,2v}$, $\sum_{k=1}^{K} \|(1-t)0 + t\widehat{\Delta}^{(k)}_{r,v}\|_F = M \max(a_{n,p,v}, b_{n,p})$ and $\|(1-t)0 +$

$t\widehat{\Delta}_{r,v}^{(k)}\|_{op} \le C_1$; that is, $(1-t)0+t\widehat{\Delta}_{r,v} \in \mathbf{M}_{2r,2v}$. Therefore, $F\{(1-t)0+t\widehat{\Delta}_{r,v}\} > 0$ by (C.3). On the other hand, by convexity, we have $F\{(1-t)0+t\widehat{\Delta}_{r,v}\} \le (1-t)F(0)+tF(\widehat{\Delta}_{r,v}) \le 0$; to see the last inequality, we notice $F(0) = 0$ and $F(\widehat{\Delta}_{r,v}) \le 0$ because $\widehat{\Theta}_{r,v}$ is the minimizer. The contradiction has been derived, and it remains to prove (C.3).

For any $\Delta \in \mathbf{M}_{2r,2v}$, we have

$$\begin{aligned} F(\Delta) &= \sum_{k=1}^{K} n_k \left[\operatorname{tr}(\Delta^{(k)} S^{(k)}) - \{\log|\Theta_{r,v}^{(k)} + \Delta^{(k)}| - \log|\Theta_{r,v}^{(k)}|\}\right] \\ &\ge C_2 \sum_{k=1}^{K} n_k \|\Delta^{(k)}\|_F^2 + \sum_{k=1}^{K} n_k \operatorname{tr}\{\Delta^{(k)}(S^{(k)} - \Sigma_{r,v}^{(k)})\} \\ &= C_2 \sum_{k=1}^{K} n_k \|\Delta^{(k)}\|_F^2 + \underbrace{\sum_{k=1}^{K} n_k \operatorname{tr}\{\Delta^{(k)}(\Sigma_0^{(k)} - \Sigma_{r,v}^{(k)})\}}_{\text{I}} \\ &\quad + \underbrace{\sum_{k=1}^{K} n_k \operatorname{tr}\{\Delta^{(k)}(S^{(k)} - \Sigma_0^{(k)})\}}_{\text{II}}, \end{aligned} \tag{C.4}$$

where $\Sigma_{r,v}^{(k)} = (\Theta_{r,v}^{(k)})^{-1}$. The inequality is a result of the similar inequality of the log-determinant function in the proof of Theorem 3.1 in Chapter B and the definition of $\mathbf{M}_{2r,2v}$ ($\|\Delta^{(k)}\|_{op} \le C_1$). We consider I and II separately.

To bound I,

$$\begin{aligned} |\text{I}| &\le \sum_{k=1}^{K} n_k \|\Delta^{(k)}\|_F \|\Sigma_0^{(k)} - \Sigma_{r,v}^{(k)}\|_F \\ &\le C_3 n d_{r,v} \sum_{k=1}^{K} \|\Delta^{(k)}\|_F. \end{aligned} \tag{C.5}$$

The first inequality is the Cauchy–Schwarz inequality, and the second inequality is a direct result of $\|\Sigma_0^{(k)} - \Sigma_{r,v}^{(k)}\|_F = \|(\Theta_0^{(k)})^{-1} - (\Theta_{r,v}^{(k)})^{-1}\|_F \le C_3 \|\Theta_0^{(k)} - \Theta_{r,v}^{(k)}\|_F$, which can be derived by Taylor expansion.

By the definition of $\mathbf{E}_{2r,2v}$ and Lemma B.2 in Appendix B.2, we can write $\Delta^{(k)} = D_\Delta^{(k)} + L_\Delta^{(k)}$, so that $D_\Delta^{(k)} \in \mathbf{D}^p$, $L_\Delta^{(k)} \in \mathbf{S}^{p,9v_k}$ and $\|\Delta^{(k)}\|_F^2 \ge C_4(\|D_\Delta^{(k)}\|_F^2 + \|L_\Delta^{(k)}\|_F^2)$ for

some positive constant C_4. The bound of II can be obtained by considering the bound for each category,

