HIGH DIMENSIONAL CLASSIFICATION AND VARIABLE SELECTION

by

Quefeng Li

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

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2013

Date of final oral examination: 04/12/13

The dissertation is approved by the following members of the Final Oral Committee:
Professor Jun Shao, Department of Statistics
Professor Menggang Yu, Department of Biostatistics & Medical Informatics
Professor Yazhen Wang, Department of Statistics
Professor Chunming Zhang, Department of Statistics
Professor Bret Hanlon, Department of Statistics

To Mingquan, Zhiye and Yaoyao

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Professor Jun Shao, for his supervision of my study and research. During the past five years, Professor Shao has not only guided me through several exciting research projects, but also impacted me greatly on how to be a better researcher in general.

I would also like to thank other members in my dissertation committee: Professor Menggang Yu, Professor Yazhen Wang, Professor Chunming Zhang and Professor Bret Hanlon for their insightful comments and the time they dedicated to review my dissertation. Their suggestions have helped me greatly improve the work. Moreover, I would like to express my thanks to Professor Sijian Wang for his indispensable advice in one of the research projects. I thank the Department of Statistics at University of Wisconsin – Madison, which provided such a wonderful PhD program and continued to support me during the past five years.

Finally, I owe my deepest thanks to my father Mingquan Li and my mother Zhiye Xing, for their tremendous care and support throughout my life. In addition, a very special thanks to Yaoyao Xu, for her love and care in the past three years.

CONTENTS

Contents iii

List of Tables v

List of Figures vi

Abstract viii

I Introduction 1

1 Introduction 2

1.1 High Dimensional Discriminant Analysis 2

1.2 Variable Selection of High Dimensional Linear Regression 3

1.3 Structure of This Dissertation 5

II Sparse Quadratic Discriminant Analysis 7

2 Motivation 8

3 Preliminary Results 11

4 Sparse Estimators and SQDA 17

5 Numerical Work 26

5.1 A simulation comparison of the SLDA and SQDA 26

5.2 A Real Data Example 29

6 Proofs 32

6.1 Proofs of Lemmas 32

6.2 Proofs of Theorems 41

III Regularized LASSO in High Dimensional Linear Regression **55**

7 Motivation 56

8 The Methodology 60

9 Asymptotic Results 65

9.1 Results Under A Sparse Covariance Matrix 65

9.2 Results Under A Sparse Inverse of Covariance Matrix 70

10 Comparison of LASSO and RLASSO 74

11 Numerical Results 79

12 Proofs 85

IV Conclusion **96**

Bibliography 99

LIST OF TABLES

5.1 Misclassification Rate and Simulation Standard Error for Comparison of SQDA and SLDA 27

5.2 Average Misclassification Rates of Nine Classifiers for Colon Data 30

5.3 Quantiles of Misclassified Objects by the SLDA and SQDA for Colon Data 30

11.1 Simulation Results of Eight Methods under Model 1 82

11.2 Simulation Results of Eight Methods under Model 2 83

11.3 Simulation Results of Eight Methods under Model 3 83

11.4 Simulation Results of Eight Methods under Model 4 83

LIST OF FIGURES

5.1 Misclassification Rates of the SLDA and SQDA for Colon Data 31

10.1 Comparison of LASSO and RLASSO 77

ABSTRACT

Recent advances in biotechnology and other disciplines have led to the generation of many high-dimensional data, which raises challenges to develop new statistical methodologies to handle them.

This dissertation focuses on two aspects of high-dimensional data inference: (1) classification based on high-dimensional covariates; (2) variable selection of high-dimensional linear regression model. Both aspects have great importance in high-dimensional data inference and are related with each other. Variable selection plays a critical rule to reduce the dimension of data. It usually boosts the signal to noise ratio and results in a simpler model that becomes much easier to interpret. Classification has many important applications in practice, such as face detection, hand-writing recognition, etc.

For the high-dimensional classification problem, I have developed a new Sparse Quadratic Discriminant Analysis (SQDA) approach, which extends the application of traditional low-dimensional Quadratic Discriminant Analysis. The theoretical properties of the new SQDA approach is thoroughly addressed. Simulation studies have been conducted to compare SQDA with many other well-known classifiers in the literature. This new approach has also been applied to analyze one dataset from a colon cancer study.

For the variable selection problem, a Regularized LASSO approach has been proposed, which alleviates the strong conditions for the classical LASSO method to perform well. It has been found that the new Regularized LASSO approach includes many other well-known variable selection methods as its special cases, which makes

it a very general approach. The asymptotic properties of Regularized LASSO is thoroughly studied. It has been shown that the Regularized LASSO asymptotically identifies the correct model under mild assumptions. The new method has also been investigated through simulation studies, where it outperforms many other variable selection methods.

Part I

Introduction

1 INTRODUCTION

1.1 High Dimensional Discriminant Analysis

Classifying two classes based on n observations of p variables has long been a classical problem in statistics. In a novel paper, Fisher [20] assumed that the two classes share the same covariance and proposed an approach based on a linear combination the difference between two population means, which was referred as Linear Discriminant Analysis (LDA). LDA was further extended to Quadratic Discriminant Analysis, which allows the two classes to have different covariance.

An advantage of LDA and QDA is that they have an explicit form, which makes them widely used. When the dimension p is fixed, they both have adequate performance of classification. (e.g., see Anderson [2]) However, in many contemporary data sets, the number of covariates is much larger than the number of observations, i.e. $p \gg n$. This new feature brings much difficulty of the direct application of LDA and QDA to the high dimensional data.

In fact, when $p > n$, Bickel and Levina [4] and Shao et al. [30] showed that the LDA may be asymptotically as bad as random guessing. In other words, its misclassification rate tends to 1/2. This necessitates the modification of LDA to accommodate high dimensional data. There were some excellent advances in the literature. Shao et al. [30] proposed to replace Maximum Likelihood Estimators (MLE) in the LDA rule with the corresponding sparse estimators to obtain a Sparse LDA (SLDA) rule. They showed that SLDA is consistent to the Bayes rule under certain sparsity conditions on the difference between two populations means and on their common covariance. Cai

and Liu [10] modified classical LDA by obtaining a sparse representation of direction of the classification line. They showed that, when $p \gg n$, their method is also Fisher consistent, namely its misclassification rate converges to that of the optimal Bayes rule. Fan et al. [16] proposed a Regularized Optimal Affine Discriminant (ROAD) method by directly minimize the misclassification rate. Their method is also shown to be Fisher consistent.

However, all above method needs to assume the two classes have the same covariance. Very little is known for the asymptotic performance of the Quadratic Discriminant Analysis, even for the case where $n > p \to \infty$, Cheng [11] established some asymptotic results for the QDA, but under the much simplified situation where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are diagonal.

This motivates the study of Fisher consistency of general QDA and its extension to the case of $p \gg n$. The asymptotic misclassification rate of the general QDA will be investigated for both cases of $p < n$ and $p \gg n$. It will be shown that the classical QDA is still consistent when p diverges in a much smaller rate than n. When $p \gg n$, in order to achieve Fisher consistency, a sparse QDA (SQDA) method will be developed by using sparse estimate of means and covariances.

1.2 Variable Selection of High Dimensional Linear Regression

The second part of the dissertation focuses on variable selection of high dimensional linear regression model.

For the low dimensional setting, where $n \to \infty$ and p is fixed or $p \to \infty$ at a rate much slower than n, there are many variable selection tools, e.g. forward/backward selection based on AIC/BIC or other information criteria. For variable selection in the case of $p > n$ with $p = O(n^l)$ for some $l > 1$ or $O(e^{n^\nu})$ for some $\nu \in (0, 1)$ (ultra-high dimension), some excellent advances in asymptotic theory have been made recently. See, for example, Fan and Peng [19], Hunter and Li [23], Meinshausen and Bühlmann [24], Zhao and Yu [38], Zou [39], Wang et al. [33], Fan and Lv [17], Zhang and Huang [36], Meinshausen and Yu [25], Wang [32], Fan and Song [15], and a review paper by Fan and Lv [18].

A large part of these advances is to study the asymptotic behavior of LASSO, which is a novel method proposed by Tibshirani [31]. It is now well known that the LASSO method requires a very strong irrepresentable condition (see Zhao and Yu [38]) to achieve model selection consistency. Even if adding a thresholding step after LASSO, it is still too conservative especially when the number of explanatory variables p is much larger than the number of observations n. Another well-known method, the sure independence screening (SIS), applies thresholding to an estimator of marginal covariate effect vector and, therefore, is not selection consistent unless the zero components of the marginal covariate effect vector are asymptotically the same as the zero components of the regression effect vector.

Since the weakness of LASSO is caused by the fact that it utilizes the covariate sample covariance matrix that is not well behaved when p is larger than n, a new regularized LASSO (RLASSO) method will be proposed by replacing the covariate sample covariance matrix in LASSO with a regularized estimator of covariate covariance

matrix and adding a thresholding step. Using a regularized estimator of covariate covariance matrix, we can consistently estimate the regression effects and, hence, the new method also extends and improves the SIS method that estimates marginal covariate effects. The selection consistency of RLASSO will be established under conditions that the regression effect vector is sparse and the covariate covariance matrix or its inverse is sparse. It will be shown that some well-known variable selection methods, such as LASSO, LASSO followed by thresholding, SIS and scout method (see Witten and Tibshirani [35]) can be regarded as special cases of RLASSO. Some simulation results for comparing variable selection performances of RLASSO and various other methods will also be presented.

1.3 Structure of This Dissertation

This dissertation is organized as follows. Part II focuses on the problem of Sparse Quadratic Discriminant Analysis. In Chapter 3, we introduce some notation and preliminary results, including a result showing that the classical QDA has the smallest asymptotic misclassification rate when $p \to \infty$ but at a rate much slower than n, and an example indicating that it is necessary to regulate the difference of covariance matrices. The main results are presented in Chapter 4, where we first state some sparsity conditions on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ and construct sparse estimators of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ based on the training data, which results in our proposed SQDA classification rule. Asymptotic properties of sparse estimators and the SQDA are established under the sparsity conditions and some conditions on the divergence rate of p. Chapter 5

contains a simulation comparison between the SQDA and SLDA, and a microarray data example, in which we compare the SQDA with the SLDA and some other popular classifiers in the literature. All technical proofs are given Chapter 6.

Part III elaborates the results of Regularized LASSO (RLASSO) for variable selection of high dimensional linear regression model. The general methodology is introduced in Chapter 8. Chapter 9 contains results on the selection-consistency of the proposed method. A comparison of LASSO and RLASSO is given in Chapter 10. Chapter 11 provides some simulation results on the performance of the proposed method and several other variable selection methods. All corresponding proofs are given in Chapter 12.

Finally, I conclude this dissertation in Part IV.

Part II

Sparse Quadratic Discriminant Analysis

2 MOTIVATION

Consider the problem of classifying a p-dimensional normally distributed vector $\boldsymbol{x}$ into one of two classes represented by two p-dimensional normal distributions, $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\boldsymbol{\mu}_k$'s are mean vectors and $\boldsymbol{\Sigma}_k$'s are positive definite covariance matrices. If $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, $k = 1, 2$, are known, then an optimal classification rule having the smallest possible misclassification rate can be constructed. However, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, $k = 1, 2$, are usually unknown and a classification rule has to be constructed using a training sample to estimate unknown parameters. In the traditional setup where the dimension p of $\boldsymbol{x}$ is fixed, the well-known linear discriminant analysis (LDA) for the case of $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ or quadratic discriminant analysis (QDA) for the case of $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ has the smallest asymptotic misclassification rate in the sense that its misclassification rate converges to that of the optimal rule as the training sample size $n \to \infty$. In fact, Shao, Wang, Deng, and Wang [30] showed that the LDA still has the smallest asymptotic misclassification rate when p diverges to infinity at a rate slower than $\sqrt{n}$ as $n \to \infty$. A similar result for the QDA is established in this paper.

Nowadays, much more characteristics are collected simultaneously, which results in a high dimensional $\boldsymbol{x}$. In many recent applications, p is much larger than the training sample size n, which is referred to as the large-p-small-n problem or ultra-high dimension problem when p is of the order e^{n^ν} with a constant $\nu \in (0, 1)$. An example is a study with genetic or microarray data. In our example presented in Section 5.2, for instance, to classify tumor and normal colon tissues by Oligonucleotide microarray technique, $p = 2000$ genes are involved whereas the size of the sample is only $n = 62$. Other examples include data from radiology, biomedical imaging, signal processing,

climate, and finance. When $p > n$, Bickel and Levina [4] and Shao, Wang, Deng, and Wang [30] showed that the LDA may be asymptotically as bad as random guessing.

Some improvements over the LDA for large p problems have been made in recent years. See, for example, Bickel and Levina [4], Fan and Fan [14], Guo, Hastie, and Tibshirani [22], Clemmensen, Hastie, and Ersbøll [12], Qiao, Zhou, and Huang [27], and Zhang and Wang [37]. Moreover, Shao, Wang, Deng, and Wang [30] proposed a sparse LDA (SLDA) by thresholding and showed that it has the smallest asymptotic misclassification rate under some sparsity conditions on unknown parameters. To derive an asymptotically optimal classification rule is more difficult than variable selection, because we must identify not only components of $\boldsymbol{x}$ having mean effects for classification, but also components of $\boldsymbol{x}$ correlated with those having mean effects.

To the best of our knowledge, most theoretical work on the asymptotic misclassification rate of discriminant analysis assumes a common covariance matrix. Very little has been done for the QDA, even for the case where $p < n$. Cheng [11] established some asymptotic results for the QDA, but under the much simplified situation where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal.

The purpose of this paper is to construct a sparse QDA (SQDA) and establish its asymptotic optimality under some sparsity conditions on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, $k = 1, 2$. Although our proposed SQDA is based on the well-known thresholding methodology, the study of asymptotic properties of the SQDA is much more complicated and difficult than that for the SLDA studied in Shao, Wang, Deng, and Wang [30]. First, the misclassification rate of the LDA has a closed form, but the misclassification rate of the QDA does not, since it involves a probability related to a complicated quadratic

form of $\boldsymbol{x}$. Second, for a good performance of the QDA, we need sparsity conditions on each of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, and the difference $\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$. Otherwise, the QDA may be asymptotically as bad as random guessing. This is quite different from the LDA, in which we only need sparsity of $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. To accommodate our method to this issue, we construct mean estimators by thresholding and covariance matrix estimators by double thresholding, one for the covariance matrices and another for their differences. Finally, because of the existence of quadratic forms of $\boldsymbol{x}$, we have to handle convergence of estimated covariance matrices in terms of not only the usual L_2 norm, but also L_1 norm, the Frobenius norm, and another norm defined in Chapter 3. For the SLDA, however, only L_2 norm is needed. As by-products, we derived some results on convergence of estimated covariance matrices in terms of several norms that may be useful in other studies.

3 PRELIMINARY RESULTS

We start with some notation. For any vector $\boldsymbol{a}$, $\boldsymbol{a}'$ denotes its transpose and $\|\boldsymbol{a}\|$ denotes its L_2 norm. For any symmetric $p \times p$ matrix $\boldsymbol{A}$ whose (i,j)th element is a_{ij}, we define the following norms: $\|\boldsymbol{A}\|_G = \sum_{i=1}^p \sum_{j=1}^p |a_{ij}|$, $\|\boldsymbol{A}\|_F = (\sum_{i=1}^p \sum_{j=1}^p a_{ij}^2)^{1/2}$, $\|\boldsymbol{A}\|_1 = \max_{1\le i\le p} \sum_{j=1}^p |a_{ij}|$, and $\|\boldsymbol{A}\|_2 = \max_{1\le j\le p} |\lambda_{p,j}(\boldsymbol{A})|$, where $\lambda_{p,j}(\boldsymbol{A})$ is the jth smallest eigenvalue of $\boldsymbol{A}$. Note that $\|\boldsymbol{A}\|_2$ is the L_2 matrix norm, $\|\boldsymbol{A}\|_1$ is the L_1 matrix norm, which is the same as the L_∞ norm since $\boldsymbol{A}$ is symmetric; $\|\boldsymbol{A}\|_F$ is the Frobenius norm related to the L_2 norm and $\|\boldsymbol{A}\|_G$ is a counterpart of $\|\boldsymbol{A}\|_F$ related to the L_1 norm; $\|\boldsymbol{A}\|_2 \le \|\boldsymbol{A}\|_1 \le \|\boldsymbol{A}\|_G$ and $\|\boldsymbol{A}\|_2 \le \|\boldsymbol{A}\|_F$. The following lemma is useful in studying relationship among various matrix norms.

Lemma 1. *Let $\boldsymbol{L}$, $\boldsymbol{C}$, and $\boldsymbol{R}$ be symmetric $p \times p$ matrices. Then,*

$$\|\boldsymbol{LCR}\|_G \le \|\boldsymbol{L}\|_1 \|\boldsymbol{C}\|_G \|\boldsymbol{R}\|_1 \quad \textit{and} \quad \|\boldsymbol{LCR}\|_F \le \|\boldsymbol{L}\|_2 \|\boldsymbol{C}\|_F \|\boldsymbol{R}\|_2.$$

Let $\boldsymbol{\mu}_k$ be the mean and $\boldsymbol{\Sigma}_k$ be the covariance matrix of the p-dimensional normal distribution, $k = 1, 2$, $\boldsymbol{I}$ be the identity matrix of order p, and

$$\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1, \quad \boldsymbol{\Delta} = \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1, \quad \boldsymbol{\nabla} = \boldsymbol{\Sigma}_2^{-1} - \boldsymbol{\Sigma}_1^{-1}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1^{1/2} - \boldsymbol{I}.$$

Throughout this paper, we assume the following regularity condition on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$: there are positive constants m and M such that

(C1) all absolute values of components of $\boldsymbol{\mu}_k \leq M$;

(C2) $m \leq$ all eigenvalues of $\boldsymbol{\Sigma}_k \leq M$;

(C3) $m \leq \liminf_{p\to\infty} D_p$, where $D_p = \sqrt{\|\boldsymbol{\Delta}\|_F^2 + \|\boldsymbol{\delta}\|^2}$.

Condition (C3) avoids the trivial case where the two classes are the same as $p \to \infty$.

When $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are known, the optimal classification rule, which is often called the Bayes rule, classifies $\boldsymbol{x}$ to class 2 if and only if

$$(\boldsymbol{x} - \boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\boldsymbol{x} - \boldsymbol{\mu}_1) - 2\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta} - \log(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|) < 0. \qquad (3.1)$$

The Bayes rule in (3.1) has misclassification rate of

$$R_B = \frac{R_{B1} + R_{B2}}{2}, \qquad R_{Bk} = P\left(\text{incorrectly classify } \boldsymbol{x} \text{ to class } k\right). \qquad (3.2)$$

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, then the probabilities in R_B are related to normal distributions. Otherwise, these probabilities have no known form and we need the following result for establishing asymptotic results.