$$\begin{aligned}
|\mathrm{II}| &\leq \sum_{k=1}^{K} n_k |\operatorname{tr}\{\Delta^{(k)}(S^{(k)} - \Sigma_0^{(k)})\}| \\
&\leq \sum_{k=1}^{K} n_k \left[|\operatorname{tr}\{D_\Delta^{(k)}(S^{(k)} - \Sigma_0^{(k)})\}| + |\operatorname{tr}\{L_\Delta^{(k)}(S^{(k)} - \Sigma_0^{(k)})\}| \right] \\
&\leq \sum_{k=1}^{K} n_k \left[\|D_\Delta^{(k)}\|_F \left\{ \sum_{j=1}^{p} (s_j^{(k)} - \sigma_{0j}^{(k)})^2 \right\}^{1/2} + \|L_\Delta^{(k)}\|_* \|S^{(k)} - \Sigma_0^{(k)}\|_{op} \right] \\
&\leq \sum_{k=1}^{K} n_k \left[p^{1/2} \|D_\Delta^{(k)}\|_F \max_{1 \leq j \leq p} |s_j^{(k)} - \sigma_{0j}^{(k)}| + (9v_k)^{1/2} \|L_\Delta^{(k)}\|_F \|S^{(k)} - \Sigma_0^{(k)}\|_{op} \right] \\
&\leq C_5 \sum_{k=1}^{K} n_k \left[\{(p \log p)/n_k\}^{1/2} \|D_\Delta^{(k)}\|_F + v_k^{1/2} (p/n_k)^{1/2} \|L_\Delta^{(k)}\|_F \right] \\
&\leq C_6 \sum_{k=1}^{K} n_k \max(a_{n_k,p,v_k}, b_{n_k,p}) \|\Delta^{(k)}\|_F, \\
&\leq C_6 n \max(a_{n,p,v}, b_{n,p}) \sum_{k=1}^{K} \|\Delta^{(k)}\|_F \qquad \text{(C.6)}
\end{aligned}$$

in which $s_j^{(k)}$ and $\sigma_{0j}^{(k)}$ are the jth diagonal elements of $S^{(k)}$ and $\Sigma_0^{(k)}$ respectively, $a_{n_k,p,v_k} = v_k^{1/2}(p/n_k)^{1/2}$ and $b_{n_k} = \{(p \log p)/n_k\}^{1/2}$. The second to the second last inequalities here depend on single category analysis, and omitted details can be found in the proof of Theorem 3.1 in Chapter B.

We replace I and II in (C.4) with their lower bounds and obtain

$$\begin{aligned}
F(\Delta) &\geq C_2 \sum_{k=1}^{K} n_k \|\Delta^{(k)}\|_F^2 - C_3 n d_{r,v} \sum_{k=1}^{K} \|\Delta^{(k)}\|_F - C_6 n \max(a_{n,p,v}, b_{n,p}) \sum_{k=1}^{K} \|\Delta^{(k)}\|_F \\
&\geq C_7 n \sum_{k=1}^{K} \|\Delta^{(k)}\|_F^2 - C_8 n \max(a_{n,p,v}, b_{n,p}) \sum_{k=1}^{K} \|\Delta^{(k)}\|_F \\
&\geq C_7 n K^{-1} \left(\sum_{k=1}^{K} \|\Delta^{(k)}\|_F \right)^2 - C_8 n \max(a_{n,p,v}, b_{n,p}) \sum_{k=1}^{K} \|\Delta^{(k)}\|_F \\
&= n \left(\sum_{k=1}^{K} \|\Delta^{(k)}\|_F \right)^2 \left\{ C_7 K^{-1} - C_8 \max(a_{n,p,v}, b_{n,p}) \left(\sum_{k=1}^{K} \|\Delta^{(k)}\|_F \right)^{-1} \right\} \\
&= n \left(\sum_{k=1}^{K} \|\Delta^{(k)}\|_F \right)^2 \left\{ C_7 K^{-1} - C_8 M^{-1} \right\} \\
&> 0,
\end{aligned}$$

for sufficiently large M. The second inequality is due to Condition 4.2 and $d_{r,v} = O\{\max(a_{n,p,v}, b_{n,p})\}$. This proves (C.3) and completes the proof of Theorem 4.1. The proof of Corollary 4.1 is similar to this proof and is omitted. □

C.2 Proof of Theorem 4.2

Proof of Theorem 4.2. To prove this theorem, according to the discussion in Section 4.4, we need to show that a pair $(r, v) \in \mathbf{A}_1 \cup \mathbf{A}_2$ is never chosen; that is, the objective function (4.9) evaluated at $\widehat{\Theta}_{r,v}$ with $(r, v) \in \mathbf{A}_1 \cup \mathbf{A}_2$, is always larger than that evaluated at $\widehat{\Theta}$.