Lemma 2. *Suppose that (C1)-(C2) hold. Let* $\boldsymbol{z} \sim N_p(\boldsymbol{0}, \boldsymbol{I})$ *and* $T_p = \boldsymbol{z}'\boldsymbol{\Lambda}\boldsymbol{z} - 2\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{z}$. *If* $D_p \to \infty$ *as* $p \to \infty$, *then* $[T_p - \mathrm{E}(T_p)]/\sqrt{\mathrm{Var}(T_p)} \xrightarrow{D} N(0,1)$, *where* $\xrightarrow{D}$ *denotes convergence in distribution.*

In practice, since $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are unknown, the optimal rule cannot be used. To estimate $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, we assume that there is a training sample $\boldsymbol{X} = \{\boldsymbol{x}_{ki}, i = 1, ..., n_k, k = 1, 2\}$, where n_k is the sample size for class k, $\boldsymbol{x}_{ki} \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $k = 1, 2$, all $\boldsymbol{x}_{ki}$'s are independent, and $\boldsymbol{X}$ is independent of $\boldsymbol{x}$ to be classified. For any unknown

$\boldsymbol{a}$ or $\boldsymbol{A}$, let $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{A}}$ be their estimators based on the training sample $\boldsymbol{X}$. Then the sample analog of the optimal rule classifies $\boldsymbol{x}$ to class 2 if and only if

$$(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1) - 2\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1) + \hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \log(|\hat{\boldsymbol{\Sigma}}_1|/|\hat{\boldsymbol{\Sigma}}_2|) < 0. \tag{3.3}$$

Its conditional misclassification rate, given $\boldsymbol{X}$, is

$$R(\boldsymbol{X}) = \frac{R_1(\boldsymbol{X}) + R_2(\boldsymbol{X})}{2},$$

where

$$R_k(\boldsymbol{X}) = P\left(\text{incorrectly classify } \boldsymbol{x} \text{ to class } k \mid \boldsymbol{X}\right), \tag{3.4}$$

and the probability is with respect to $\boldsymbol{x}$ conditional on $\boldsymbol{X}$. Unlike the LDA case where $R(\boldsymbol{X})$ has a simple explicit form, the probability $R_k(\boldsymbol{X})$ is complicated and does not have an explicit form.

When $p < n$, the well known QDA is defined by (3.3) with $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$ being the maximum likelihood estimators (MLE) based on $\boldsymbol{X}$. When it is known that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, $\hat{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{\Sigma}}_2$ and (3.3) reduces to the LDA. The following result establishes some asymptotic properties of the QDA when $p < n$. The limiting process considered in this paper is $n = n_1 + n_2 \to \infty$ with $n_1/n \to$ a constant strictly between 0 and 1. Throughout, p is considered as a function of n and p may diverge to ∞ at a certain rate as $n \to \infty$.

Theorem 3.1. *Suppose that conditions (C1)-(C3) hold.*
(i) When D_p is bounded as $p \to \infty$, if $p = o(n^{1/5})$ and

(C4) the density function of T_p is bounded by a constant not depending on p, where T_p is defined as in Lemma 2,

then

$$R_{\mathrm{QDA}}(\boldsymbol{X}) - R_B \xrightarrow{P} 0, \tag{3.5}$$

where $R_{\mathrm{QDA}}(\boldsymbol{X})$ is the conditional misclassification rate of the QDA given the training data $\boldsymbol{X}$, R_B is the optimal rate of the Bayes rule, and $\xrightarrow{P}$ denotes convergence in probability.

(ii) When $D_p \to \infty$ as $p \to \infty$ and $p < n$, if $p^2/(nD_p^2) \to 0$, then (3.5) holds.

(C4) holds in the special case where $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. In fact, it holds when $\boldsymbol{\Lambda}$ has some eigenvalues that are always equal to 0. Another situation in which (C4) holds is when there are at least two eigenvalues of $\boldsymbol{\Lambda}$ in $(-\infty, m]$ or $[m, \infty)$ (see the proof of Theorem 3).

When $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ and $p > n$, the results in Bickel and Levina [4] and Shao, Wang, Deng, and Wang [30] indicated that some sparsity conditions on $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ are necessary in order to obtain an asymptotically optimal classification rule. When $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ and $p > n$, we need sparsity conditions on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, $k = 1, 2$, since all of them are involved in the Bayes rule (3.1) and their estimators are involved in the classifier (3.3). Furthermore, as the following discussion indicates, some condition on $\boldsymbol{\Delta}$, the difference between two covariance matrices, is also necessary.

We consider a special case where $p/n \to \infty$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are known, but $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are unknown. Then, we only need to estimate the means. In this case, the QDA

classifies $\boldsymbol{x}$ to class 2 if and only if

$$\hat{T}_p - \mathrm{E}(\hat{T}_p|\boldsymbol{X}) < -\hat{\phi}_p,$$

where $\hat{T}_p = (\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)'\boldsymbol{\nabla}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1) - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)$, $\mathrm{E}(\hat{T}_p|\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{\Lambda}) + (\hat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1) - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_1)$ when $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, and $\hat{\phi}_p = \mathrm{tr}(\boldsymbol{\Lambda}) - \log(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|) + (\hat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1) - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_1) + \hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}$. We now show that the conditional misclassification rate of QDA converges to 1/2 when $\|\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1\|_F \to \infty$ and $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ (but we do not know $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$).

Following the proof of Lemma 2, we can show that

$$[\hat{T}_p - \mathrm{E}(\hat{T}_p|\boldsymbol{X})]/[\mathrm{Var}(\hat{T}_p|\boldsymbol{X})]^{1/2} \xrightarrow{D|\boldsymbol{X}} N(0,1),$$

where $\mathrm{Var}(\hat{T}_p|\boldsymbol{X}) = 2\|\boldsymbol{\Lambda}\|_F^2 + 4\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}$ and the convergence is with respect to the distribution of the new observation $\boldsymbol{x}$, conditioned on $\boldsymbol{X}$. Under (C2),

$$-1 < \frac{m}{M} - 1 \le \lambda_{p,j} \le \frac{M}{m} - 1, \quad j = 1, ..., p, \tag{3.6}$$

which implies

$$\left|\mathrm{tr}(\boldsymbol{\Lambda}) - \log(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|)\right| \le \frac{M^2}{2m^2}\|\boldsymbol{\Lambda}\|_F^2.$$

Then, $|\hat{\phi}_p|/[\mathrm{Var}(\hat{T}_p|\boldsymbol{X})]^{1/2}$ is bounded by

$$\frac{\frac{M^2}{2m^2}\|\boldsymbol{\Lambda}\|_F^2 + |2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_1) - \hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_1^{-1}\hat{\boldsymbol{\delta}}| + |(\hat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}_1-\boldsymbol{\mu}_1)|}{\sqrt{2\|\boldsymbol{\Lambda}\|_F^2 + 4\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}}}. \tag{3.7}$$

We consider $\boldsymbol{\Sigma}_1 = \boldsymbol{I}$, a diagonal $\boldsymbol{\Sigma}_2$ with jth diagonal $\sigma_{2j}^2 = 2$ for $j = 1, ..., K$, $\sigma_{2j}^2 = (\sqrt{17} - 3)/2$ for $j = K+1, ..., 2K$, $\sigma_{2j}^2 = 1$ for $j = 2K+1, ..., p$, and $n_1 = n_2 = n/2$. Then, $\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1 \sim N(0, \frac{2}{n}\boldsymbol{I})$ and $\hat{\boldsymbol{\mu}}_2 - \boldsymbol{\mu}_2 \sim N(0, \frac{2}{n}\boldsymbol{\Sigma}_2)$. Let ϵ_{1j} and ϵ_{2j} be independent standard normal random variables. Then,

$$\begin{aligned}\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_1^{-1}\hat{\boldsymbol{\delta}} - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1) &= \frac{2}{n}\sum_{j=1}^{p}\left[(\sigma_{2j}\epsilon_{2j} - \epsilon_{1j})^2 + 2\frac{\epsilon_{1j}(\sigma_{2j}\epsilon_{2j} - \epsilon_{1j})}{\sigma_{2j}^2}\right] \\ &= \frac{2}{n}\sum_{j=1}^{p}\left[\left(1 - \frac{2}{\sigma_{2j}^2}\right)\epsilon_{1j}^2 + 2\left(\frac{1}{\sigma_{2j}} - \sigma_{2j}\right)\epsilon_{1j}\epsilon_{2j} + \sigma_{2j}^2\epsilon_{2j}^2\right],\end{aligned}$$

which has mean 0 for the particular set of σ_{2j}^2's we have chosen. Hence, by the Central Limit Theorem, $\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_1^{-1}\hat{\boldsymbol{\delta}} - 2\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1) = O_P\left(\sqrt{p}/n\right)$. Also, $4\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}} = O_P\left(p/n\right)$, $\|\boldsymbol{\Lambda}\|_F^2 = (11 - \sqrt{17})K/8 = O(K)$, and $(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\hat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1) = O_P\left(K/n\right)$. Therefore, the quantity in (3.7) is bounded by

$$\frac{|\hat{\phi}_p|}{\sqrt{\mathrm{Var}(\hat{T}_p|\boldsymbol{X})}} \leq \frac{O(K) + O_P(\sqrt{p}/n) + O_P(K/n)}{\sqrt{O(K) + O_P(p/n)}},$$

which is $o(1/\sqrt{n})$ if we choose $K = o(\sqrt{p}/n)$. This together with the asymptotic normality of $\hat{T}_p$ shows that the conditional misclassification rate of the QDA converges to 1/2, provided that $K = o(\sqrt{p}/n)$.

This example shows that if no condition is imposed on $\boldsymbol{\Delta} = \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1$, the QDA could asymptotically be as bad as random guessing, which is caused by the accumulated errors in estimating components of $\boldsymbol{\mu}_k$'s corresponding to small Δ_{ij}'s when $\|\boldsymbol{\Delta}\|_F$ diverges with a certain rate.

4 SPARSE ESTIMATORS AND SQDA

We first define some sparsity measures on population parameters. For $\boldsymbol{\mu}_k$, we adopt the sparsity measure in Bickel and Levina [5],

$$d_p = \max_{k=1,2} \sum_{j=1}^{p} |\mu_{kj}|^{2g}, \tag{4.1}$$

where μ_{kj} is the jth component of $\boldsymbol{\mu}_k$ and g is a constant in $[0, 1)$. As $n \to \infty$, d_p may diverge to ∞, but if its divergence rate is much slower than p, then $\boldsymbol{\mu}_k$'s are sparse. If $g = 0$, then d_p is the maximum of the numbers of non-zero components of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Similarly, we consider the following sparsity measure for covariance matrices:

$$c_p = \max_{k=1,2} \max_{i=1,\ldots,p} \sum_{j=1}^{p} |\sigma_{kij}|^h, \tag{4.2}$$

where σ_{kij} is the (i, j)th element of $\boldsymbol{\Sigma}_k$ and h is a constant in $[0, 1)$. When c_p is much smaller than p, $\boldsymbol{\Sigma}_k$'s are sparse in terms of off-diagonal values, but the diagonal elements of $\boldsymbol{\Sigma}_k$'s are not sparse.

As discussed in the end of Chapter 3, we need to regulate the magnitude of $\boldsymbol{\Delta}$ in some sense. We consider the following sparsity measure:

$$c_{1p} = \sum_{1 \le i,\, j \le p} |\Delta_{ij}|^{\eta}, \tag{4.3}$$

where Δ_{ij} is the (i, j)th element of $\boldsymbol{\Delta}$ and η is a constant in $[0, 1)$. If c_{1p} is much smaller than p, then $\boldsymbol{\Delta}$ is sparse. Unless otherwise mentioned, we eliminate the case

of $c_{1p} = 0$, where there is no need to consider QDA.

We allow $p > n$ to be ultra-high, but assume the following condition on the divergence of p as $n \to \infty$:

(C5) $$n^{-1} \log p \to 0.$$

Condition (C5) allows that p diverges at the rate e^{n^ν} for some $\nu \in (0, 1)$.

We need to construct sparse estimators of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, and $\boldsymbol{\Delta}$, since these estimators are all used in the QDA rule (3.3). This is different from the LDA where only sparse estimators of $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}$ are needed. The constructed sparse estimators must be asymptotically valid in terms of several measures.

The estimation of $\boldsymbol{\mu}_k$ is simple. A sparse estimator of $\boldsymbol{\mu}_k$ is obtained by thresholding the MLE $\bar{\boldsymbol{x}}_k = n_k^{-1} \sum_{i=1}^{n_k} \boldsymbol{x}_{ki}$ at

$$t_n = M_0 \left(n^{-1} \log p \right)^\alpha \tag{4.4}$$

for some constants $\alpha \in (0, 1/2)$ and $M_0 > 0$. That is, the thresholded estimator of $\boldsymbol{\mu}_k$ is $\hat{\boldsymbol{\mu}}_k$ whose jth component is $\bar{x}_{kj} I(|\bar{x}_{kj}| > t_n)$, where $I(A)$ is the indicator function of the event A and $\bar{x}_{kj}$ is the jth component of $\bar{\boldsymbol{x}}_k$. The parameter $\boldsymbol{\delta}$ is then estimated by $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1$.

Let q_{kn} be the number of components of $\boldsymbol{\mu}_k$ whose absolute values are larger than t_n/r with a constant $r > 1$, $k = 1, 2$. From the proof of Theorem 3 in Shao, Wang, Deng, and Wang [30], if

(S1) $$b_n = \max\{d_p t_n^{2(1-g)}, q_{1n}/n, q_{2n}/n\} \to 0,$$

where d_p is given by (4.1), then

$$\|\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|^2 = O_P\left(b_n\right). \tag{4.5}$$

Note that (4.5) also holds with $\|\hat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k\|^2$ replaced by $\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2$.

The estimation of $\boldsymbol{\Sigma}_k$ is more complicated, since we need estimators of $\boldsymbol{\Sigma}_k$'s to be sparse in terms of off-diagonal elements as well as a spare estimator of $\boldsymbol{\Delta}$ in view of the discussion in the end of Chapter 3. We propose an estimator by two steps of thresholding. Let $\boldsymbol{S}_k$ be the MLE of $\boldsymbol{\Sigma}_k$ based on $\{\boldsymbol{x}_{ki}, i = 1, ..., n_k\}$ and let s_{kij} be the (i, j)th element of $\boldsymbol{S}_k$, $k = 1, 2$. $\boldsymbol{S}_2 - \boldsymbol{S}_1$ is a natural estimator of $\boldsymbol{\Delta}$, but it is not sparse. In the first step, small elements of $\boldsymbol{S}_2 - \boldsymbol{S}_1$ are thresholded to 0. That is, we replace s_{1ij} and s_{2ij} by $\bar{s}_{ij} = (n_1 s_{1ij} + n_2 s_{2ij})/n$ whenever $|s_{1ij} - s_{2ij}|$ is less than or equal to the threshold value

$$t_{1n} = M_1 \left(n^{-1} \log p\right)^{1/2}, \tag{4.6}$$

where M_1 is a constant. This produces an estimator of $\boldsymbol{\Sigma}_k$, $\tilde{\boldsymbol{\Sigma}}_k$, whose (i, j)th element $\tilde{s}_{kij} = \bar{s}_{ij}$ when $|s_{1ij} - s_{2ij}| \leq t_{1n}$ and $\tilde{s}_{kij} = s_{kij}$ otherwise, $k = 1, 2$. Although $\tilde{\boldsymbol{\Sigma}}_2 - \tilde{\boldsymbol{\Sigma}}_1$ is sparse, each $\tilde{\boldsymbol{\Sigma}}_k$ may not be sparse in terms of its off-diagonal elements. Hence, we apply the second step of thresholding to the elements of $\tilde{\boldsymbol{\Sigma}}_k$, which results in the estimator $\hat{\boldsymbol{\Sigma}}_k$ whose (i, j)th element is $\tilde{s}_{kij} I(|\tilde{s}_{kij}| > t_{2n})$, $k = 1, 2$, where t_{2n} is given by (4.6) with M_1 replaced by a possibly different constant M_2. The resulting estimator $\hat{\boldsymbol{\Sigma}}_k$ is sparse in terms of its off-diagonal elements and $\hat{\boldsymbol{\Sigma}}_2 - \hat{\boldsymbol{\Sigma}}_1$ is sparse.

Note that

$$\begin{aligned}\max_{i,j}|\tilde{s}_{1ij}-\sigma_{1ij}| &= \max_{i,j}\{|s_{1ij}-\sigma_{1ij}|I(|s_{1ij}-s_{2ij}|\geq t_{1n}) \\ &\quad + |\bar{s}_{ij}-\sigma_{1ij}|I(|s_{1ij}-s_{2ij}|<t_{1n})\} \\ &\leq \max_{i,j}\{|s_{1ij}-\sigma_{1ij}|+|s_{1ij}-s_{2ij}|I(|s_{1ij}-s_{2ij}|<t_{1n})\} \\ &\leq \max_{i,j}|s_{1ij}-\sigma_{1ij}|+M_1(n^{-1}\log p)^{1/2} \\ &= O_P\left((n^{-1}\log p)^{1/2}\right),\end{aligned}$$

where the last equality follows from result (12) of Bickel and Levina [5]. Similarly,

$$\max_{i,j}|\tilde{s}_{2ij}-\sigma_{2ij}| = O_P\left((n^{-1}\log p)^{1/2}\right).$$

Therefore, following the proof of Theorem 1 in Bickel and Levina [5], we can establish that, if (C2) and (C5) hold and

(S2) $$a_n = c_p(n^{-1}\log p)^{(1-h)/2}\to 0,$$

where c_p is given by (4.2), then

$$\|\hat{\mathbf{\Sigma}}_k-\mathbf{\Sigma}_k\|_2 = O_P\left(a_n\right), \qquad k=1,2. \tag{4.7}$$

Hence, $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$ are asymptotically invertible and result (12.3) also holds with $\|\hat{\mathbf{\Sigma}}_k-\mathbf{\Sigma}_k\|_2$ replaced by $\|\hat{\mathbf{\Sigma}}_k^{-1}-\mathbf{\Sigma}_k^{-1}\|_2$. In fact, from the proof in Bickel and Levina [5], (12.3) still holds with $\|\hat{\mathbf{\Sigma}}_k-\mathbf{\Sigma}_k\|_2$ replaced by $\|\hat{\mathbf{\Sigma}}_k-\mathbf{\Sigma}_k\|_1$. The following lemma, which is useful in our proofs, gives the upper bound of $\|\hat{\mathbf{\Sigma}}_k^{-1}-\mathbf{\Sigma}_k^{-1}\|_1$.

Lemma 3. *Under (C2) and (C5), if $a_n v_p \to 0$, where $v_p = \max\{\|\mathbf{\Sigma}_1^{-1}\|_1, \|\mathbf{\Sigma}_2^{-1}\|_1\}$ and a_n is as in (S2), it holds that*

$$\|\hat{\mathbf{\Sigma}}_k^{-1} - \mathbf{\Sigma}_k^{-1}\|_1 = O_P\left(a_n v_p^2\right), \qquad k = 1, 2.$$

We estimate $\mathbf{\Delta}$ by $\hat{\mathbf{\Delta}} = \hat{\mathbf{\Sigma}}_2 - \hat{\mathbf{\Sigma}}_1$ and $\mathbf{\nabla}$ by $\hat{\mathbf{\nabla}} = \hat{\mathbf{\Sigma}}_2^{-1} - \hat{\mathbf{\Sigma}}_1^{-1}$. The previous discussion shows the upper bounds of $\hat{\mathbf{\Delta}} - \mathbf{\Delta}$ and $\hat{\mathbf{\nabla}} - \mathbf{\nabla}$ in both L_1 and L_2 norms. The following theorem shows a stronger result: the sparse estimator $\hat{\mathbf{\Delta}}$ converges to $\mathbf{\Delta}$ in the $\|\cdot\|_G$ norm under a sparsity condition on $\mathbf{\Delta}$, and $\hat{\mathbf{\nabla}}$ also converges to $\mathbf{\nabla}$ in $\|\cdot\|_G$ norm. The result is not only useful for the purpose of establishing asymptotic properties of the SQDA, but also interesting on its own.

Theorem 4.1. *Assume that (C2)-(C3) and (C5) hold.*
(i) If $a_{1n} = c_{1p}(n^{-1}\log p)^{(1-\eta)/2} \to 0$, then

$$\|\hat{\mathbf{\Delta}} - \mathbf{\Delta}\|_G = O_P\left(a_{1n}\right).$$

(ii) If

(S3) $$\tau_n = c_{1p} c_p v_p^3 (n^{-1}\log p)^{(1-\max\{h,\eta\})/2} \to 0,$$

where v_p is as defined in Lemma 3, then

$$\|\hat{\mathbf{\nabla}} - \mathbf{\nabla}\|_G = O_P(\tau_n).$$

We define the SQDA to be the classification rule (3.3) with the sparse estimators

previously described. Note that we allow the number of non-zero estimators (for the mean differences or covariances) to be much larger than n. This is different from variable selection and is necessary when there are many components of $\boldsymbol{x}$ that have no mean effects for classification but are correlated with those having mean effects.