We tackle this by proving $\forall (r, v) \in \mathbf{A}_1 \cup \mathbf{A}_2$,

$$\left\{ \sum_{k=1}^{K} n_k f^{(k)}(\widehat{\Theta}_{r,v}^{(k)}) + \lambda \tau(r, v) \right\} - \left\{ \sum_{k=1}^{K} n_k f^{(k)}(\widehat{\Theta}_{r_0,v_0}^{(k)}) + \lambda \tau(r_0, v_0) \right\} > 0. \tag{C.7}$$

This is sufficient because the objective function evaluated at $\widehat{\Theta}_{r_0,v_0}$ is no smaller than the minimized objective function, or the objective function evaluated at $\widehat{\Theta}$.

To this end, we firstly look at the penalty term and the negative log-likelihood term separately, and then (C.7) is proven for $\mathbf{A}_1$ and $\mathbf{A}_2 \cap \mathbf{A}_1^c$.

Step 1: The penalty. The range of the penalty of arbitrary (r, v) is considered. For the purpose of brevity, let $v_{\max} = \max_k v_k$. We have

$$\begin{aligned}
\tau(r,v) &> 2p(v_{\max} - r) - (v_{\max} - r)(v_{\max} - r - 1) + 2p(r+1) - r(r-1) \\
&= 2p(v_{\max} + 1) - v_{\max}(v_{\max} - 1) + 2v_{\max} r - 2r^2 \\
&\geq 2p(v_{\max} + 1) - v_{\max}(v_{\max} - 1) \\
&= v_{\max}\{2p - (v_{\max} - 1)\} + 2p \\
&\geq pv_{\max} \qquad \text{(C.8)}
\end{aligned}$$

and

$$\begin{aligned}
\tau(r,v) &\leq 2Kp(v_{\max} - r) - K(v_{\max} - r)(v_{\max} - r - 1) + 2p(r+1) - r(r-1) \\
&= 2Kp(v_{\max} + 1) - Kv_{\max}(v_{\max} - 1) \\
&\quad + 2Kv_{\max} r + 2(1-K)p(r+1) - (1+K)r^2 + (1-K)r \\
&< 2Kp(v_{\max} + 1) + 2Kv_{\max} r \\
&< 2Kp(2v_{\max} + 1) \qquad \text{(C.9)}
\end{aligned}$$

Then, we can find out the bound of $\tau(r, v) - \tau(r_0, v_0)$. Let $v_{0,\max} = \max_k v_{0k}$.

When $(r, v) \in \mathbf{A}_1$, we have $v_{\max}/v_{0,\max} \to \infty$, $v_{\max}/\log p \to \infty$; thus,

$$\begin{aligned}
\tau(r,v) - \tau(r_0,v_0) &> p(v_{\max} - 4Kv_{0,\max} - 2K) \\
&\geq (pv_{\max})/2, \qquad \text{(C.10)}
\end{aligned}$$

when n and p are sufficiently large.

When $(r, v) \in \mathbf{A}_2 \cap \mathbf{A}_1^c$, we have $v_{\max} = O(v_{0,\max})$; thus,

$$\begin{aligned}
|\tau(r,v) - \tau(r_0,v_0)| &\leq \max\{\tau(r,v) - \tau(r_0,v_0), \tau(r_0,v_0) - \tau(r,v)\} \\
&< 2Kp \max\{2v_{\max} + 1, 2v_{0,\max} + 1\} \\
&= O(na_{n,p,v_0}^2). \qquad \text{(C.11)}
\end{aligned}$$

Step 2: The negative log-likelihood. To obtain the lower bound of the left-hand-side of (C.7), we notice, since $\widehat{\Theta}_{r_0,v_0}$ minimizes (4.9) when the ranks are (r_0, v_0),

$$\sum_{k=1}^{K} n_k f^{(k)}(\widehat{\Theta}_{r,v}^{(k)}) - \sum_{k=1}^{K} n_k f^{(k)}(\widehat{\Theta}_{r_0,v_0}^{(k)}) \geq \sum_{k=1}^{K} n_k f^{(k)}(\widehat{\Theta}_{r,v}^{(k)}) - \sum_{k=1}^{K} n_k f^{(k)}(\Theta_0^{(k)}). \tag{C.12}$$

Therefore, a lower bound of the right-hand-side of (C.12) suffices, and we proceed to find this lower bound.

In a slight abuse of notation, let $F^{(k)}(\Delta^{(k)}) = f^{(k)}(\Theta_0^{(k)} + \Delta^{(k)}) - f^{(k)}(\Theta_0^{(k)})$ and $\widehat{\Delta}_{r,v}^{(k)} = \widehat{\Theta}_{r,v}^{(k)} - \Theta_0^{(k)}$; that is, we now center at $\Theta_0^{(k)}$ instead of $\Theta_{r,v}^{(k)}$. The problem becomes finding out the lower bound of $F(\widehat{\Delta}_{r,v}) = \sum_{k=1}^{K} n_k F^{(k)}(\widehat{\Delta}_{r,v}^{(k)})$.