Under some conditions, we now establish that the conditional misclassification rate of the SQDA converges to the same limit as R_B, the misclassification rate of the Bayes rule. To this end, we study the difference between the left hand sides of (3.1) and (3.3). The results are stated in the following lemmas. Note that Lemmas 6-7 (together with Lemmas 2-3 and Theorem 2) are used to handle the quadratic term and the difference in covariance matrices in the SQDA rule.

Lemma 4. *Under sparsity conditions (S1), (S2), (C2)-(C3) and (C5), if* $\|\boldsymbol{\delta}\|$ *is bounded, then, when* $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$,

$$\left|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1) - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right| = O_P\left(\max\{\sqrt{b_n}, a_n\}\right).$$

Lemma 5. *Under sparsity conditions (S1), (S2), (C2)-(C3) and (C5), if* $\|\boldsymbol{\delta}\|$ *is bounded, then,*

$$|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}| = O_P\left(\max\{\sqrt{b_n}, a_n\}\right).$$

Lemma 6. *Under sparsity conditions (S1), (S3), (C2)-(C3) and (C5), when* $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$,

$$\left|(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right| = O_P\left(\max\{\sqrt{b_n}, \tau_n\}\right).$$

Lemma 7. *Under sparsity conditions (S1), (S3), (C2)-(C3) and (C5),*

$$\left|\mathrm{tr}(\hat{\mathbf{\Lambda}}) - \mathrm{tr}(\mathbf{\Lambda})\right| = O_P(\tau_n) \quad \textit{and} \quad \left|\log(|\hat{\mathbf{\Sigma}}_1|/|\hat{\mathbf{\Sigma}}_2|) - \log(|\mathbf{\Sigma}_1|/|\mathbf{\Sigma}_2|)\right| = O_P(\tau_n).$$

By using Lemmas 4-7, we establish the following result for the SQDA.

Theorem 4.2. *Suppose that conditions (C1)-(C3) and (C5) hold.*
(i) When D_p is bounded as $p \to \infty$, if (C4) holds and

$$\max\{\sqrt{b_n}, a_n, \tau_n\} \to 0,$$

then

$$R_{\mathrm{SQDA}}(\boldsymbol{X}) - R_B \xrightarrow{P} 0, \tag{4.8}$$

where $R_{\mathrm{SQDA}}(\boldsymbol{X})$ is the conditional misclassification rate of the SQDA given $\boldsymbol{X}$ and R_B is the optimal misclassification rate of the Bayes rule in (3.1).
(ii) When $D_p \to \infty$ as $p \to \infty$, if $a_n \to 0$ and

$$\max\{b_n, a_{1n}\}/D_p^2 \to 0,$$

then (4.8) holds.

The tuning parameters M_0, M_1, and M_2 can be selected by minimizing the leave-one-out cross-validation estimate of the misclassification rate. We examine this method numerically in Chapter 5.

When $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, our SQDA is actually almost the same as the SLDA in Shao,

Wang, Deng, and Wang [30]. By choosing the tuning parameter M_1 large enough, the covariance matrix estimate in the SQDA is identical to that in the SLDA. As for the estimation of $\boldsymbol{\delta}$, Shao, Wang, Deng, and Wang [30] threshold $\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_1$ to obtain the estimate $\tilde{\boldsymbol{\delta}}$, while we separately threshold $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$ to obtain the estimate $\hat{\boldsymbol{\delta}}$. Under sparsity condition (S1),

$$\|\tilde{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2 = O_P\left(b_n\right) = \|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2.$$

For a fixed n, with any data-driven selection of the tuning parameters, the SQDA may be slightly better or worse than the SLDA.

On the other hand, if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, the SQDA may be much better than the SLDA in terms of the asymptotic misclassification rate. In Shao, Wang, Deng, and Wang [30], an estimator of the covariance matrix (assuming that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$) is obtained by thresholding

$$\boldsymbol{S} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(\boldsymbol{x}_{ki} - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_{ki} - \bar{\boldsymbol{x}}_k)',$$

which converges in L_2 norm to $\boldsymbol{\Sigma}^* = \gamma\boldsymbol{\Sigma}_1 + (1-\gamma)\boldsymbol{\Sigma}_2$ when actually $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$, where $n_1/n \to \gamma \in (0,1)$. Let s_{ij} and σ^*_{ij} be the (i,j)th element of $\boldsymbol{S}$ and $\boldsymbol{\Sigma}^*$, respectively. Then,

$$\begin{aligned}
\max_{i,j}|s_{ij} - \sigma^*_{ij}| &\leq \max_{i,j}\left(\left|\frac{n_1 s_{1ij}}{n} - \gamma\sigma_{1ij}\right|\right) + \left|\frac{n_2 s_{2ij}}{n} - (1-\gamma)\sigma_{2ij}\right| \\
&\leq \max_{i,j}[|s_{1ij} - \sigma_{1ij}| + |s_{2ij} - \sigma_{2ij}| + |n_1/n - \gamma||\sigma_{1ij}| \\
&\quad + |n_2/n - (1-\gamma)||\sigma_{2ij}|]
\end{aligned}$$

$$= O_P\left([n^{-1}\log p]^{1/2}\right) + O(|n_1/n - \gamma|).$$

If $n_2/n - \gamma = O((n^{-1}\log p)^{1/2})$, then $\max_{i,j}|s_{ij} - \sigma^*_{ij}| = O_P\left([n^{-1}\log p]^{1/2}\right)$. With this result, we can show that $\|\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_2 = O_P(a_n)$, where $\tilde{\boldsymbol{\Sigma}}$ is $\boldsymbol{S}$ thresholded at $M_3(\log p/n)^{1/2}$ with a constant $M_3 > 0$. Then, under the regularity conditions stated in Theorem 3 of Shao, Wang, Deng, and Wang [30],

$$R_{\text{SLDA}}(\boldsymbol{X}) - \Phi(-\sqrt{\boldsymbol{\delta}'\boldsymbol{\Sigma}^*\boldsymbol{\delta}}/2) \xrightarrow{P} 0.$$

If $\|\boldsymbol{\delta}\|$ is bounded as $p \to \infty$, then

$$\liminf_{p\to\infty} \Phi(-\sqrt{\boldsymbol{\delta}'\boldsymbol{\Sigma}^*\boldsymbol{\delta}}/2) > \liminf_{p\to\infty} R_B \geq 0,$$

because R_B is the misclassification rate of the Bayes rule. These results, together with the result in Theorem 3, imply that

$$\lim_{n\to\infty} P\Big(R_{\text{SLDA}}(\boldsymbol{X}) > R_{\text{SQDA}}(\boldsymbol{X}) + \epsilon_0\Big) = 1$$

for some fixed $\epsilon_0 > 0$. Thus, the SQDA is better than the SLDA in terms of misclassification rate.

If $\|\boldsymbol{\delta}\| \to \infty$, then R_B, $R_{\text{SLDA}}(\boldsymbol{X})$, and $R_{\text{SQDA}}(\boldsymbol{X})$ all converge to 0, and the asymptotic relative performance between the SLDA and SQDA depends on $\|\boldsymbol{\delta}\|$ and $\|\boldsymbol{\Delta}\|_F$ in a complicated manner. We compare the SLDA and SQDA in a simulation study in the next section.

5 NUMERICAL WORK

We first compare the SLDA and SQDA in a simulation study. Then, we consider a microarray data example, in which we compare the SQDA with the SLDA and some other popular classifiers in the literature.

5.1 A simulation comparison of the SLDA and SQDA

We consider the following two scenarios for the mean vectors:

$$A. \quad \boldsymbol{\mu}_1 = (1, \mathbf{0}_{p-1})', \quad \boldsymbol{\mu}_2 = (2, \mathbf{0}_{p-1})', \quad \|\boldsymbol{\delta}\| = 1,$$

$$B. \quad \boldsymbol{\mu}_1 = (\boldsymbol{e}_5, \mathbf{0}_{p-5})', \quad \boldsymbol{\mu}_2 = (3\boldsymbol{e}_5, \mathbf{0}_{p-5})', \quad \|\boldsymbol{\delta}\| = 4.47,$$

where $\boldsymbol{e}_t$ denotes a t-dimensional vector of 1's and $\mathbf{0}_t$ denotes a t-dimensional vector of 0's. For the covariance matrices, we consider the following three cases:

$$1. \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} \boldsymbol{B} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-5} \end{pmatrix}, \quad \|\boldsymbol{\Delta}\|_F = 0,$$

$$2. \quad \boldsymbol{\Sigma}_1 = \boldsymbol{I}_p, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} \boldsymbol{B} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-5} \end{pmatrix}, \quad \|\boldsymbol{\Delta}\|_F = 8.92,$$

$$3. \quad \boldsymbol{\Sigma}_1 = \boldsymbol{I}_p, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2\boldsymbol{B} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-5} \end{pmatrix}, \quad \|\boldsymbol{\Delta}\|_F = 16.82,$$

where $\boldsymbol{I}_t$ denotes the identity matrix of order t and

$$\boldsymbol{B} = \begin{pmatrix} 4 & 1 & 0.5 & 0 & 0 \\ 1 & 4 & 1 & 0.5 & 0 \\ 0.5 & 1 & 4 & 1 & 0.5 \\ 0 & 0.5 & 1 & 4 & 1 \\ 0 & 0 & 0.5 & 1 & 4 \end{pmatrix}.$$

In each scenario, we consider $n_1 = n_2 = 20$ and $p = 50$ or 200. The tuning parameters M_0, M_1 and M_2 are chosen by minimizing the leave-one-out cross-validation estimate of the misclassification rate. The same procedure is used to choose thresholds in the SLDA. The misclassification rates of the SLDA and SQDA calculated based on 50 simulations are shown in Table 5.1.

Table 5.1: Misclassification Rate (in %) and Simulation Standard Error (in brackets)

		mean scenario A		mean scenario B	
		SLDA	SQDA	SLDA	SQDA
covariance case 1	$p = 50$	40.9 (9.1)	36.9 (7.0)	19.1 (6.4)	19.2 (6.1)
	$p = 200$	40.1 (10.1)	44.5 (9.2)	22.9 (6.8)	19.7 (7.3)
covariance case 2	$p = 50$	40.7 (9.2)	23.1 (7.0)	10.1 (4.9)	8.8 (4.2)
	$p = 200$	44.4 (10.4)	32.6 (8.4)	18.3 (6.9)	16.9 (6.4)
covariance case 3	$p = 50$	42.2 (9.0)	10.4 (4.5)	15.7 (5.8)	6.1 (3.2)
	$p = 200$	45.7 (10.7)	20.8 (6.3)	26.8 (8.5)	14.0 (4.8)

The following is a summary of the results in Table 5.1.

I. Mean scenario A.

1. For covariance case 1 ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$), both SLDA and SQDA do not perform well,

since the signal strength $\|\boldsymbol{\delta}\| = 1$ is low and $\|\boldsymbol{\Delta}\|_F = 0$. The SQDA may be slightly better or worse than the SLDA.

2. Under covariance case 2 or 3, $\|\boldsymbol{\Delta}\|_F$ is much larger than that in case 1, although $\|\boldsymbol{\delta}\|$ is small. The SQDA is clearly better than the SLDA.

II. Mean scenario B.

1. Both the SLDA and SQDA are substantially better when $\|\boldsymbol{\delta}\|$ is larger. For covariance case 1 ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$), the SLDA and SQDA have similar performances.
2. Under covariance case 2, the performance of the SLDA is better than those in the other cases, including covariance case 1 in which $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. This is probably because, under covariance case 2, $\boldsymbol{\Sigma}_1$ is more sparse than the $\boldsymbol{\Sigma}_1$ in covariance case 1. This shows the importance of the sparsity of covariance matrices. Under covariance case 2, the SQDA is only slightly better than the SLDA. This may be because $\|\boldsymbol{\Delta}\|_F$ is not large enough to payoff the cost of estimating more quantities by applying the SQDA.
3. Under covariance case 3, $\|\boldsymbol{\Delta}\|_F$ is large and the SLDA starts to break down, while the performance of the SQDA is slightly better than that in case 2 and is much better than that of the SLDA.

To conclude, a large difference in $\boldsymbol{\mu}_k$'s is needed to have a good performance of the SLDA. The same is true for the SQDA, but the SQDA may still be good when there is a large difference in covariance matrices. The SLDA and SQDA are similar when $\|\boldsymbol{\Delta}\|_F$ is moderate but the SQDA outperforms the SLDA substantially when $\|\boldsymbol{\Delta}\|_F$ is large.

5.2 A Real Data Example

Alon, Barkai, Notterman, Gish, Ybarra, Mack, and Levine [1] studied gene expression difference between tumor and normal colon tissues using the Oligonucleotide microarray technique. The dataset contains $n_1 = 20$ observations from normal tissues and $n_2 = 42$ observations from tumor tissues. A total of $p = 2000$ genes with highest minimal intensity is included in the study. Dettling [13] used this dataset to compare the performance of seven different classifiers, namely, the Boosting, Bagging and boosting (BagBoost), Support Vector Machine (SVM), random forest (RanFor), the k nearest neighbor (kNN), the nearest shrunken centroid classifier (PAM) and diagonal LDA (DLDA), which applies the LDA by assuming that $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$ is a diagonal matrix. The dataset was randomly split into a training set of 13 observations from normal tissues and 29 observations from tumor tissues and a test set of 7 observations from normal tissues and 13 observations from tumor tissues. For each aforementioned classifier, a misclassification rate was calculated by classifying observations in the test set using the rule constructed based on the training set. To reduce variability, Dettling [13] independently repeated this process 50 times and reported the average misclassification rates of the seven classifiers over the 50 random splitting. The results are listed in our Table 5.2.

To compare, we added the average misclassification rates of the SLDA and SQDA calculated using the same random splitting process but a different random seed since we do not have information about the random seed in Dettling [13]. We used the same procedure as in the simulation study to choose the tuning parameters in the SQDA and SLDA. The results are given in Table 5.2. In this example, the SQDA is

the best among all classifiers. The SLDA, slightly behind the PAM, is actually the third winner.

The absolute gain in misclassification rate for the SQDA over the SLDA is 1.8%, which represents a relative gain of $(12.2\% - 10.4\%)/12.2\% = 14.8\%$. To further compare the SLDA and SQDA, Figure 5.1 displays a boxplot of the SLDA and SQDA misclassification rates in 50 replications and Table 5.3 lists some quantiles of numbers of misclassified subjects by the SLDA and SQDA in 50 replications.

Table 5.2: Average Misclassification Rates (in %) of Nine Classifiers for Colon Data

BagBoost	RanFor	SVM	kNN	DLDA	Boosting	PAM	SLDA	SQDA
16.10	14.86	15.05	16.38	12.86	19.14	11.90	12.20	10.40

Table 5.3: Quantiles of Misclassified Objects by the SLDA and SQDA for Colon Data

	Min.	25%	Median	75%	Max.
SLDA	0	1.25	2.5	3	6
SQDA	1	1	2	2	5

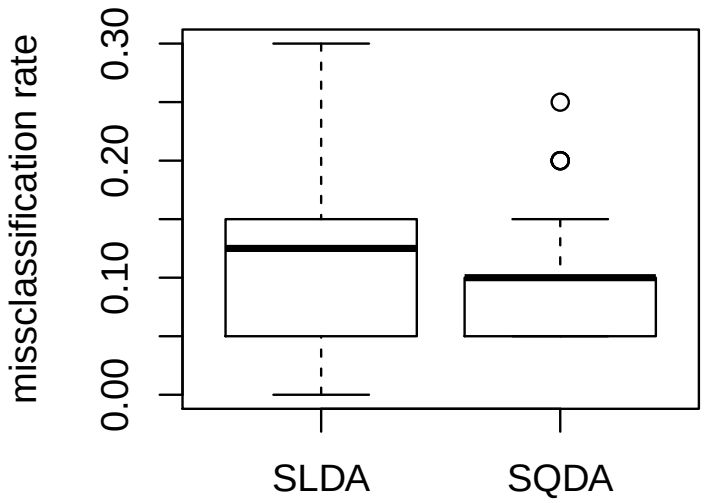

Figure 5.1: Misclassification Rates of the SLDA and SQDA for Colon Data

6 PROOFS

6.1 Proofs of Lemmas

Proof of Lemma 1. The first result follows from

$$\begin{aligned}
\|\boldsymbol{LCR}\|_G &= \sum_i \sum_j \left| \sum_k \sum_h l_{ik} c_{kh} r_{hj} \right| \\
&\leq \sum_i \sum_j \sum_k \sum_h |l_{ik} c_{kh} r_{hj}| \\
&= \sum_i \sum_k \sum_h \left(|l_{ik} c_{kh}| \cdot \sum_j |r_{hj}| \right) \\
&\leq \sum_i \sum_k \sum_h |l_{ik} c_{kh}| \left(\max_h \sum_j |r_{hj}| \right) \\
&= \|\boldsymbol{R}\|_1 \sum_h \sum_k |c_{kh}| \cdot \sum_i |l_{ik}| \\
&\leq \|\boldsymbol{R}\|_1 \sum_h \left(\sum_k |c_{kh}| \right) \left(\max_k \sum_i |l_{ik}| \right) \\
&= \|\boldsymbol{R}\|_1 \|\boldsymbol{L}\|_1 \|\boldsymbol{C}\|_G,
\end{aligned}$$

where the last equality follows from the fact that $\boldsymbol{L}$ is symmetric.

Represent $\boldsymbol{L}$ and $\boldsymbol{R}$ as their spectral decompositions, $\boldsymbol{L} = \boldsymbol{P}'\boldsymbol{D_L}\boldsymbol{P}$ and $\boldsymbol{R} = \boldsymbol{Q}'\boldsymbol{D_R}\boldsymbol{Q}$. Then, the second result follows from

$$\begin{aligned}
\|\boldsymbol{LCR}\|_F^2 &= \mathrm{tr}(\boldsymbol{RCL}^2\boldsymbol{CR}) = \mathrm{tr}(\boldsymbol{RCP}'\boldsymbol{D}_{\boldsymbol{L}}^2\boldsymbol{PCR}) \\
&= \mathrm{tr}(\boldsymbol{D}_{\boldsymbol{L}}^2\boldsymbol{PCR}^2\boldsymbol{CP}') \leq \|\boldsymbol{L}\|_2^2 \mathrm{tr}(\boldsymbol{PCR}^2\boldsymbol{CP}')
\end{aligned}$$

$$= \|\boldsymbol{L}\|_2^2 \mathrm{tr}(\boldsymbol{R}^2\boldsymbol{C}^2) = \|\boldsymbol{L}\|_2^2 \mathrm{tr}(\boldsymbol{Q}'\boldsymbol{D}_{\boldsymbol{R}}^2\boldsymbol{Q}\boldsymbol{C}^2)$$
$$= \|\boldsymbol{L}\|_2^2 \mathrm{tr}(\boldsymbol{D}_{\boldsymbol{R}}^2\boldsymbol{Q}\boldsymbol{C}^2\boldsymbol{Q}') \le \|\boldsymbol{L}\|_2^2\|\boldsymbol{R}\|_2^2 \mathrm{tr}(\boldsymbol{Q}\boldsymbol{C}^2\boldsymbol{Q}')$$
$$= \|\boldsymbol{L}\|_2^2\|\boldsymbol{R}\|_2^2 \mathrm{tr}(\boldsymbol{C}^2) = \|\boldsymbol{L}\|_2^2\|\boldsymbol{R}\|_2^2\|\boldsymbol{C}\|_F^2.$$

Proof of Lemma 2. Write $\boldsymbol{\Lambda}$ as its spectral decomposition $\boldsymbol{\Lambda} = \boldsymbol{U}'\boldsymbol{D}_{\boldsymbol{\Lambda}}\boldsymbol{U}$, where $\boldsymbol{D}_{\boldsymbol{\Lambda}}$ is a diagonal matrix whose jth diagonal element is $\lambda_{p,j} = \lambda_{p,j}(\boldsymbol{\Lambda})$. Then,

$$T_p = \sum_{j=1}^{p} \lambda_{p,j}\tilde{z}_{p,j}^2 - 2a_{p,j}\tilde{z}_{p,j}, \tag{6.1}$$

where $\tilde{z}_{p,j}$ is the jth component of $\boldsymbol{U}\boldsymbol{z}$ and $a_{p,j}$ is the jth component of $\boldsymbol{U}\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}$. Let $\zeta_{p,j} = \lambda_{p,j}\tilde{z}_{p,j}^2 - 2a_{p,j}\tilde{z}_{p,j} - \lambda_{p,j}$. Then $\mathrm{E}\zeta_{p,j} = 0$, $\sigma_{p,j}^2 = \mathrm{E}\zeta_{p,j}^2 = 2\lambda_{p,j}^2 + 4a_{p,j}^2$, and $T_p - \mathrm{E}(T_p) = \sum_{j=1}^{p}\zeta_{p,j}$. In the following, we show that $\{\zeta_{p,j},\ j = 1, 2, ..., p\}$ satisfy condition (B_γ) on page 43 of Saulis and Statulevicius [29] for $k \ge 3$. Actually, there exist constants $C_1 > 0$ and $M_0 > 0$ such that

$$\begin{aligned}
|\mathrm{E}\zeta_{p,j}^k| &= |\mathrm{E}(\lambda_{p,j}\tilde{z}_{p,j}^2 - 2a_{p,j}\tilde{z}_{p,j} - \lambda_{p,j})^k| \\
&\le \mathrm{E}|\lambda_{p,j}\tilde{z}_{p,j}^2 - 2a_{p,j}\tilde{z}_{p,j} - \lambda_{p,j}|^k \\
&\le \mathrm{E}\left(|\lambda_{p,j}\tilde{z}_{p,j}^2| + |2a_{p,j}\tilde{z}_{p,j}| + |\lambda_{p,j}|\right)^k \\
&\le 3^{k-1}\mathrm{E}\left(|\lambda_{p,j}|^k|\tilde{z}_{p,j}|^{2k} + |2a_{p,j}|^k|\tilde{z}_{p,j}|^k + |\lambda_{p,j}|^k\right) \\
&\le 3^{k-1}\left(|\lambda_{p,j}|^k(2k-1)!! + |2a_{p,j}|^k(k-1)!! + |\lambda_{p,j}|^k\right) \\
&\le 3^{k-1}C_1 k! 2^k(|\lambda_{p,j}|^k + |a_{p,j}|^k) \\
&= (1/3)C_1 6^k k!(|\lambda_{p,j}|^k + |a_{p,j}|^k)
\end{aligned}$$

$$
\begin{aligned}
&= (1/3)C_1 6^k k! \sigma_{p,j}^2 \frac{|\lambda_{p,j}|^k + |a_{p,j}|^k}{2\lambda_{p,j}^2 + 4a_{p,j}^2} \\
&\leq (1/3)C_1 6^k k! \sigma_{p,j}^2 \cdot 1/2[\max\{|\lambda_{p,j}|, |a_{p,j}|\}]^{k-2} \\
&\leq C_1 6^{k-1} M_0^{k-2} k! \sigma_{p,j}^2 \\
&\leq (6M_0 \cdot \max\{6C_1, 1\})^{k-2} k! \sigma_{p,j}^2.
\end{aligned}
$$

Therefore, $\{\zeta_{p,j}, j = 1, 2, ..., p\}$ satisfy the condition (B_γ) with constants $\gamma = 0$ and $K = 6M_0 \max\{6C_1, 1\}$.

Then, Theorem 3.1 of Saulis and Statulevicius [29] implies that

$$
\sup_x \left| F_{Z_p}(x) - \Phi(x) \right| \leq \frac{324\sqrt{2}K_p}{B_p},
$$

where $K_p = 2\max\{K, \sqrt{6}M_0\}$, $B_p^2 = \sum_{j=1}^{p} 2\lambda_{p,j}^2 + 4a_{p,j}^2 = 2\|\mathbf{\Lambda}\|_F^2 + 4\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}$, and F_{Z_p} denotes the distribution function of $Z_p = [T_p - \mathrm{E}(T_p)]/\sqrt{\mathrm{Var}(T_p)}$. In other words,

$$
\sup_x \left| F_{Z_p}(x) - \Phi(x) \right| \lesssim \frac{1}{[2\|\mathbf{\Lambda}\|_F^2 + 4\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}]^{1/2}} \lesssim \frac{1}{D_p},
$$

which converges to 0 as $p \to \infty$, because under (C2), $\|\mathbf{\Lambda}\|_F \asymp \|\mathbf{\Delta}\|_F$, $\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta} \asymp \boldsymbol{\delta}'\boldsymbol{\delta}$.

Proof of Lemma 3. Under (C2), $\hat{\boldsymbol{\Sigma}}_k$ is asymptoticly invertible by (12.3). Hence,

$$
\hat{\boldsymbol{\Sigma}}_k^{-1} = \boldsymbol{\Sigma}_k^{-1} + \hat{\boldsymbol{\Sigma}}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}, \quad \hat{\boldsymbol{\Sigma}}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} = \hat{\boldsymbol{\Sigma}}_k^{-1}(\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k)\boldsymbol{\Sigma}_k^{-1}.
$$

Then,

$$\|\hat{\boldsymbol{\Sigma}}_k^{-1}\|_1 \leq \|\boldsymbol{\Sigma}_k^{-1}\|_1 + \|\hat{\boldsymbol{\Sigma}}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\|_1 \leq (1 + \|\hat{\boldsymbol{\Sigma}}_k^{-1}\|_1 \|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k\|_1)\boldsymbol{\Sigma}_k^{-1}.$$

Since $\|\hat{\boldsymbol{\Sigma}}_k - \boldsymbol{\Sigma}_k\|_1 = O_P(a_n)$, $\|\boldsymbol{\Sigma}_k^{-1}\|_1 = O_P(v_p)$, and $a_n v_p \to 0$, it holds that

$$1/2\|\hat{\boldsymbol{\Sigma}}_k^{-1}\|_1 \leq (1 + \|\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k\|_1 \|\boldsymbol{\Sigma}_k^{-1}\|_1)^{-1} \|\hat{\boldsymbol{\Sigma}}_k^{-1}\|_1 \leq \|\boldsymbol{\Sigma}_k^{-1}\|_1.$$

Hence, $\|\hat{\boldsymbol{\Sigma}}_k^{-1}\|_1 \leq 2\|\boldsymbol{\Sigma}_k^{-1}\|_1$. Then,

$$\|\hat{\boldsymbol{\Sigma}}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\|_1 = \|\hat{\boldsymbol{\Sigma}}_k^{-1}(\boldsymbol{\Sigma}_k - \hat{\boldsymbol{\Sigma}}_k)\boldsymbol{\Sigma}_k^{-1}\|_1 = O_P\left(a_n v_p^2\right).$$

Proof of Lemma 4. Note that

$$\begin{aligned}|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1) - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)| &\leq |(\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{x} - \boldsymbol{\mu}_1)| \\ &\quad + |\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)|.\end{aligned}$$

By Chebyshev's inequality,

$$\begin{aligned}|(\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{x} - \boldsymbol{\mu}_1)| &= O_P\left(\sqrt{\mathrm{Var}\left((\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{x} - \boldsymbol{\mu}_1) \mid \boldsymbol{X}\right)}\right) \\ &= O_P\left(\sqrt{(\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta})'\boldsymbol{\Sigma}_1(\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta})}\right) \\ &= O_P\left(\|\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}\|\right) \\ &= O_P\left(\|(\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\Sigma}_2^{-1})\hat{\boldsymbol{\delta}}\| + \|\boldsymbol{\Sigma}_2^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|\right)\end{aligned}$$

$$= O_P\left(\|\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\Sigma}_2^{-1}\|_2\|\hat{\boldsymbol{\delta}}\| + \|\boldsymbol{\Sigma}_2^{-1}\|_2\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|\right)$$
$$= O_P(a_n) + O_P(\sqrt{b_n}),$$

where the third equality follows from (C2) and the last equality follows from (C2), $\|\hat{\boldsymbol{\delta}}\|^2 = O_P(1)$, and results (4.5) and (12.3). From result (12.3) and (C2),

$$|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)| \leq [\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)]^{1/2} = O_P(\sqrt{b_n}).$$

Proof of Lemma 5. The result follows from

$$|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}| \leq |\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}| + |\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}|,$$

$$|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}| \leq \|\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\Sigma}_2^{-1}\|_2\|\hat{\boldsymbol{\delta}}\|^2 = O_P\left(a_n\right)$$

and

$$|\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}| \leq |\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})| + |\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})|$$
$$= O_P\left(\|\boldsymbol{\delta}\|\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|\right) + O_P\left(\|\hat{\boldsymbol{\delta}}\|\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|\right)$$
$$= O_P(\sqrt{b_n}).$$

Proof of Lemma 6. Consider

$$|(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_1)'\boldsymbol{\nabla}(\boldsymbol{x} - \boldsymbol{\mu}_1)| \leq I + II + III,$$

where $I = |(\boldsymbol{x} - \boldsymbol{\mu}_1)'(\hat{\boldsymbol{\nabla}} - \boldsymbol{\nabla})(\boldsymbol{x} - \boldsymbol{\mu}_1)|$, $II = 2|(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x} - \boldsymbol{\mu}_1)|$, and $III =$

$|(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)|$. Let y_j be the jth component of $\boldsymbol{x} - \boldsymbol{\mu}_1$ and d_{ij} be the (i,j)th element of $\hat{\boldsymbol{\nabla}} - \boldsymbol{\nabla}$. Then

$$I \leq \sum_{i=1}^{p}\sum_{j=1}^{p} |d_{ij}y_iy_j| = O_P\left(\sum_{i=1}^{p}\sum_{j=1}^{p}|d_{ij}|\right) = O_P\left(\|\hat{\boldsymbol{\nabla}} - \boldsymbol{\nabla}\|_G\right),$$

where the first equality follows from the fact that $\mathrm{E}\left(|y_iy_j|\big|\boldsymbol{X}\right) \leq \sqrt{\mathrm{E}(y_i^2\big|\boldsymbol{X})\mathrm{E}(y_j^2\big|\boldsymbol{X})} = \sqrt{\sigma_{1ii}}\sqrt{\sigma_{1jj}} \leq M$. It then follows from Theorem 2(ii) that $I = O_P(\tau_n)$. Next,

$$\begin{aligned} II &= \left|(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x} - \boldsymbol{\mu}_1)\right| \\ &= O_P\left(\sqrt{\mathrm{Var}\left((\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x} - \boldsymbol{\mu}_1) \mid \boldsymbol{X}\right)}\right) \\ &= O_P\left(\sqrt{(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}\boldsymbol{\Sigma}_1\hat{\boldsymbol{\nabla}}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)}\right) \\ &= O_P(\sqrt{b_n}), \end{aligned}$$

since the eigenvalues of $\boldsymbol{\Sigma}_1$ and $\hat{\boldsymbol{\nabla}}$ are both bounded. Similarly,

$$III = \left|(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)\right| = O_P(b_n).$$

Proof of Lemma 7. From the property of the trace operation, we obtain that

$$\begin{aligned} \left|\mathrm{tr}(\hat{\boldsymbol{\Lambda}}) - \mathrm{tr}(\boldsymbol{\Lambda})\right| &= \left|\mathrm{tr}(\hat{\boldsymbol{\Delta}}\hat{\boldsymbol{\Sigma}}_2^{-1}) - \mathrm{tr}(\boldsymbol{\Delta}\boldsymbol{\Sigma}_2^{-1})\right| \\ &\leq \left|\mathrm{tr}((\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})\boldsymbol{\Sigma}_2^{-1})\right| + \left|\mathrm{tr}(\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\Sigma}_2^{-1}))\right| \\ &\leq \|(\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta})\boldsymbol{\Sigma}_2^{-1}\|_G + \|\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\Sigma}_2^{-1})\|_G \end{aligned}$$

$$
\begin{aligned}
&\leq \|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_G \|\boldsymbol{\Sigma}_2^{-1}\|_1 + \|\hat{\boldsymbol{\Delta}}\|_G \|\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\Sigma}_2^{-1}\|_1 \\
&= O_P\left(a_{1n} v_p\right) + O_P\left(c_{1p} a_n v_p^2\right) \\
&= O_P\left(\tau_n\right),
\end{aligned}
$$

where the third inequality follows from Lemma 1 and the second equality follows from Lemma 3 and Theorem 2. This proves the first result.

Note that $\log(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|) = \log|\boldsymbol{I} + \boldsymbol{\Lambda}|$ and $\log(|\hat{\boldsymbol{\Sigma}}_1|/|\hat{\boldsymbol{\Sigma}}_2|) = \log|\boldsymbol{I} + \hat{\boldsymbol{\Lambda}}|$. We employ a Taylor expansion of $f(t) = \log|\boldsymbol{I} + t\boldsymbol{\Lambda}|$ as appeared in equation (9) of Rothman, Bickel, Levina, and Zhu [28],

$$
\log|\boldsymbol{I} + \boldsymbol{\Lambda}| = \operatorname{tr}(\boldsymbol{\Lambda}) - \boldsymbol{l}'\boldsymbol{K}\boldsymbol{l}, \quad \boldsymbol{K} = \int_0^1 (1-v)(\boldsymbol{I} + v\boldsymbol{\Lambda})^{-1} \otimes (\boldsymbol{I} + v\boldsymbol{\Lambda})^{-1} dv,
$$

where $\boldsymbol{l}$ is $\boldsymbol{\Lambda}$ vectorized to be a $p^2 \times 1$ vector and $\otimes$ denotes the Kronecker product of matrices. Let $\hat{\boldsymbol{K}}$ be $\boldsymbol{K}$ with $\boldsymbol{\Lambda}$ and $\boldsymbol{l}$ replaced by $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{l}}$. Then,

$$
\left|\log|\boldsymbol{I} + \hat{\boldsymbol{\Lambda}}| - \operatorname{tr}(\hat{\boldsymbol{\Lambda}}) - \log|\boldsymbol{I} + \boldsymbol{\Lambda}| + \operatorname{tr}(\boldsymbol{\Lambda})\right| \leq \left|\hat{\boldsymbol{l}}'\hat{\boldsymbol{K}}\hat{\boldsymbol{l}} - \boldsymbol{l}'\boldsymbol{K}\boldsymbol{l}\right| \leq I + II + III,
$$

where $I = |\boldsymbol{l}'(\hat{\boldsymbol{K}} - \boldsymbol{K})\boldsymbol{l}|$, $II = |(\hat{\boldsymbol{l}} - \boldsymbol{l})'\hat{\boldsymbol{K}}\hat{\boldsymbol{l}}|$, and $III = |\boldsymbol{l}'\hat{\boldsymbol{K}}(\hat{\boldsymbol{l}} - \boldsymbol{l})|$. Note that

$$
\begin{aligned}
\|\hat{\boldsymbol{K}} - \boldsymbol{K}\|_2 &\leq \int_0^1 (1-v) \left\| [(\boldsymbol{I} + v\hat{\boldsymbol{\Lambda}})^{-1} - (\boldsymbol{I} + v\boldsymbol{\Lambda})^{-1}] \otimes (\boldsymbol{I} + v\hat{\boldsymbol{\Lambda}})^{-1} \right\|_2 dv \\
&\quad + \int_0^1 (1-v) \left\| (\boldsymbol{I} + v\boldsymbol{\Lambda})^{-1} \otimes [(\boldsymbol{I} + v\hat{\boldsymbol{\Lambda}})^{-1} - (\boldsymbol{I} + v\boldsymbol{\Lambda})^{-1}] \right\|_2 dv \\
&\lesssim \int_0^1 (1-v) \|(\boldsymbol{I} + v\hat{\boldsymbol{\Lambda}})^{-1} - (\boldsymbol{I} + v\boldsymbol{\Lambda})^{-1}\|_2 dv
\end{aligned}
$$

$$\lesssim \int_0^1 (1-v)v\|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\|_2 dv$$
$$= \|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\|_2/6,$$

where the first $\lesssim$ follows from the fact that eigenvalues of a Kronecker product of symmetric matrices are products of the eigenvalues of the symmetric matrices, and the second $\lesssim$ follows from the fact that the spectra of $(\boldsymbol{I} + v\boldsymbol{\Lambda})^{-1}$ and $(\boldsymbol{I} + v\hat{\boldsymbol{\Lambda}})^{-1}$ are both positive and bounded. Thus,

$$I = |\boldsymbol{l}'(\hat{\boldsymbol{K}} - \boldsymbol{K})\boldsymbol{l}| \leq \|\hat{\boldsymbol{K}} - \boldsymbol{K}\|_2 \boldsymbol{l}'\boldsymbol{l} \lesssim \|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\|_2 \boldsymbol{l}'\boldsymbol{l} = \|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\|_2 \|\boldsymbol{\Lambda}\|_F^2.$$

Since $\hat{\boldsymbol{K}}$ is positive definite with bounded spectrum,

$$\begin{aligned} II &= |(\hat{\boldsymbol{l}} - \boldsymbol{l})'\hat{\boldsymbol{K}}\hat{\boldsymbol{l}}| \\ &\leq |(\hat{\boldsymbol{l}} - \boldsymbol{l})'\hat{\boldsymbol{K}}(\hat{\boldsymbol{l}} - \boldsymbol{l})|^{1/2} \, |\hat{\boldsymbol{l}}'\hat{\boldsymbol{K}}\hat{\boldsymbol{l}}|^{1/2} \\ &\lesssim |(\hat{\boldsymbol{l}} - \boldsymbol{l})'(\hat{\boldsymbol{l}} - \boldsymbol{l})|^{1/2} \, |\hat{\boldsymbol{l}}'\hat{\boldsymbol{l}}|^{1/2} \\ &= \|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\|_F \|\hat{\boldsymbol{\Lambda}}\|_F. \end{aligned}$$

Similarly,

$$III = |\boldsymbol{l}'\hat{\boldsymbol{K}}(\hat{\boldsymbol{l}} - \boldsymbol{l})| \lesssim \|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\|_F \|\boldsymbol{\Lambda}\|_F.$$

From

$$\begin{aligned} \hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda} = \hat{\boldsymbol{\Sigma}}_1^{1/2}(\hat{\boldsymbol{\nabla}} - \boldsymbol{\nabla})\hat{\boldsymbol{\Sigma}}_1^{1/2} + (\hat{\boldsymbol{\Sigma}}_1^{1/2} - \boldsymbol{\Sigma}_1^{1/2})\boldsymbol{\nabla}\hat{\boldsymbol{\Sigma}}_1^{1/2} \\ + \boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\nabla}(\hat{\boldsymbol{\Sigma}}_1^{1/2} - \boldsymbol{\Sigma}_1^{1/2}), \end{aligned} \tag{6.2}$$

$$\begin{aligned}\hat{\boldsymbol{\nabla}} - \boldsymbol{\nabla} &= \hat{\boldsymbol{\Sigma}}_1^{-1}(\boldsymbol{\Delta} - \hat{\boldsymbol{\Delta}})\hat{\boldsymbol{\Sigma}}_2^{-1} + (\boldsymbol{\Sigma}_1^{-1} - \hat{\boldsymbol{\Sigma}}_1^{-1})\boldsymbol{\Delta}\hat{\boldsymbol{\Sigma}}_2^{-1} \\ &\quad + \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Delta}(\boldsymbol{\Sigma}_2^{-1} - \hat{\boldsymbol{\Sigma}}_2^{-1}),\end{aligned} \tag{6.3}$$

and Lemma 1,

$$\begin{aligned}\|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\|_F &\lesssim \|\hat{\boldsymbol{\nabla}} - \boldsymbol{\nabla}\|_F + 2\|\hat{\boldsymbol{\Sigma}}_1^{1/2} - \boldsymbol{\Sigma}_1^{1/2}\|_2\|\boldsymbol{\nabla}\|_F \\ &\lesssim \|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F + \|\hat{\boldsymbol{\Sigma}}_1^{-1} - \boldsymbol{\Sigma}_1^{-1}\|_2\|\boldsymbol{\Delta}\|_F \\ &\quad + \|\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\Sigma}_2^{-1}\|_2\|\boldsymbol{\Delta}\|_F + 2\|\hat{\boldsymbol{\Sigma}}_1^{1/2} - \boldsymbol{\Sigma}_1^{1/2}\|_2\|\boldsymbol{\Delta}\|_F.\end{aligned}$$

From (C2) and (4.3), $\|\boldsymbol{\Delta}\|_F^2 = \sum \Delta_{ij}^2 \leq (2M)^{2-\eta} \sum |\Delta_{ij}|^\eta = (2M)^{2-\eta} c_{1p}$. Using this fact and a similar proof to that of Theorem 2(i), we can show that

$$\begin{aligned}\|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F &= O_P\left(\sqrt{c_{1p}}\,(n^{-1}\log p)^{1/2-\eta/4}\right) \\ &= O_P\left(\sqrt{c_{1p}}\,(n^{-1}\log p)^{(1-\eta)/2}\right).\end{aligned} \tag{6.4}$$

Also, $\|\hat{\boldsymbol{\Sigma}}_1^{1/2} - \boldsymbol{\Sigma}_1^{1/2}\|_2 \lesssim [\|\hat{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1\|_2]^{1/2} = O_p\left(\sqrt{a_n}\right)$. Hence,

$$\|\hat{\boldsymbol{\Lambda}} - \boldsymbol{\Lambda}\|_F = O_P\left(\sqrt{c_{1p}}\left(n^{-1}\log p\right)^{(1-\eta)/2} + \sqrt{c_{1p}a_n}\right).$$

Therefore,

$$II + III = O_P\left(c_{1p}\left(n^{-1}\log p\right)^{(1-\eta)/2} + c_{1p}\sqrt{a_n}\right) = O_P(\tau_n).$$

As for term I, by (12.18),

$$\begin{aligned}\|\hat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\|_2 &\lesssim \|\hat{\boldsymbol{\nabla}}-\boldsymbol{\nabla}\|_2+2\|\hat{\boldsymbol{\Sigma}}_1^{1/2}-\boldsymbol{\Sigma}_1^{1/2}\|_2\\ &\lesssim \|\hat{\boldsymbol{\Sigma}}_1^{-1}-\boldsymbol{\Sigma}_1^{-1}\|_2+\|\hat{\boldsymbol{\Sigma}}_2^{-1}-\boldsymbol{\Sigma}_2^{-1}\|_2+2\|\hat{\boldsymbol{\Sigma}}_1^{1/2}-\boldsymbol{\Sigma}_1^{1/2}\|_2\\ &= O_P\left(\sqrt{a_n}\right),\end{aligned}$$

Hence, $I = O_P(\tau_n)$. This together with the proved first result imply the second result.