We firstly notice

$$\begin{aligned} F(\widehat{\Delta}_{r,v}) &= \sum_{k=1}^{K} n_k \left[\operatorname{tr}(\widehat{\Delta}_{r,v}^{(k)} S^{(k)}) - \{\log|\Theta_0^{(k)} + \widehat{\Delta}_{r,v}^{(k)}| - \log|\Theta_0^{(k)}|\} \right] \\ &\geq C_9 \sum_{k=1}^{K} n_k \|\widehat{\Delta}_{r,v}^{(k)}\|_F^2 + \underbrace{\sum_{k=1}^{K} n_k \operatorname{tr}\{\widehat{\Delta}_{r,v}^{(k)}(S^{(k)} - \Sigma_0^{(k)})\}}_{\text{III}}. \end{aligned} \tag{C.13}$$

The inequality is a result of the similar inequality of the log-determinant function in Theorem 3.1 of Chapter B and Lemma C.1; although look alike, (C.13) differs from (C.4) in that the latter results from the definition of $\mathbf{M}_{2r,2v_k}$ instead of Lemma C.1.

Now we look at III, which is similar to II except that $\widehat{\Delta}_{r,v}^{(k)} = \widehat{\Theta}_{r,v}^{(k)} - \Theta_0^{(k)}$. We can write $\widehat{\Delta}_{r,v}^{(k)} = D_\Delta^{(k)} + L_\Delta^{(k)}$, so that $D_\Delta^{(k)} \in \mathbf{D}^p$, $L_\Delta^{(k)} \in \mathbf{S}^{p,3(v_{0k}+v_k)}$ and $\|\widehat{\Delta}_{r,v}^{(k)}\|_F^2 \geq C_4(\|D_\Delta^{(k)}\|_F^2 +$

$\|L_\Delta^{(k)}\|_F^2$). Then, the bound of III can be obtained,

$$\begin{aligned}
|\text{III}| &\leq \sum_{k=1}^{K} n_k |\operatorname{tr}\{\widehat{\Delta}_{r,v}^{(k)}(S^{(k)} - \Sigma_0^{(k)})\}| \\
&\leq \sum_{k=1}^{K} n_k \left[|\operatorname{tr}\{D_\Delta^{(k)}(S^{(k)} - \Sigma_0^{(k)})\}| + |\operatorname{tr}\{L_\Delta^{(k)}(S^{(k)} - \Sigma_0^{(k)})\}| \right] \\
&\leq C_{10} \sum_{k=1}^{K} n_k \left[\{(p \log p)/n_k\}^{1/2} \|D_\Delta^{(k)}\|_F + (v_{0k} + v_k)^{1/2} (p/n_k)^{1/2} \|L_\Delta^{(k)}\|_F \right] \\
&\leq C_{11} \sum_{k=1}^{K} n_k \max\{a_{n_k,p,(v_{0k}+v_k)}, b_{n_k,p}\} \|\widehat{\Delta}_{r,v}^{(k)}\|_F, \\
&\leq C_{12} n \max(a_{n,p,v_0}, a_{n,p,v}, b_{n,p}) \sum_{k=1}^{K} \|\widehat{\Delta}_{r,v}^{(k)}\|_F, \qquad \text{(C.14)}
\end{aligned}$$

in which $a_{n_k,p,(v_{0k}+v_k)} = (v_{0k} + v_k)^{1/2}(p/n_k)^{1/2}$. The third inequality is derived in the same manner as (C.6).

By plugging (C.14) into (C.13) and some simple calculation, we have

$$F(\widehat{\Delta}_{r,v}) \geq n \left\{ C_{13} K^{-1} \left(\sum_{k=1}^{K} \|\widehat{\Delta}_{r,v}^{(k)}\|_F \right)^2 - C_{12} \max(a_{n,p,v_0}, a_{n,p,v}, b_{n,p}) \sum_{k=1}^{K} \|\widehat{\Delta}_{r,v}^{(k)}\|_F \right\}. \qquad \text{(C.15)}$$

Step 3: The inequality (C.7).