6.2 Proofs of Theorems

Proof of Theorem 1. In this proof, $\hat{\boldsymbol{\Sigma}}_1$, $\hat{\boldsymbol{\Sigma}}_2$, $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ denote the MLEs of the corresponding parameters without thresholding.

(i) Note that

$$|(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)-(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\nabla}(\boldsymbol{x}-\boldsymbol{\mu})| \le I+II+III,$$

where $I = |(\boldsymbol{x}-\boldsymbol{\mu}_1)'(\hat{\boldsymbol{\nabla}}-\boldsymbol{\nabla})(\boldsymbol{x}-\boldsymbol{\mu}_1)|$, $II = 2|(\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x}-\boldsymbol{\mu}_1)|$, and $III = |(\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_1)|$. When $p < n$, both $\|\hat{\boldsymbol{\Sigma}}_k-\boldsymbol{\Sigma}_k\|_2$ and $\|\hat{\boldsymbol{\Sigma}}_k^{-1}-\boldsymbol{\Sigma}_k^{-1}\|_2$ are $O_P(\sqrt{p/n})$. Hence,

$$\begin{aligned}I &= O_P\left(E[|(\boldsymbol{x}-\boldsymbol{\mu}_1)'(\hat{\boldsymbol{\nabla}}-\boldsymbol{\nabla})(\boldsymbol{x}-\boldsymbol{\mu}_1)| \mid \boldsymbol{X}]\right)\\ &= O_P\left(\sum_{j=1}^{p}|\lambda_{p,j}(\boldsymbol{\Sigma}_1^{1/2}(\hat{\boldsymbol{\nabla}}-\boldsymbol{\nabla})\boldsymbol{\Sigma}_1^{1/2})|\right)\\ &= O_P(p\|\boldsymbol{\Sigma}_1^{1/2}(\hat{\boldsymbol{\nabla}}-\boldsymbol{\nabla})\boldsymbol{\Sigma}_1^{1/2}\|_2)\end{aligned}$$

$$= O_P(p\|\hat{\boldsymbol{\nabla}} - \boldsymbol{\nabla}\|_2)$$
$$= O_P(p\sqrt{p/n}).$$

Also,

$$II = O_P\left(\sqrt{(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)}\right) = O_P\left(\sqrt{p/n}\right)$$

and

$$III = O_P\left((\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)\right) = O_P(p/n).$$

Hence,

$$|(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1) - (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\nabla}(\boldsymbol{x} - \boldsymbol{\mu})| = O_P(p\sqrt{p/n}). \tag{6.5}$$

By Chebyshev's inequality,

$$\begin{aligned}
|(\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{x} - \boldsymbol{\mu}_1)| &= O_P\left(\sqrt{\mathrm{Var}\left((\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{x} - \boldsymbol{\mu}_1) \mid \boldsymbol{X}\right)}\right) \\
&= O_P\left(\sqrt{(\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta})'\boldsymbol{\Sigma}_1(\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta})}\right) \\
&= O_P\left(\|\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}\|\right) \\
&= O_P\left(\|(\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\Sigma}_2^{-1})\hat{\boldsymbol{\delta}}\| + \|\boldsymbol{\Sigma}_2^{-1}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|\right) \\
&= O_P\left(\|\hat{\boldsymbol{\Sigma}}_2^{-1} - \boldsymbol{\Sigma}_2^{-1}\|_2\|\hat{\boldsymbol{\delta}}\| + \|\boldsymbol{\Sigma}_2^{-1}\|_2\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|\right) \\
&= O_P\left(p/\sqrt{n}\right) + O_P\left(\sqrt{p/n}\right).
\end{aligned}$$

Also,

$$|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)| \leq [\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)]^{1/2} = O_P(\sqrt{p/n}). \tag{6.6}$$

Hence, $|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1)-\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)|$ is bounded by

$$|(\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}-\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1})(\boldsymbol{x}-\boldsymbol{\mu}_1)|+|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{\mu}_1-\hat{\boldsymbol{\mu}}_1)|=O_P(p/\sqrt{n}).$$

From

$$|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}}-\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}|\le\|\hat{\boldsymbol{\Sigma}}_2^{-1}-\boldsymbol{\Sigma}_2^{-1}\|_2\|\hat{\boldsymbol{\delta}}\|^2=O_P\left(p\sqrt{p/n}\right)$$

and

$$\begin{aligned}|\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}|&\le|\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})|+|\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})|\\&=O_P\left(\|\boldsymbol{\delta}\|\|\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}\|\right)+O_P\left(\|\hat{\boldsymbol{\delta}}\|\|\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}\|\right)\\&=O_P(p/\sqrt{n}),\end{aligned}$$

we obtain that

$$|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}|\le|\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}}-\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}|+|\hat{\boldsymbol{\delta}}'\boldsymbol{\Sigma}_2^{-1}\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}|=O_P(p\sqrt{p/n}).\quad(6.7)$$

Next,

$$\begin{aligned}|\mathrm{tr}(\hat{\boldsymbol{\Lambda}})-\mathrm{tr}(\boldsymbol{\Lambda})|&=|\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{\Sigma}}_2^{-1}-\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1})|\\&\le|\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_2^{-1/2}(\hat{\boldsymbol{\Sigma}}_1-\boldsymbol{\Sigma}_1)\hat{\boldsymbol{\Sigma}}_2^{-1/2})|+|\mathrm{tr}(\boldsymbol{\Sigma}_1^{1/2}(\hat{\boldsymbol{\Sigma}}_2^{-1}-\boldsymbol{\Sigma}_2^{-1})\boldsymbol{\Sigma}_1^{1/2})|\\&\le p\|\hat{\boldsymbol{\Sigma}}_2^{-1/2}(\hat{\boldsymbol{\Sigma}}_1-\boldsymbol{\Sigma}_1)\hat{\boldsymbol{\Sigma}}_1^{-1/2}\|_2+p\|\boldsymbol{\Sigma}_1^{1/2}(\hat{\boldsymbol{\Sigma}}_2^{-1}-\boldsymbol{\Sigma}_2^{-1})\boldsymbol{\Sigma}_1^{1/2}\|_2\\&=O_P(p\sqrt{p/n}).\end{aligned}$$

Similar to the proof of Lemma 7, we have

$$\left|\log|\boldsymbol{I}+\hat{\boldsymbol{\Lambda}}|-\mathrm{tr}(\hat{\boldsymbol{\Lambda}})-\log|\boldsymbol{I}+\boldsymbol{\Lambda}|+\mathrm{tr}(\boldsymbol{\Lambda})\right| \lesssim A,$$

where $A=\|\hat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\|_2\|\boldsymbol{\Lambda}\|_F^2+\|\hat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\|_F\|\boldsymbol{\Lambda}\|_F+\|\hat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\|_F\|\hat{\boldsymbol{\Lambda}}\|_F$. By (12.18),

$$\|\hat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\|_2 \lesssim \|\hat{\boldsymbol{\nabla}}-\boldsymbol{\nabla}\|_2+2\|\hat{\boldsymbol{\Sigma}}_1^{1/2}-\boldsymbol{\Sigma}_1^{1/2}\|_2 = O_P((p/n)^{1/4}).$$

By (12.18)-(12.19),

$$\begin{aligned}\|\hat{\boldsymbol{\Lambda}}-\boldsymbol{\Lambda}\|_F &\lesssim \|\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}\|_F+\|\hat{\boldsymbol{\Sigma}}_1^{-1}-\boldsymbol{\Sigma}_1^{-1}\|_2\|\boldsymbol{\Delta}\|_F \\ &\quad+\|\hat{\boldsymbol{\Sigma}}_2^{-1}-\boldsymbol{\Sigma}_2^{-1}\|_2\|\boldsymbol{\Delta}\|_F+2\|\hat{\boldsymbol{\Sigma}}_1^{1/2}-\boldsymbol{\Sigma}_1^{1/2}\|_2\|\boldsymbol{\Delta}\|_F \\ &= O_P(\sqrt{p}(p/n)^{1/4}).\end{aligned}$$

Then, $A=O_P(p(p/n)^{1/4})$, since $\|\boldsymbol{\Lambda}\|_F=O(\sqrt{p})$. This together with (6.5), (6.6) and (6.7) prove that the difference between the quantity on the left hand side of (3.1) and the quantity on the left hand side of (3.3) is $O_P\left(p(p/n)^{1/4}\right)$, which converges to 0. The rest of the proof is similar to the proof of Theorem 3(i).

(ii) From the proof of part (i), we have $\|\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}\|^2=O_P\left(p/n\right)$ and $\|\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}\|_F^2=O_P\left(p^2/n\right)$. Hence,

$$\|\hat{D}_p-D_p\|^2=O_P\left(p^2/n\right).$$

The rest of the proof is similar to the proof of Theorem 3(ii).

Proof of Theorem 2. (i) Let $\tilde{\boldsymbol{\Delta}}=\tilde{\boldsymbol{\Sigma}}_2-\tilde{\boldsymbol{\Sigma}}_1$ and Δ_{ij}^S be the (i,j)th element of $\boldsymbol{S}_2-\boldsymbol{S}_1$. We adopt the technique in the proof of Theorem 1 in Bickel and Levina [5].

Note that,

$$\begin{aligned}\|\tilde{\mathbf{\Delta}} - \mathbf{\Delta}\|_G &= \sum_{i,j} |\Delta_{ij}^S I(|\Delta_{ij}^S| \geq t_{1n}) - \Delta_{ij}| \\ &\leq \sum_{i,j} \left|\Delta_{ij}^S I(|\Delta_{ij}^S| \geq t_{1n}) - \Delta_{ij} I(|\Delta_{ij}| \geq t_{1n})\right| \qquad (6.8) \\ &\quad + \sum_{i,j} |\Delta_{ij}|\, I(|\Delta_{ij}| < t_{1n}). \qquad (6.9)\end{aligned}$$

By (4.3), the sum in (6.9) is bounded by

$$t_{1n}^{1-\eta} \sum_{i,j} |\Delta_{ij}|^\eta = t_{1n}^{1-\eta} c_{1p}.$$

The sum in (12.17) is bounded by

$$\begin{aligned}&\sum_{i,j} |\Delta_{ij}^S - \Delta_{ij}| I(|\Delta_{ij}^S| \geq t_{1n}, |\Delta_{ij}| \geq t_{1n}) \qquad (6.10) \\ &+ \sum_{i,j} |\Delta_{ij}| I(|\Delta_{ij}^S| < t_{1n}, |\Delta_{ij}| \geq t_{1n}) \qquad (6.11) \\ &+ \sum_{i,j} |\Delta_{ij}^S| I(|\Delta_{ij}^S| \geq t_{1n}, |\Delta_{ij}| < t_{1n}). \qquad (6.12)\end{aligned}$$

The quantity in (6.10) is bounded by

$$\begin{aligned}\sum_{i,j} |\Delta_{ij}^S - \Delta_{ij}| I(|\Delta_{ij}| \geq t_{1n}) &\leq \max_{i,j} |\Delta_{ij}^S - \Delta_{ij}| \sum_{i,j} \frac{|\Delta_{ij}|}{t_{1n}} \\ &= O_P\left((n^{-1} \log p)^{1/2} c_{1p} t_{1n}^{-\eta}\right) \\ &= O_P\left(a_{1n}\right).\end{aligned}$$

The quantity in (6.11) is bounded by

$$\begin{aligned}
&\max_{i,j} |\Delta_{ij}^S - \Delta_{ij}| \sum_{i,j} I(|\Delta_{ij}| \geq t_{1n}) + t_{1n} \sum_{i,j} I(|\Delta_{ij}| \geq t_{1n}) \\
&\leq \max_{i,j} |\Delta_{ij}^S - \Delta_{ij}| \, t_{1n}^{-\eta} c_{1p} + t_{1n}^{1-\eta} c_{1p} \\
&= O_P\left(a_{1n}\right).
\end{aligned}$$

The quantity in (6.12) is bounded by

$$\begin{aligned}
&\sum_{i,j} |\Delta_{ij}^S - \Delta_{ij}| I(|\Delta_{ij}^S| \geq t_{1n}, |\Delta_{ij}| < t_{1n}) + \sum_{i,j} |\Delta_{ij}| I(|\Delta_{ij}| < t_{1n}) \\
&\leq I + II + t_{1n}^{1-\eta} c_{1p}
\end{aligned}$$

for some $\gamma \in (0, 1)$, where

$$\begin{aligned}
I &= \sum_{i,j} |\Delta_{ij}^S - \Delta_{ij}| I(|\Delta_{ij}^S| \geq t_{1n}, |\Delta_{ij}| \leq \gamma t_{1n}) \\
&\leq \max_{i,j} |\Delta_{ij}^S - \Delta_{ij}| \sum_{i,j} I(|\Delta_{ij}^S - \Delta_{ij}| > (1-\gamma) t_{1n})
\end{aligned}$$

and

$$\begin{aligned}
II &= \sum_{i,j} |\Delta_{ij}^S - \Delta_{ij}| I(|\Delta_{ij}^S| \geq t_{1n}, \gamma t_{1n} < |\Delta_{ij}| < t_{1n}) \\
&\leq \max_{i,j} |\Delta_{ij}^S - \Delta_{ij}| \sum_{i,j} I(|\Delta_{ij}^S| \geq t_{1n}, \gamma t_{1n} < |\Delta_{ij}| < t_{1n}) \\
&\leq \max_{i,j} |\Delta_{ij}^S - \Delta_{ij}| (\gamma t_{1n})^{-\eta} c_{1p}
\end{aligned}$$

$$= O_P\left(a_{1n}\right).$$

Note that

$$P\left(I > 0\right) = P\left(\max_{i,j} |\Delta_{ij}^S - \Delta_{ij}| > (1-\gamma)t_{1n}\right) \leq 2p^2 e^{-n\zeta(1-\gamma)^2 t_{1n}^2/4}$$

for some $\zeta > 0$. Since $t_{1n} = M_1\sqrt{\log p/n}$ and $0 < 1-\gamma < 1$, $2\log p - n\zeta(1-\gamma)^2 t_{1n}^2/4 \to -\infty$, if M_1 is sufficiently large. Hence, $I = 0$ with probability tending to 1. Combining these results, we conclude that

$$\|\tilde{\mathbf{\Delta}} - \mathbf{\Delta}\|_G = O_P\left(a_{1n}\right). \tag{6.13}$$

Consider

$$\begin{aligned}
\|\hat{\mathbf{\Delta}} - \tilde{\mathbf{\Delta}}\|_G &= \sum_{i,j} |\Delta_{ij}^S + s_{1ij}| I(|s_{1ij}| \geq t_{2n}, |s_{2ij}| < t_{2n}, |\Delta_{ij}^S| \geq t_{1n}) \\
&\quad + \sum_{i,j} |\Delta_{ij}^S - s_{2ij}| I(|s_{1ij}| < t_{2n}, |s_{2ij}| \geq t_{2n}, |\Delta_{ij}^S| \geq t_{1n}) \\
&\leq 2t_{2n} \sum_{i,j} I(|\Delta_{ij}^S| \geq t_{1n}) \\
&\leq 2t_{2n}(III + IV),
\end{aligned}$$

where

$$III = \sum_{i,j} I(|\Delta_{ij}^S - \Delta_{ij}| \geq (1-\gamma)t_{1n})$$

and

$$IV = \sum_{i,j} I(|\Delta_{ij}| \geq \gamma t_{1n}) \leq \sum_{i,j} \frac{|\Delta_{ij}|}{\gamma t_{1n}}$$

for some $\gamma \in (0,1)$. In analogous to the aforementioned analysis of term I, $III = 0$ with probability tending to 1. On the other hand, $IV \leq c_{1p}/(\gamma t_{1n})^{\eta}$. Hence, $2t_{2n}IV = O_P(a_{1n})$. This shows that $\|\tilde{\mathbf{\Delta}} - \hat{\mathbf{\Delta}}\|_G = O_P(a_{1n})$, which together with result (12.14) imply that $\|\hat{\mathbf{\Delta}} - \mathbf{\Delta}\|_G = O_P(a_{1n})$.

(ii) It follows from Lemma 1 and result (12.19) that

$$\begin{aligned}\|\tilde{\mathbf{\nabla}} - \mathbf{\nabla}\|_G \leq{}& \|\hat{\mathbf{\Sigma}}_1^{-1}\|_1 \|\mathbf{\Delta} - \hat{\mathbf{\Delta}}\|_G \|\hat{\mathbf{\Sigma}}_2^{-1}\|_1 + \|\hat{\mathbf{\Sigma}}_1^{-1} - \mathbf{\Sigma}_1^{-1}\|_1 \|\mathbf{\Delta}\|_G \|\hat{\mathbf{\Sigma}}_2^{-1}\|_1 \\ &+ \|\mathbf{\Sigma}_1^{-1}\|_1 \|\mathbf{\Delta}\|_G \|\hat{\mathbf{\Sigma}}_1^{-1} - \mathbf{\Sigma}_1^{-1}\|_1.\end{aligned}$$

By Lemma 3,

$$\|\hat{\mathbf{\Sigma}}_1^{-1} - \mathbf{\Sigma}_1^{-1}\|_1 = O_P\left(a_n v_p^2\right).$$

Then, it holds that

$$\begin{aligned}\|\hat{\mathbf{\nabla}} - \mathbf{\nabla}\|_G &= O_P\left(\|\hat{\mathbf{\Delta}} - \mathbf{\Delta}\|_G v_p^2 + a_n c_{1p} v_p^3\right) \\ &= O_P\left(c_{1p}(n^{-1}\log p)^{(1-\eta)/2} v_p^2 + a_n c_{1p} v_p^3\right) \\ &= O_P(\tau_n),\end{aligned}$$

where the second equality follows from part (i).

Proof of Theorem 3. (i) When D_p is bounded. Let T_p be defined as in Lemma 2

and

$$\hat{T}_p|\boldsymbol{X} = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1)'\hat{\boldsymbol{\nabla}}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1) - 2\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_1)$$
$$\hat{\varphi}_p|\boldsymbol{X} = \mathrm{tr}(\hat{\boldsymbol{\Lambda}}) + \hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}} - \log(|\hat{\boldsymbol{\Sigma}}_1|/|\hat{\boldsymbol{\Sigma}}_2|)$$
$$\varphi_p = \mathrm{tr}(\boldsymbol{\Lambda}) + \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta} - \log(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_2|)$$

Denote $F_p(\cdot)$ the cumulative distribution function (c.d.f) of $T_p - \mathrm{E}(T_p)$ and $\hat{F}_p(\cdot)$ the conditional c.d.f of $\hat{T}_p|\boldsymbol{X} - \mathrm{E}(\hat{T}_p|\boldsymbol{X})$. From Lemma 4-7, we have

$$[\hat{T}_p|\boldsymbol{X} - E(\hat{T}_p|\boldsymbol{X})] - [T_p - \mathrm{E}(T_p)] \xrightarrow{P} 0 \quad \text{and} \quad \hat{\varphi}_p|\boldsymbol{X} - \varphi_p \xrightarrow{P} 0.$$

It follows from (3.3) that

$$R_2 - R_{B2} = \hat{F}_p(-\hat{\varphi}_p) - F_p(-\varphi_p),$$

where R_{B2} and R_2 are defined in (3.2) and (3.4).

Next, we show that if $\max\{\sqrt{b_n}, a_n, \tau_n\} \to 0$,

$$\hat{F}_p(-\hat{\varphi}_p) - F_p(-\varphi_p) \xrightarrow{P} 0. \tag{6.14}$$

Similarly, we can show that $R_1 - R_{B1} \xrightarrow{P} 0$. Then, (i) of Theorem 4.2 is proved.

We prove (6.14) by a subsequence argument. For any subsequence $\{p_k\} \subset \{p\}$,

there is a further subsequence $\{p_{k_t}\} \subset \{p_k\}$ such that

$$[\hat{T}_{p_{k_t}}|\boldsymbol{X} - E(\hat{T}_{p_{k_t}}|\boldsymbol{X})] - [T_{p_{k_t}} - \mathrm{E}(T_{p_{k_t}})] \xrightarrow{a.s.} 0, \tag{6.15}$$

$$\hat{\varphi}_{p_{k_t}}|\boldsymbol{X} - \varphi_{p_{k_t}} \xrightarrow{a.s.} 0. \tag{6.16}$$

Then,

$$\begin{aligned}
&|\hat{F}_{p_{k_t}}(-\hat{\varphi}_{p_{k_t}}) - F_{p_{k_t}}(-\hat{\varphi}_{p_{k_t}}) + F_{p_{k_t}}(-\hat{\varphi}_{p_{k_t}}) - F_{p_{k_t}}(-\varphi_{p_{k_t}})| \\
&\leq \sup_x |\hat{F}_{p_{k_t}}(x) - F_{p_{k_t}}(x)| + \sup_x |F'_{p_{k_t}}(x)|(\varphi_{p_{k_t}} - \hat{\varphi}_{p_{k_t}}) \\
&\leq (1 + \sup_x |F'_{p_{k_t}}(x)|) d_L(\hat{F}_{p_{k_t}}, F_{p_{k_t}}) + \sup_x |F'_{p_{k_t}}(x)|(\varphi_{p_{k_t}} - \hat{\varphi}_{p_{k_t}}),
\end{aligned}$$

where the last inequality follows from a well-known inequality on page 43 of Petrov [26] and $d_L(\hat{F}_{p_{k_t}}, F_{p_{k_t}})$ is the Levy metric between $\hat{F}_{p_{k_t}}$ and $F_{p_{k_t}}$.

Under (C4), it holds that

$$\sup_x |F'_{p_{k_t}}(x)| \leq C, \tag{6.17}$$

where C does not depend on the index p_{k_t}. (6.15) implies that $d_L(\hat{F}_{p_{k_t}}, F_{p_{k_t}}) \to 0$. Then, this together with (6.16) and (6.17) proves that

$$|\hat{F}_{p_{k_t}}(-\hat{\varphi}_{p_{k_t}}) - F_{p_{k_t}}(-\varphi_{p_{k_t}})| \xrightarrow{a.s.} 0.$$

By the above subsequence argument, we prove (6.14).

In the following, we show that (6.17) holds in some meaningful cases.

Case 1: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Note that,

$$\begin{aligned} T_p &= \sum_{j=1}^{p} \lambda_{p,j} \tilde{z}_{p,j}^2 - 2a_{p,j}\tilde{z}_{p,j} \\ &= \sum_{\lambda_{p,j}\neq 0} \lambda_{p,j} \left(\tilde{z}_{p,j} - \frac{a_{p,j}}{\lambda_{p,j}} \right)^2 - \sum_{\lambda_{p,j}\neq 0} \frac{a_{p,j}^2}{\lambda_{p,j}} - \sum_{\lambda_{p,j}=0} a_{p,j}\tilde{z}_{p,j}, \end{aligned} \tag{6.18}$$

where $\tilde{z}_{p,j}, j = 1, \ldots, p$ are i.i.d from $N(0,1)$, $\lambda_{p,j}$ is the jth smallest eigenvalue of $\boldsymbol{\Lambda}$ and $a_{p,j}$ is the jth component of $\boldsymbol{U}\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta}$, where $\boldsymbol{U}'\boldsymbol{\Lambda}\boldsymbol{U} = \mathrm{diag}(\lambda_{p,1}, \ldots, \lambda_{p,p})$.

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, T_p reduces to a normal random variable with mean 0 and

$$\mathrm{Var}(T_p) = \boldsymbol{\delta}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\delta},$$

which is bounded below by a constant under (C3). Then, (6.17) holds.

Case 2: There are at least two eigenvalues of $\boldsymbol{\Lambda}$ in $(-\infty, m]$ or $[m, \infty)$. Without loss of generality, assume the two eigenvalues $\lambda_{p,1} \geq m$ and $\lambda_{p,2} \geq m$.

Let

$$\begin{aligned} Y_{p,1} &= \lambda_{p,1} \left(\tilde{z}_{p,1} - \frac{a_{p,1}}{\lambda_{p,1}} \right)^2 \\ Y_{p,2} &= \lambda_{p,2} \left(\tilde{z}_{p,2} - \frac{a_{p,2}}{\lambda_{p,2}} \right)^2 \end{aligned}$$

and $Y_p = Y_{p,1} + Y_{p,2}$. Denote $f_{p,1}$ the p.d.f of $Y_{p,1}$, $f_{p,2}$ the p.d.f of $Y_{p,2}$; $\tilde{f}_{p,1}$ the p.d.f of the noncentral χ^2-distribution with 1 degree of freedom and noncentrality parameter $a_{p,1}^2/\lambda_{p,1}^3$, $\tilde{f}_{p,2}$ the p.d.f of the noncentral χ^2-distribution with 1 degree of freedom and noncentrality parameter $a_{p,2}^2/\lambda_{p,2}^3$.

Then,

$$\frac{f_{p,1}(y)}{\tilde{f}_{p,1}(y)} = \frac{\frac{1}{2\lambda_{p,1}} e^{-(y/\lambda_{p,1}+a_{p,1}^2/\lambda_{p,1}^2)/2} (\lambda_{p,1} a_{p,1}^{-2} y)^{-1/4}}{\frac{1}{2} e^{-(y+a_{p,1}^2/\lambda_{p,1}^3)/2} (\lambda_{p,1}^3 a_{p,1}^{-2} y)^{-1/4}}$$
$$= \frac{1}{\lambda_{p,1}^{1/2}} e^{-y(1+\lambda_{p,1}^{-1})/2} e^{a_{p,1}^2 \lambda_{p,1}^{-3}(1-\lambda_{p,1})/2}.$$

Under (C1) and (C2), $a_{p,1} = O(1)$. This together with $\lambda_{p,1} \geq m$ shows that

$$\sup_{y>0} \frac{f_{p,1}(y)}{\tilde{f}_{p,1}(y)} \leq c_1,$$

where c_1 does not depend on p. Similarly, we have

$$\sup_{y>0} \frac{f_{p,2}(y)}{\tilde{f}_{p,2}(y)} \leq c_2.$$

Let $f_{p,0}$ denote the p.d.f of Y_p. By convolution formula,

$$f_{p,0}(y) = \int_0^y f_{p,1}(y-t) f_{p,2}(t) dt.$$

Hence,

$$\sup_{y>0} f_{p,0}(y) \leq \sup_{y>0} \int_0^y c_1 c_2 \tilde{f}_{p,1}(y-t) \tilde{f}_{p,2}(t) dt$$
$$= c_1 c_2 \sup_{y>0} \tilde{f}_0(y),$$

where $\tilde{f}_0(y)$ is the p.d.f of noncentral χ^2-distribution with 2 degrees of freedom and

noncentrality parameter $(a_{p,1}^2\lambda_{p,1}^{-3} + a_{p,2}^2\lambda_{p,2}^{-3})$. Therefore,

$$\tilde{f}_0(y) = \frac{1}{2}e^{-(y+a_{p,1}^2\lambda_{p,1}^{-3}+a_{p,2}^2\lambda_{p,2}^{-3})/2} I_0\left(\sqrt{y(a_{p,1}^2\lambda_{p,1}^{-3} + a_{p,2}^2\lambda_{p,2}^{-3})}\right),$$

where $I_0(\cdot)$ is the Bessel function of the first kind. By the property of $I_0(\cdot)$, $\sup_{y>0} I_0\left(\sqrt{y(a_{p,1}^2\lambda_{p,1}^{-3} + a_{p,2}^2\lambda_{p,2}^{-3})}\right) \leq 1$ and $a_{p,1}^2\lambda_{p,1}^{-3} + a_{p,2}^2\lambda_{p,2}^{-3} = O(1)$ by (C1) and (C2). Then, it holds that

$$\sup_{y>0} f_{p,0}(y) \leq c_1 c_2 \sup_{y>0} \tilde{f}_0(y) \leq C,$$

where C does not depend on p.

Then, it follows from (6.18) that T_p is a sum of p independent random variables. By the property of convolution, the density of T_p

$$\sup_x |F_p'(x)| \leq \sup_x |\tilde{f}_{p,0}(x)| \leq C.$$

Therefore, (6.17) holds.

(ii) When $D_p \to \infty$, by Lemma 2, the misclassification rate of Bayes rule $R_B \to 0$. Hence, it suffices to show $R_{\text{SQDA}}(\boldsymbol{X}) \xrightarrow{P} 0$. Let

$$\hat{Z}_p|\boldsymbol{X} = \frac{\hat{T}_p|\boldsymbol{X} - \mathrm{E}(\hat{T}_p|\boldsymbol{X})}{[\mathrm{Var}(\hat{T}_p|\boldsymbol{X})]^{1/2}} \quad \text{and} \quad \hat{\xi}_p = \frac{\hat{\varphi}_p|\boldsymbol{X}}{[\mathrm{Var}(\hat{T}_p|\boldsymbol{X})]^{1/2}},$$

where $\hat{T}_p|\boldsymbol{X}$ and $\hat{\varphi}_p|\boldsymbol{X}$ are as defined in (i). Note that

$$\mathrm{E}(\hat{T}_p|\boldsymbol{X}) = \mathrm{tr}(\hat{\boldsymbol{\Lambda}}) \quad \text{and} \quad \mathrm{Var}(\hat{T}_p|\boldsymbol{X}) = 2\mathrm{tr}(\hat{\boldsymbol{\Lambda}}^2) + 4\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}}.$$

For $\hat{D}_p = \sqrt{\|\hat{\boldsymbol{\delta}}\|^2 + \|\hat{\boldsymbol{\Delta}}\|_F^2}$,

$$\begin{aligned}\hat{D}_p &\geq D_p - \sqrt{\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2 + \|\hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta}\|_F^2} \\ &= D_p - \sqrt{O_P\left(\max\{b_n, a_{1n}\}\right)} \\ &= D_p \left(1 - \sqrt{O_P\left(\max\{b_n, a_{1n}\}/D_p^2\right)}\right),\end{aligned}$$

where the first identity follows by (4.5), (12.10) and (C5). Therefore, $\hat{D}_p \to \infty$ if $\max\{b_n, a_{1n}\}/D_p^2 \to 0$. Using the subsequence argument as in (i) and a proof analogous to Lemma 2,

$$\left| F_{\hat{Z}_p|\boldsymbol{X}}(-\hat{\xi}_p) - \Phi(-\hat{\xi}_p) \right| \xrightarrow{P} 0.$$

Since

$$\hat{\xi}_p = \frac{\hat{\varphi}_p|\boldsymbol{X}}{[\mathrm{Var}(\hat{T}_p)]^{1/2}} = \frac{\mathrm{tr}(\hat{\boldsymbol{\Lambda}}) - \log(|\hat{\boldsymbol{\Sigma}}_1|/|\hat{\boldsymbol{\Sigma}}_2|) + \hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}}}{[2\mathrm{tr}(\hat{\boldsymbol{\Lambda}})^2 + 4\hat{\boldsymbol{\delta}}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\Sigma}}_1\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\delta}}]^{1/2}} \asymp \hat{D}_p,$$

there exists a constant c_2 such that

$$\Phi(-\hat{\xi}_p) \leq \Phi(-c_2\hat{D}_p) \leq \Phi\left(-c_2 D_p \left(1 - \sqrt{O_P\left(\max\{b_n, a_{1n}\}/D_p^2\right)}\right)\right).$$

Hence, if $\max\{b_n, a_{1n}\}/D_p^2 \to 0$, $F_{\hat{Z}_p|\boldsymbol{X}}(-\hat{\xi}_p) = R_2(\boldsymbol{X}) \xrightarrow{P} 0$, where $R_2(\boldsymbol{X})$ is defined in (i). Similarly, we can prove that $R_1(\boldsymbol{X}) \to 0$, and the result follows.

Part III

Regularized LASSO in High Dimensional Linear Regression

7 MOTIVATION

In many statistical applications, one investigates the effect of a vector $\boldsymbol{x}$ of p explanatory variables on a response variable y based on n independently observed data $\{y_i, \boldsymbol{x}_i, i = 1, ..., n\}$ following a linear model

$$y_i = \mu + \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma_i\varepsilon_i, \quad i = 1, ..., n, \tag{7.1}$$

where y_i is the ith observed response, $\boldsymbol{x}_i$ is the p-dimensional observed explanatory variables associated with y_i, $\boldsymbol{x}_i$'s are independent and identically distributed (iid), $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ is a p-dimensional vector of unknown parameters called regression effects, μ is an unknown parameter (intercept), σ_i's are positive unknown parameters, ε_i's are iid unobserved random errors with mean 0 and variance 1, $\boldsymbol{x}_i$'s and ε_i's are independent, and $\boldsymbol{A'}$ denotes the usual transpose of a vector or matrix $\boldsymbol{A}$. The theory of linear models is well established for traditional applications where the dimension p is fixed and the sample size $n > p$. With modern technologies, however, in many biological, medical, social, and economical studies, p is comparable with or much larger than n. Note that variable j, the jth component of $\boldsymbol{x}$, has no effect on the response when $\beta_j = 0$. When the number of variables p is large but many variables have no effect on the response, which is often true in applications, variable selection, i.e., identifying zero components of $\boldsymbol{\beta}$, is usually applied prior to statistical inference. Without loss of generality we assume that $\boldsymbol{x}_i$'s have mean 0 and variance 1 and they are standardized so that $\sum_{i=1}^n \boldsymbol{x}_i = \boldsymbol{0}$ and the diagonal elements of $\boldsymbol{S} = \sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'/n$ are equal to 1, since this does not affect variable selection.

There is a rich literature on asymptotic theory for variable selection in the case where $n \to \infty$ and p is fixed or $p \to \infty$ at a rate much slower than n. For variable selection in the case of $p > n$ with $p = O(n^l)$ for some $l > 1$ or $O(e^{n^\nu})$ for some $\nu \in (0, 1)$ (ultra-high dimension), some excellent advances in asymptotic theory have been made recently. See, for example, Fan and Peng [19], Hunter and Li [23], Meinshausen and Bühlmann [24], Zhao and Yu [38], Zou [39], Wang, Li, and Tsai [33], Fan and Lv [17], Zhang and Huang [36], Meinshausen and Yu [25], Wang [32], Fan and Song [15], and a review paper by Fan and Lv [18].

Let $\mathcal{M}_{\boldsymbol{\beta}} = \{j : \beta_j \neq 0\}$ denote the index set for nonzero components of $\boldsymbol{\beta}$ and let $\widehat{\mathcal{M}}_{\boldsymbol{\beta}}$ denote the set of indices of nonzero components of $\boldsymbol{\beta}$ selected by a variable selection method using data. The selection method or $\widehat{\mathcal{M}}_{\boldsymbol{\beta}}$ is selection-consistent if

$$P\left(\widehat{\mathcal{M}}_{\boldsymbol{\beta}} = \mathcal{M}_{\boldsymbol{\beta}}\right) \to 1, \tag{7.2}$$

where the limit is taken as $n \to \infty$ with $p = p_n$ that may also diverge to ∞ and the probability is with respect to the randomness of data $\{y_i, \boldsymbol{x}_i, i = 1, ..., n\}$. Selection-consistency is an important property, since it leads to oracle properties of estimation and inference procedures (see, e.g., Fan and Lv [17]). Although some results on the selection-consistency can be found in the previously cited papers, they are established under some conditions that may not generally hold. For example, it is well known that the LASSO method (Tibshirani [31]) requires very strong conditions for its selection-consistency (see our discussions in Chapter 10). Another example is the

sure independent screening (SIS) in Fan and Lv [17] who showed that if

$$\min_{j\in\mathcal{M}_\beta}\left|\sum_{k\in\mathcal{M}_\beta}\beta_k\rho_{kj}\right|\geq c_0n^{-\kappa} \tag{7.3}$$

for some constant $c_0>0$ and $0\leq\kappa<1/2$, where ρ_{kj} is the correlation coefficient between the jth and kth component of $\boldsymbol{x}$, then, under some regularity conditions, the SIS is screening consistent in the sense that $P\left(\mathcal{M}_{\boldsymbol{\beta}}\subset\widehat{\mathcal{M}}_{\boldsymbol{\beta}}\right)\to 1$, i.e., useful components of $\boldsymbol{x}$ are selected by the SIS with probability tending to 1. However, condition (7.3) may be questionable sometimes, e.g., Model 4 in Chapter 11. The SIS is selection-consistent according to (7.2) if, in addition to (7.3),

$$\max_{j\notin\mathcal{M}_\beta}\left|\sum_{k\in\mathcal{M}_\beta}\beta_k\rho_{kj}\right|=o(n^{-\kappa}) \tag{7.4}$$

holds. However, condition (7.4) is rarely satisfied in practice, since it imposes a very strong structure on the correlation coefficients ρ_{kj}. That is why the SIS has a reputation to be screening consistent only, not selection consistent. The reason why the SIS needs conditions (7.3) and (7.4) is because it is based on thresholding components of an estimated marginal effect vector $\boldsymbol{\beta}_M=\mathrm{E}(\boldsymbol{x}y)=\mathrm{Cov}(\boldsymbol{x},y)$, which is different from the regression effect vector $\boldsymbol{\beta}=\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_M$ in general, where $\boldsymbol{\Sigma}=\mathrm{E}(\boldsymbol{x}\boldsymbol{x}')$ is the covariance matrix of $\boldsymbol{x}$ assumed to be positive definite (note that $\mathrm{E}(\boldsymbol{x})=0$ is assumed). Zero components of $\boldsymbol{\beta}$ may not be zero components of $\boldsymbol{\beta}_M$ and vice versa.