When $(r, v) \in \mathbf{A}_1$, we can simplify (C.15) as

$$\begin{aligned}
F(\widehat{\Delta}_{r,v}) &\geq n \left\{ C_{13} K^{-1} \left(\sum_{k=1}^{K} \|\widehat{\Delta}_{r,v}^{(k)}\|_F \right)^2 - C_{12} a_{n,p,v} \sum_{k=1}^{K} \|\widehat{\Delta}_{r,v}^{(k)}\|_F \right\} \\
&\geq -C_{14} n a_{n,p,v}^2. \qquad \text{(C.16)}
\end{aligned}$$

The second inequality is by minimizing the quadratic function with respect to $\sum_{k=1}^{K} \|\widehat{\Delta}_{r,v}^{(k)}\|_F$.

By (C.10), (C.16) and Condition 4.4,

$$\begin{aligned}
&\left\{\sum_{k=1}^{K} n_k f^{(k)}(\widehat{\Theta}_{r,v}^{(k)}) + \lambda\tau(r,v)\right\} - \left\{\sum_{k=1}^{K} n_k f^{(k)}(\widehat{\Theta}_{r_0,v_0}^{(k)}) + \lambda\tau(r_0,v_0)\right\} \\
\geq\ & -C_{14} n a_{n,p,v}^2 + \lambda\left\{\tau(r,v) - \tau(r_0,v_0)\right\} \\
\geq\ & 0, \qquad\qquad (C.17)
\end{aligned}$$

when n and p are sufficiently large.

When $(r,v) \in \mathbf{A}_2 \cap \mathbf{A}_1^c$, let $d_{r,v} = \eta_{n,p,v} \max(a_{n,p,v_0}, b_{n,p})$, where $\eta_{n,p,v} \to \infty$, we can simplify (C.15) as

$$\begin{aligned}
F(\widehat{\Delta}_{r,v}) \ &\geq\ n\left\{C_{13}K^{-1}\left(\sum_{k=1}^{K}\|\widehat{\Delta}_{r,v}^{(k)}\|_F\right)^2 - C_{15}\max(a_{n,p,v_0}, b_{n,p})\sum_{k=1}^{K}\|\widehat{\Delta}_{r,v}^{(k)}\|_F\right\} \\
&=\ n\left(\sum_{k=1}^{K}\|\widehat{\Delta}_{r,v}^{(k)}\|_F\right)^2\left\{C_{13}K^{-1} - C_{15}\max(a_{n,p,v_0}, b_{n,p})\left(\sum_{k=1}^{K}\|\widehat{\Delta}_{r,v}^{(k)}\|_F\right)^{-1}\right\} \\
&\geq\ n\left(\sum_{k=1}^{K}\|\widehat{\Delta}_{r,v}^{(k)}\|_F\right)^2\left(C_{13}K^{-1} - C_{15}\eta_{n,p,v}^{-1}\right) \\
&\geq\ C_{16} n (d_{r,v})^2, \qquad\qquad (C.18)
\end{aligned}$$

when n and p are sufficiently large. The first inequality is due to $\max_k v_k = O\{\max(\max_k v_{0k}, \log p)\}$. The second inequality is because $\sum_{k=1}^{K}\|\widehat{\Delta}_{r,v}^{(k)}\|_F \geq d_{r,v}$.

By (C.11), (C.18) and Condition 4.4,

$$\begin{aligned}
&\left\{\sum_{k=1}^{K} n_k f^{(k)}(\widehat{\Theta}_{r,v}^{(k)}) + \lambda\tau(r,v)\right\} - \left\{\sum_{k=1}^{K} n_k f^{(k)}(\widehat{\Theta}_{r_0,v_0}^{(k)}) + \lambda\tau(r_0,v_0)\right\} \\
\geq\ & C_{16} n (d_{r,v})^2 + \lambda\left\{\tau(r,v) - \tau(r_0,v_0)\right\} \\
\geq\ & 0, \qquad\qquad (C.19)
\end{aligned}$$

when n and p are sufficiently large.

Results (C.17) and (C.19) together say that, with a tuning parameter satisfying Condition 4.4, $(r,v) \in \mathbf{A}_1 \cup \mathbf{A}_2$ is never chosen over (r_0, v_0), and this completes the proof. □

Lemma C.1. *Let $\widehat{\Theta}_{r,v}$ be the solution of (4.7), then for every k we have $\|\widehat{\Theta}_{r,v}^{(k)} - \Theta_0^{(k)}\|_{op} < C$ for some constant C, with probability tending to* 1.

Proof. To prove this lemma, we follow the steps of the proof of Lemma B.1 in Appendix B.2 and modify in the following way: (i) the objective function is now a summation over K categories; (ii) the eigenvalue used to establish the contradiction should be the maximum eigenvalue of any category; (iii) the quantities required to remain unaffected when alter the eigenvalue should contain various categories.

Due to the similarity, the details are omitted. □