The purpose of this paper is to derive a variable selection method that is selection-consistent without requiring conditions (7.3)-(7.4) or those described in Chapter 10 needed for LASSO. These conditions are replaced by a more practical condition, a

sparsity condition on the covariance matrix $\mathbf{\Sigma} = \mathrm{E}(\boldsymbol{x}'\boldsymbol{x})$ (Bickel and Levina [3] or Cai, Zhang, and Zhou [8]), or a sparsity condition on the inverse of $\mathbf{\Sigma}$ (Cai, Liu, and Luo [7]). Our key idea is that, since the LASSO utilizes a least squares minimization involving the covariate sample covariance matrix that is not well behaved when p is larger than n, we replace the least squares component in the minimization of LASSO by a regularized least squares component using results in the high-dimensional covariance matrix estimation (Bickel and Levina [3], Cai, Zhang, and Zhou [8], Cai, Liu, and Luo [7]). Furthermore, a thresholding step is added to the resulting estimator to improve its variable selection performance.

8 THE METHODOLOGY

We first introduce a simple procedure that is selection-consistent. The idea is simple. If p is fixed, then we can select variables by thresholding the least squares estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_{\text{lse}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{S}^{-1}\boldsymbol{X}'\boldsymbol{y}/n,$$

where $\boldsymbol{y}$ is the n-dimensional vector of y_i's, $\boldsymbol{X}$ is the $n \times p$ matrix whose ith row is $\boldsymbol{x}_i$, and $\boldsymbol{S} = \boldsymbol{X}'\boldsymbol{X}/n$. But when $p > n$, $\boldsymbol{S}$ is singular and even if we use a generalized inverse as its inverse, $\hat{\boldsymbol{\beta}}_{\text{lse}}$ does not have a good behavior because $\boldsymbol{S}$ is not a good estimator of the covariate covariance matrix $\boldsymbol{\Sigma} = \text{E}(\boldsymbol{S})$. If $\boldsymbol{\Sigma}$ is sparse in some sense, then we may estimate $\boldsymbol{\Sigma}$ by a regularized or sparse estimator $\hat{\boldsymbol{\Sigma}}$ that is L_2-consistent in the sense that

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 = o_p(1),$$

where $\|\boldsymbol{A}\|_2$ is the L_2 norm of a matrix $\boldsymbol{A}$. Such an estimator can be obtained using results in high-dimensional covariance matrix estimation (Bickel and Levina [3]; Cai and Liu [9]). Then we estimate $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}}_{\text{slse}} = \hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}'\boldsymbol{y}/n \tag{8.1}$$

and construct a variable selection method by thresholding $\hat{\boldsymbol{\beta}}_{\text{slse}}$, i.e., the selected components of $\boldsymbol{\beta}$ are the components of $\hat{\boldsymbol{\beta}}_{\text{slse}}$ whose absolute values are larger than a threshold.

A rigorous proof of the selection-consistency of this method under some conditions

is a special case of the general results we establish in Theorems 1-2. This method is computationally simple, if $\hat{\boldsymbol{\Sigma}}$ is obtained by thresholding elements of $\boldsymbol{S}$. In what follows we derive a general method for variable selection, and show that it is selection-consistent and has better finite sample properties than the simple method by thresholding $\hat{\boldsymbol{\beta}}_{\text{slse}}$ in (8.1).

Note that $\boldsymbol{\Sigma}\boldsymbol{\beta} = \boldsymbol{\beta}_M$ and the true $\boldsymbol{\beta}$ is a solution to

$$\min_{\boldsymbol{\beta}} \left(\frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{2} - \boldsymbol{\beta}_M'\boldsymbol{\beta} \right)$$

for any fixed $\boldsymbol{\beta}_M$ and $\boldsymbol{\Sigma}$, and the ordinary least squares estimator solves the same minimization problem with $\boldsymbol{\beta}_M$ and $\boldsymbol{\Sigma}$ replaced by $\boldsymbol{X}'\boldsymbol{y}/n$ and $\boldsymbol{S}$, respectively, i.e., $\hat{\boldsymbol{\beta}}_{\text{lse}}$ is a solution to

$$\min_{\boldsymbol{\beta}} \left(\frac{\boldsymbol{\beta}'\boldsymbol{S}\boldsymbol{\beta}}{2} - \frac{\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}}{n} \right). \tag{8.2}$$

When $p > n$, the estimator $\hat{\boldsymbol{\beta}}_{\text{slse}}$ in (8.1) is an improvement to the least squares estimator by replacing $\boldsymbol{S}$ in (8.2) with an L_2-consistent sparse estimator $\hat{\boldsymbol{\Sigma}}$.

The well-known LASSO method (Tibshirani (1996)) adds an L_1 penalty term in minimization, i.e., the optimization problem in (8.2) is modified to

$$\min_{\boldsymbol{\beta}} \left(\frac{\boldsymbol{\beta}'\boldsymbol{S}\boldsymbol{\beta}}{2} - \frac{\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta}}{n} + \lambda_n \|\boldsymbol{\beta}\|_1 \right), \tag{8.3}$$

where $\|\boldsymbol{\beta}\|_1$ is the L_1-norm of the vector $\boldsymbol{\beta}$ and $\lambda_n \geq 0$ is a tuning parameter. It is known that the LASSO is not selection-consistent especially when $p > n$, unless a very strong assumption on $\boldsymbol{X}$ is imposed. Using the same idea in improving the least squares estimator, we consider regularizing the LASSO by replacing $\boldsymbol{S}$ in (8.3) by an

L_2-consistent sparse estimator $\hat{\boldsymbol{\Sigma}}$. This leads to solving the minimization problem

$$\min_{\boldsymbol{\beta}} \left(\frac{\boldsymbol{\beta}' \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta}}{2} - \frac{\boldsymbol{y}' \boldsymbol{X} \boldsymbol{\beta}}{n} + \lambda_n \|\boldsymbol{\beta}\|_1 \right). \tag{8.4}$$

In addition to the regularization step in the estimation of $\boldsymbol{\Sigma}$ and the L_1 penalization, our proposed method consists of a third step of regularization, i.e., thresholding the solution from the minimization problem (8.4). Let $\tilde{\boldsymbol{\beta}}$ be a solution to (8.4). Our final estimator of $\boldsymbol{\beta}$ is obtained by thresholding $\tilde{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\beta}_1 I(|\tilde{\beta}_1| > t_n), ..., \tilde{\beta}_p I(|\tilde{\beta}_p| > t_n) \right)', \tag{8.5}$$

where $\tilde{\beta}_j$ is the jth component of $\tilde{\boldsymbol{\beta}}$, $I(A)$ is the indicator function of the event A, and t_n is an appropriate threshold chosen as described in Chapter 8. The reason to implement the thresholding step is that the L_1 penalty in (8.4) may not shrink all small $\tilde{\beta}_j$'s to 0. Since our method can be viewed as adding two regularization steps to the LASSO, it will be referred to as the regularized LASSO (RLASSO).

If we choose $\lambda_n = 0$ in (8.4), then RLASSO reduces to thresholding $\hat{\boldsymbol{\beta}}_{\text{slse}}$ discussed earlier. The relationship between the RLASSO and other available methods can be described as follows. First, if we ignore the regularization step in the estimation of $\boldsymbol{\Sigma}$, i.e., we use $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S}$, then RLASSO becomes thresholding the LASSO estimator discussed in Meinshausen and Yu [25]; of course, if the last step of thresholding is also ignored, then RLASSO becomes LASSO. Second, if we choose $\hat{\boldsymbol{\Sigma}}$ to be the $p \times p$ identity matrix (assuming that all covariates are standardized), which can be viewed as a particular type of regularization by ignoring all correlations among components

of $\boldsymbol{x}_i$, and if we also choose $\lambda_n = 0$, then RLASSO becomes SIS. Finally, if we choose $\hat{\boldsymbol{\Sigma}}$ to be the inverse of the graphical LASSO estimator of $\boldsymbol{\Sigma}^{-1}$ and if we ignore the last step of thresholding, then RLASSO becomes the "scout" method proposed by Witten and Tibshirani [35].

In general, there are two ways to obtain regularized estimator of $\boldsymbol{\Sigma}$ depending on whether $\boldsymbol{\Sigma}$ is sparse or its inverse $\boldsymbol{\Sigma}^{-1}$ is sparse. If $\boldsymbol{\Sigma}$ is sparse, we may apply thresholding $\boldsymbol{S}$ as proposed by Bickel and Levina [3] or the adaptive thresholding method in Cai and Liu [9]. Both methods provide L_2-consistent estimators of $\boldsymbol{\Sigma}$. Define $\hat{\sigma}_{ij}$ to be the (i,j)th element of $\boldsymbol{S}$. The adaptive thresholding method estimates $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}} = (\hat{\sigma}^*_{ij})_{p\times p}$, where $\hat{\sigma}^*_{ij}$ is $\hat{\sigma}_{ij}$ being thresholded at

$$\delta \left\{ \frac{\log p}{n^2} \sum_{k=1}^{n} [x_{ki}x_{kj} - \hat{\sigma}_{ij}]^2 \right\}^{1/2}$$

by soft-thresholding or adaptive lasso thresholding. In this paper, we choose soft-thresholding. The tuning parameter $\delta \geq 2$ and it affects the probability of achieving model selection consistency in Theorem 9.1 and Theorem 9.2. Its optimal value is chosen by cross-validation in computation.

On the other hand, if $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is sparse, we may obtain a regularized estimator $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ and estimate $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Omega}}^{-1}$. Friedman, Hastie, and Tibshirani [21] proposed the graphical LASSO method to estimate $\boldsymbol{\Omega}$ by solving

$$\min_{\boldsymbol{\Omega}>0} \left\{ \text{trace}(\boldsymbol{\Omega}\boldsymbol{S}) - \log|\boldsymbol{\Omega}| + \lambda_{1n}\|\boldsymbol{\Omega}^*\|_1 \right\},$$

where $\boldsymbol{\Omega}^*$ is the off-diagonal elements of $\boldsymbol{\Omega}$, $\|\boldsymbol{A}\|_1$ is the L_1-norm of a matrix $\boldsymbol{A}$,

$\lambda_{1n} > 0$ is a tuning parameter, and $\boldsymbol{\Omega} > 0$ means that $\boldsymbol{\Omega}$ is positive definite. Rothman, Bickel, Levina, and Zhu [28] showed that $\hat{\boldsymbol{\Omega}}$ is consistent for $\boldsymbol{\Omega}$ under L_1-norm and L_2-norm if $\boldsymbol{\Omega}$ satisfies certain sparsity conditions. Cai, Liu, and Luo [7] proposed the CLIME estimator of $\boldsymbol{\Omega}$, which can be described as follows. Let $\hat{\boldsymbol{\Omega}}_1$ be the solution of the following problem:

$$\min \|\boldsymbol{\Omega}\|_1 \quad \text{subject to} \quad \|\boldsymbol{S}\boldsymbol{\Omega} - \boldsymbol{I}_p\|_\infty \leq \nu_n,$$

where $\|\boldsymbol{A}\|_\infty$ is the L_∞- or sup-norm of a matrix $\boldsymbol{A}$, $\boldsymbol{I}_k$ is the k-dimensional identity matrix, and ν_n is a tuning parameter. Let $\hat{\omega}_{ij}^1$ be the (i,j)th element of $\hat{\boldsymbol{\Omega}}_1$. The (i,j)th element of the CLIME estimator $\hat{\boldsymbol{\Omega}}$ is defined by

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

They proved that, when $\boldsymbol{\Omega}$ belongs to certain class of sparse matrices, the CLIME yields an estimator of $\boldsymbol{\Omega}$ that is consistent under L_1-norm and L_2-norm.

9 ASYMPTOTIC RESULTS

In this section we establish selection-consistency of the RLASSO. For asymptotic results when $p \to \infty$ at a rate faster than n, intuitively we need tail conditions on ε_i and x_j, the jth component of $\boldsymbol{x}$, a sparsity condition on the vector $\boldsymbol{\beta}$, and a sparsity condition on the covariance matrix $\boldsymbol{\Sigma}$ or its inverse $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$.

9.1 Results Under A Sparse Covariance Matrix

We first consider the situation where $\boldsymbol{\Sigma}$ is sparse. To measure the sparsity of $\boldsymbol{\Sigma}$, we consider

$$r_q = \max_{1 \le i \le p} \sum_{j=1}^{p} |\rho_{ij}|^q, \tag{9.1}$$

where ρ_{ij} is the (i,j)th element of $\boldsymbol{\Sigma}$ and $0 \le q < 1$ is a constant not depending on p or n. This sparsity measure was considered by Bickel and Levina [3]. When $q = 0$, r_0 is simply the maximum of numbers of nonzero components of rows of $\boldsymbol{\Sigma}$. If r_q diverges to ∞ at a rate much slower than p (e.g., condition (C4) in Theorem 1), then $\boldsymbol{\Sigma}$ is considered to be sparse. In this section, we estimate $\boldsymbol{\Sigma}$ by the adaptive thresholding estimator in Cai and Liu [9].

To measure the sparsity of $\boldsymbol{\beta}$, we consider

$$s_h = \sum_{j=1}^{p} |\beta_j|^h \tag{9.2}$$

for some $h \in [0, 1)$. In the special case where $h = 0$ in (9.2), s_0 is the number of non-zero components of $\boldsymbol{\beta}$.

Denote $\hat{\boldsymbol{\beta}}_M = \boldsymbol{X}'\boldsymbol{y}/n$. We first give two lemmas regarding $\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty$ and $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty$, which are useful for the main theorem. In what follows, a quantity is said to be a constant if it does not depend on n or p, but it may depend on some unknown population parameters. For two sequences a_n and b_n, $a_n \asymp b_n$ means that $a_n = O(b_n)$ and $b_n = O(a_n)$.

Lemma 9.1. *Assume that there exist positive constants m and M such that*
(C1) $\max_{1\le j\le p} \mathrm{E}[\exp(tx_j^2)] \le M$ *and* $\mathrm{E}[\exp(t\varepsilon_i^2)] \le M$ *for all* $|t| \le m$*;*
(C2) $\max_{1\le i\le n} \sigma_i \le M < \infty$ *and* $\max_{1\le j\le p} |\beta_j| \le M < \infty$.
Then, there exist positive constants C_1, C_2, and C_3 such that

$$P\left(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t\right) \le 2p^2 \exp(-C_1 n t^2/s_h^2) + 4p\exp(-C_2 n t^2),$$

for all $0 < t \le C_3 s_h$.

Lemma 9.2. *Assume conditions (C1)-(C2) and*
(C3) $\log p \asymp n^\tau$ *and* $\min_{jk} \mathrm{Var}(x_j x_k) \ge m$*, where* $0 < \tau < \frac{1}{3}$ *and* $m > 0$ *are constants. For any* λ_n *in* (8.4) *such that* $\lambda_n v_p \to 0$*, where* $v_p = \|\boldsymbol{\Sigma}^{-1}\|_1$*, there exist positive constants* C_4*,* C_5*,* C_6 *such that the solution to* (8.4) *satisfies*

$$P\left(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > t\right) \le 2p^2 \exp(-C_4 n t^2/(s_h v_p)^2) + 4p\exp(-C_5 n t^2/v_p^2) + C_6 n^{-1/2} p^{-(\delta-2)} (r_q v_p/t)^{1/(1-q)},$$

when n is sufficiently large.

Lemma 9.2 shows that $\tilde{\boldsymbol{\beta}}$ could be consistent to $\boldsymbol{\beta}$ under the supremum norm, when $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ are sparse, i.e., s_h and r_q diverge slowly with certain rates.

Let $\widehat{\mathcal{M}}_{\boldsymbol{\beta}}$ be the index set of nonzero components of $\hat{\boldsymbol{\beta}}$ defined in (8.5). The following result establishes the selection-consistency of $\widehat{\mathcal{M}}_{\boldsymbol{\beta}}$.

Theorem 9.1. *Assume conditions (C1)-(C3) and*
(C4) $s_h \asymp n^{\alpha_1}$, $r_q \asymp n^{\alpha_2}$, $v_p \asymp n^{\alpha_3}$, *where* α_1, α_2 *and* α_3 *are positive constants satisfying* $\alpha_1 + \alpha_3 < \frac{1}{2}(1-\tau)$ *and* $\alpha_2 + \alpha_3 < \frac{1}{2}(1-q)$.
If in (8.4), *we choose* $\lambda_n = M_1(n^{-1}\log p)$ *for some constant* $M_1 > 0$, *choose the threshold in* (8.5) *by*

$$t_n = M_2 n^{-\eta},$$

for some constant $M_2 > 0$, $0 < \eta < \min\{\frac{1}{2}(1-\tau) - \alpha_1 - \alpha_3, \frac{1}{2}(1-q) - \alpha_2 - \alpha_3\}$, *and choose* $a_n - 1 \asymp (\log n)^{-1}$, *then there exists a positive constant* C_7 *such that*

$$\begin{aligned} &1 - P\left(\mathcal{M}_{\boldsymbol{\beta},a_n t_n} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}} \subset \mathcal{M}_{\boldsymbol{\beta},t_n/a_n}\right) \\ &= O\left[\exp\left(-C_7(\log n)^{-2} n^{1-2\alpha_1-2\alpha_3-2\eta}\right) + (1/p)^{\delta-2}\left((\log n)(1/n)^{\frac{1-q}{2}-\alpha_2-\alpha_3-\eta}\right)^{1/(1-q)}\right], \end{aligned}$$

where $\mathcal{M}_{\boldsymbol{\beta},d_n}$ *denotes the index set of components of* $\boldsymbol{\beta}$ *whose absolute values are larger than* d_n. *In particular, if* $h = 0$ *in* (9.2) *and we additionally assume*
(C5) $\liminf_{n\to\infty} n^{\kappa} \min_{j\in\mathcal{M}_{\boldsymbol{\beta}}} |\beta_j| > 0$ *with* $\kappa < \min\{\frac{1}{2}(1-\tau) - \alpha_1 - \alpha_3, \frac{1}{2}(1-q) - \alpha_2 - \alpha_3\}$,
and if we choose λ_n *as above and*

$$t_n = M_2 n^{-\eta},$$

with $M_2 > 0$ *and* $\kappa < \eta < \min\{\frac{1}{2}(1-\tau) - \alpha_1 - \alpha_3, \frac{1}{2}(1-q) - \alpha_2 - \alpha_3\}$, *then*

$$1 - P\left(\widehat{\mathcal{M}}_{\boldsymbol{\beta}} = \mathcal{M}_{\boldsymbol{\beta}}\right) = O\left[\exp\left(-C_8 n^{1-2\alpha_1-2\alpha_3-2\eta}\right) + (1/p)^{\delta-2}(1/n)^{(\frac{1-q}{2}-\alpha_2-\alpha_3-\eta)/(1-q)}\right].$$

Indeed, Theorem 1 establishes a result more general than the selection-consistency defined by (7.2). That is, with probability tending to 1, the RLASSO eliminates all components of $\boldsymbol{\beta}$ whose absolute values are no larger than t_n/a_n, and retains all components of $\boldsymbol{\beta}$ whose absolute values are no smaller than $t_n a_n$. Since $a_n = 1 + c(\log n)^{-1} \to 1$, the RLASSO asymptotically eliminates or retains variables according to whether the components of $\boldsymbol{\beta}$ is smaller or larger than the threshold t_n. In particular, if $\boldsymbol{\beta}$ is sparse in the sense that s_0 (i.e., s_h in (9.2) with $h = 0$) diverges much slower than p, then the RLASSO is selection-consistent in the sense of (7.2), provided that the minimum of nonzero components of $\boldsymbol{\beta}$ does not decay too fast.

Condition (C1) requires that the distributions of x_j's and ε_i have exponential tails. Asymptotic results can also be established when the distributions of x_j's and ε_i have polynomial tails.

Lemma 9.3. *Assume that there exist constants* $M > 0$ *and* $l > 1$ *such that*
(C1′) $\max_{1\le j\le p} \mathrm{E}x_j^{4l} \le M$ *and* $\mathrm{E}\varepsilon_i^{4l} \le M$.
If (C2) also holds, then there exist some positive constants C_9 *and* C_{10} *such that*

$$P\left(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t\right) \le C_9 p^2 s_h^{2l} t^{-2l} n^{-l} + C_{10} p t^{-2l} n^{-l}.$$

Lemma 9.4. *Assume conditions (C1′), (C2) and*
(C3′) $p \asymp n^\tau$, *where* $\tau < \min\{l/2, l-1\}$.

For any λ_n in (8.4) such that $\lambda_n v_p \to 0$, there exist positive constants C_{11}, C_{12}, C_{13} such that

$$P\left(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > t\right) \leq C_{11} p^2 s_h^{2l} v_p^{2l} t^{-2l} n^{-l} + C_{12} p v_p^{2l} t^{-2l} n^{-l} + C_{13}\left(n^{-1/2} p^{-(\delta-2)} (r_q v_p / t)^{1/(1-q)} + n^{-\frac{l-1-\tau}{2}}\right),$$

when n is sufficiently large.

The following is a result similar to Theorem 1, obtained under the condition that x_j's and ε_i have sufficiently high moments and p diverges with a polynomial order of n.

Theorem 9.2. *Assume conditions (C1′), (C2), (C3′), and*
(C4′) $s_h \asymp n^{\alpha_1}$, $r_q \asymp n^{\alpha_2}$, $v_p \asymp n^{\alpha_3}$ where α_1, α_2 and α_3 are positive constants satisfying $\alpha_1 + \alpha_3 < \frac{1}{2} - \frac{\tau}{l}$ and $\alpha_2 + \alpha_3 < (\frac{1}{2} + [\delta - 2]\tau)(1 - q)$.
If in (8.4) we choose $\lambda_n = M_3 n^{-1/2}$ for some $M_3 > 0$ and choose the threshold in (8.5) by

$$t_n = M_4 n^{-\eta}$$

for some $M_4 > 0$, $0 < \eta < \min\{\frac{1}{2} - \frac{\tau}{l} - \alpha_1 - \alpha_3, (\frac{1}{2} + [\delta - 2]\tau)(1 - q) - \alpha_2 - \alpha_3\}$ and $a_n - 1 \asymp (\log n)^{-1}$, then

$$1 - P\left(\mathcal{M}_{\boldsymbol{\beta}, a_n t_n} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}} \subset \mathcal{M}_{\boldsymbol{\beta}, t_n / a_n}\right)$$
$$= O\left\{\left[(1/n)^{2l(\frac{1}{2} - \frac{\tau}{l} - \alpha_1 - \alpha_3 - \eta)} + (1/n)^{\frac{1}{1-q}\{(\frac{1}{2} + [\delta-2]\tau)(1-q) - \alpha_2 - \alpha_3 - \eta\}}\right](\log n)^{2l} + (1/n)^{\frac{l-1-\tau}{2}}\right\}.$$

In particular, if $h = 0$ in (9.2), (C5) holds for some constant $\kappa < \min\{\frac{1}{2} - \frac{\tau}{l} - \alpha_1 -$

$\alpha_3, (\frac{1}{2} + [\delta - 2]\tau)(1-q) - \alpha_2 - \alpha_3\}$, *and if we choose* λ_n *as above and*

$$t_n = M_4 n^{-\eta}$$

with $M_4 > 0$ *and* $\kappa < \eta < \min\{\frac{1}{2} - \frac{\tau}{l} - \alpha_1 - \alpha_3, (\frac{1}{2} + [\delta - 2]\tau)(1-q) - \alpha_2 - \alpha_3\}$, *then*

$$\begin{aligned} & 1 - P\left(\widehat{\mathcal{M}}_{\boldsymbol{\beta}} = \mathcal{M}_{\boldsymbol{\beta}}\right) \\ &= O\left[(1/n)^{2l(\frac{1}{2} - \frac{\tau}{l} - \alpha_1 - \alpha_3 - \eta)} + (1/n)^{\frac{1}{1-q}\{(\frac{1}{2} + [\delta-2]\tau)(1-q) - \alpha_2 - \alpha_3 - \eta)\}} + (1/n)^{\frac{l-1-\tau}{2}}\right]. \end{aligned}$$

9.2 Results Under A Sparse Inverse of Covariance Matrix

In general, there is no relationship between the sparsity of $\boldsymbol{\Sigma}$ and the sparsity of its inverse $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Hence, methods in two cases have to be developed separately, although asymptotic results are similar. To consider a sparsity measure of $\boldsymbol{\Omega}$, we still use the notation r_q in (9.1) but with ρ_{ij}'s replaced by the (i,j)th element of $\boldsymbol{\Omega}$. All asymptotic results in this section is based on this sparsity measurement of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}$ is estimated by the CLIME in Cai, Liu, and Luo [7].

Lemma 9.5. *Assume conditions (C1)-(C2) and*
(C3″) $\|\boldsymbol{\Omega}\|_1 \leq M$ *and* $\log p \asymp n^\tau$, *where* $0 < \tau < 1/4$.
For any $\lambda_n \to 0$, *there exist positive constants* C_{14}, C_{15}, C_{16} *such that*

$$\begin{aligned} P\left(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > t\right) \leq\ & 8\exp\left(-C_{14} n[t/s_h r_q]^{\frac{2}{1-q}}\right) + 4p^2 \exp\left(-C_{15} n t^2 / s_h^2\right) \\ & + 8p\exp(-C_{16} n t^2) \end{aligned}$$

for any $0 < t < 8M^{1-h}s_h$*, when* n *is sufficiently large.*

Theorem 9.3. *Assume conditions (C1), (C2), (C3″) and*
(C4″) $s_h \asymp n^{\alpha_1}$, $r_q \asymp n^{\alpha_2}$, *where* $\alpha_1 < \frac{1}{2}(1-\tau)$ *and* $\alpha_2 < \frac{1}{2}(1-q) - \alpha_1$ *are positive constants.*
If in (8.4)*, we choose* $\lambda_n = M_5(n^{-1}\log p)$ *for some constant* $M_5 > 0$ *and choose the threshold*

$$t_n = M_6 n^{-\eta},$$

with a constant $M_6 > 0$, $0 < \eta < \min\{\frac{1}{2}(1-\tau) - \alpha_1, \frac{1-q}{2} - \alpha_1 - \alpha_2\}$ *and* $a_n - 1 \asymp (\log n)^{-1}$*, then there exist constants* C_{17} *and* C_{18} *such that*

$$\begin{aligned} &1 - P\left(\mathcal{M}_{\boldsymbol{\beta},a_n t_n} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}} \subset \mathcal{M}_{\boldsymbol{\beta},t_n/a_n}\right) \\ &= O\left[\exp\left(-C_{17}\left[(\log n)^{-1} n^{\frac{1-q}{2}-\alpha_1-\alpha_2-\eta}\right]^{\frac{2}{1-q}}\right) + \exp\left(-C_{18}(\log n)^{-2} n^{2(\frac{1}{2}-\alpha_1-\eta)}\right)\right]. \end{aligned}$$

In particular, if (9.2) *holds for* $h = 0$ *and (C5) holds for some constant* $\kappa < \min\{\frac{1}{2}(1-\tau) - \alpha_1, \frac{1-q}{2} - \alpha_1 - \alpha_2\}$*, and if we choose* λ_n *as above and*

$$t_n = M_6 n^{-\eta},$$

with $M_6 > 0$ *and* $\kappa < \eta < \min\{\frac{1}{2}(1-\tau) - \alpha_1, \frac{1-q}{2} - \alpha_1 - \alpha_2\}$*, then*

$$1 - P\left(\widehat{\mathcal{M}}_{\boldsymbol{\beta}} = \mathcal{M}_{\boldsymbol{\beta}}\right) = O\left[\exp\left(-C_{19}\left[n^{\frac{1-q}{2}-\alpha_1-\alpha_2-\eta}\right]^{\frac{2}{1-q}}\right) + \exp\left(-C_{20} n^{2(\frac{1}{2}-\alpha_1-\eta)}\right)\right].$$

Lemma 9.6. *Assume conditions (C1′), (C2), and*
(C3‴) $\|\boldsymbol{\Omega}\|_1 \le M$ *and* $p \asymp n^{\tau}$*, where* $\tau < \min\{l/2, l-1\}$.

For any $\lambda_n \to 0$, there exist positive constants C_{21}, C_{22}, C_{23} and C_{24} such that

$$P\left(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > t\right) \leq C_{21}\left[\exp\left(-C_{22}n\left[t/r_q s_h\right]^{\frac{2}{1-q}}\right) + n^{-\frac{l-1-\tau}{2}}\right] + C_{23}p^2 s_h^{2l} t^{-2l} n^{-l} + C_{24} p t^{-2l} n^{-l},$$

for any $0 < t < 8M^{1-h}s_h$, when n is sufficiently large.

Theorem 9.4. *Assume conditions (C1′), (C2), (C3‴), and*
(C4‴) $s_h \asymp n^{\alpha_1}$, $r_q \asymp n^{\alpha_2}$, where $\alpha_1 < \frac{1}{2} - \frac{\tau}{l}$ and $\alpha_2 < \frac{1}{2}(1-q) - \alpha_1$ are positive constants.
If in (8.4), we choose $\lambda_n = M_7 n^{-1/2}$ for some $M_7 > 0$ and choose the threshold

$$t_n = M_8 n^{-\eta},$$

with constants $M_8 > 0$, $0 < \eta < \min\{\frac{1}{2} - \frac{\tau}{l} - \alpha_1, \frac{1}{2}(1-q) - \alpha_1 - \alpha_2\}$ and $a_n - 1 \asymp (\log n)^{-1}$, then there exists a positive constant C_{25} such that

$$1 - P\left(\mathcal{M}_{\boldsymbol{\beta}, a_n t_n} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}} \subset \mathcal{M}_{\boldsymbol{\beta}, t_n/a_n}\right) = O\left[\exp\left(-C_{25}\left[(\log n)^{-1} n^{\frac{1-q}{2} - \alpha_1 - \alpha_2 - \eta}\right]^{\frac{2}{1-q}}\right) + (1/n)^{2l(\frac{1}{2} - \frac{\tau}{l} - \alpha_1 - \eta)}(\log n)^{2l} + (1/n)^{\frac{l-1-\tau}{2}}\right].$$

In particular, if (9.2) holds for $h = 0$ and
(C5′) $\liminf_{n\to\infty} n^\kappa \min_{j \in \mathcal{M}_{\boldsymbol{\beta}}} |\beta_j| > 0$ for some constant $0 < \kappa < \min\{\frac{1}{2} - \frac{\tau}{l} - \alpha_1, \frac{1}{2}(1-q) - \alpha_1 - \alpha_2\}$, and if we choose λ_n as above and the threshold

$$t_n = M_8 n^{-\eta}$$

with $M_8 > 0$ *and* $\kappa < \eta < \min\{\frac{1}{2} - \frac{\tau}{l} - \alpha_1, \frac{1}{2}(1-q) - \alpha_2\}$*, then there exist a positive constant* C_{25} *such that*

$$\begin{aligned} & 1 - P\left(\widehat{\mathcal{M}_{\boldsymbol{\beta}}} = \mathcal{M}_{\boldsymbol{\beta}}\right) \\ = & O\left[\exp\left(-C_{26}\left[n^{\frac{1-q}{2}-\alpha_1-\alpha_2-\eta}\right]^{\frac{2}{1-q}}\right) + (1/n)^{2l(\frac{1}{2}-\frac{\tau}{l}-\alpha_1-\eta)} + (1/n)^{\frac{l-1-\tau}{2}}\right]. \end{aligned}$$

10 COMPARISON OF LASSO AND RLASSO

Meinshausen and Bühlmann [24], Zhao and Yu [38], Zhang and Huang [36] and Meinshausen and Yu [25] studied asymptotic properties of the LASSO for variable selection. It is well-known that the LASSO requires some strong conditions on $\boldsymbol{X}$ in order to have good asymptotic performance such as selection-consistency or screening-consistency. Recall that $\mathcal{M}_{\boldsymbol{\beta}}$ is the index set for non-zero components of $\boldsymbol{\beta}$ and s_0 is the number of elements in $\mathcal{M}_{\boldsymbol{\beta}}$. Without loss of generality we assume that

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix}$$

where $\boldsymbol{S}_{11} = \boldsymbol{X}_1'\boldsymbol{X}_1/n$, $\boldsymbol{S}_{21} = \boldsymbol{X}_2'\boldsymbol{X}_1/n$, $\boldsymbol{X}_1$ is the s_0 columns of $\boldsymbol{X}$ that corresponds to non-zero components of $\boldsymbol{\beta}$, and $\boldsymbol{X}_2$ is the other $p - s_0$ columns of $\boldsymbol{X}$ that corresponds to zero components of $\boldsymbol{\beta}$. Zhao and Yu [38] pointed out that the following strong irrepresentable condition (SIC) is sufficient for LASSO to achieve selection-consistency. The matrix $\boldsymbol{S}$ is said to have SIC if

$$\eta_\infty = 1 - \|\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\operatorname{sign}(\boldsymbol{\beta}_1)\|_\infty \geq \gamma, \tag{10.1}$$

where γ is a positive constant, $\boldsymbol{\beta}_1 = \{\beta_j, j \in \mathcal{M}_{\boldsymbol{\beta}}\}$, and $\operatorname{sign}(\boldsymbol{\beta}_1)$ is the vector whose components are the signs of components of $\boldsymbol{\beta}$. The SIC is also essentially necessary for LASSO to be selection-consistency (see Bühlmann and Van De Geer [6]).

However, as mentioned in Bühlmann and Van De Geer [6], the SIC is rather restrictive and could fail badly in many cases. In the following, we conduct several

simulation studies to see how strong the SIC is. We generate $\boldsymbol{x}$ from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where and $\boldsymbol{e}_k$ is the k-dimensional vector of 1's, and consider two cases:

Case 1: $n = 100$, $p = 100$, $s_0 = 4$, $k = 9$;

Case 2: $n = 200$, $p = 300$, $s_0 = 6$, $k = 14$.

The first s_0 elements of $\boldsymbol{\beta}$ are positive and the remaining elements of $\boldsymbol{\beta}$ are zero. To see how ρ affects the chance that the SIC holds, we run 1000 simulations for some values of ρ and report the proportion that the SIC holds in the following table. It is clear that the proportion drops dramatically as ρ increases.

ρ	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Case 1	0.996	0.974	0.939	0.819	0.504	0.121	0.004
Case 2	0.998	0.981	0.926	0.748	0.291	0.005	0.000

Next, we check how s_0 and k affect the chance that the SIC holds. The settings are the same except that $\rho = 0.2$ is fixed and s_0 and k vary. The results in the following table indicate that the SIC is very sensitive to the value of s_0. When s_0 is small, the SIC mostly holds. When s_0 is larger than some critical value, however, the SIC rarely holds.

$k \backslash s_0$	3	5	7	9
5	1.00	0.93	0.17	0.00
10	1.00	0.92	0.07	0.00
15	1.00	0.89	0.06	0.00
20	1.00	0.86	0.04	0.00

$n = 100$, $p = 100$

$k \backslash s_0$	3	5	7	9
5	1.00	1.00	0.98	0.65
10	1.00	1.00	0.96	0.39
15	1.00	1.00	0.96	0.31
20	1.00	1.00	0.95	0.28

$n = 200$, $p = 300$

Since RLASSO applies a second-step thresholding after solving the LASSO-like problem (8.4), it avoids the requirement of a condition like the SIC. To illustrate the gain of using RLASSO, we carry out a simulation comparison of RLASSO and LASSO in the example appeared in Section 3.2 of Zhao and Yu [38]. We consider $n = 100$, $p = 32$, and $\boldsymbol{\beta} = (7, 4, 2, 1, 1, 0, \ldots, 0)$. We first obtain a covariance matrix $\boldsymbol{\Sigma}$ generated from the Wishart(p, p) distribution. Then, we generate n iid $\boldsymbol{x}_i$'s from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and normalize $\boldsymbol{x}_i$'s to have mean 0 and variance 1. The matrix $\boldsymbol{X}$ containing $\boldsymbol{x}_i$ as its ith row is treated as fixed in 1000 simulations. In each simulation run, n iid ϵ_i's are generated from $N(0, 1)$ and y_i is obtained by (7.1) with $\mu = 0$ and $\sigma_i^2 = 0.1$ for each i. In each simulation, as Zhao and Yu did, we first run LASSO to calculate its entire path to see if there exists a model along the path that matches the true model. After that, we run RLASSO with $\hat{\boldsymbol{\Sigma}}$ in (8.4) being an adaptive thresholding estimator of $\boldsymbol{\Sigma}$ and other tuning parameters being chosen by cross-validation. After 1000 simulations, we calculate the percentages that LASSO and RLASSO selecting the correct model. Finally, we repeat independently the above process 100 times and plot in Figure 10.1 the 100 simulation percentages against $\eta_\infty = 1 - \|\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\,\mathrm{sign}(\boldsymbol{\beta}_1)\|_\infty$.

Figure 10.1 shows that the performance of LASSO is good when $\eta_\infty > 0.05$, which occurs only 45 times in the 100 designs. In other 55 cases, LASSO does not perform well. On the other hand, RLASSO performs very well in all 100 cases, regardless of whether the SIC holds or not.

Meinshausen and Yu [25] introduced another sparsity condition on $\boldsymbol{S}$, called

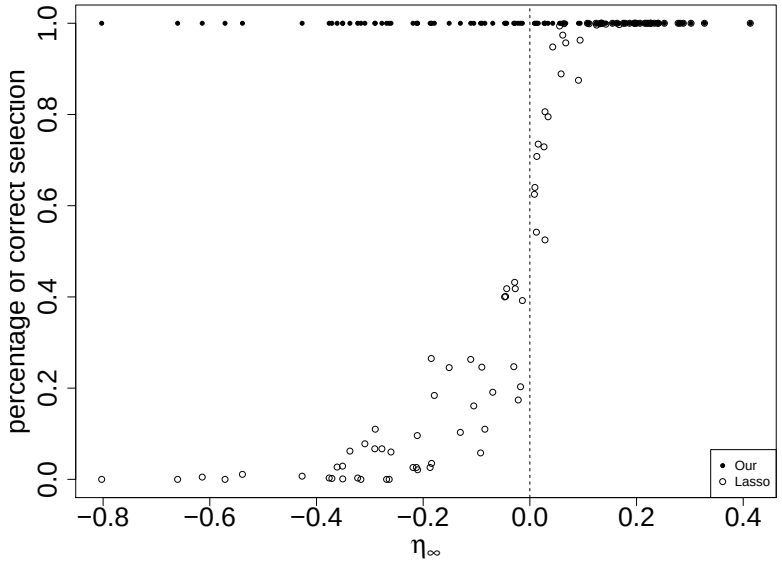

Figure 10.1: Comparison of LASSO and RLASSO

incoherent design condition (IDC). Let

$$\phi_{\min}(m) = \min_{\boldsymbol{z}:||\boldsymbol{z}||_{l_0} \leq \lceil m \rceil} \frac{\boldsymbol{z}'\boldsymbol{S}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}} \quad \text{and} \quad \phi_{\max}(m) = \max_{\boldsymbol{z}:||\boldsymbol{z}||_{l_0} \leq \lceil m \rceil} \frac{\boldsymbol{z}'\boldsymbol{S}\boldsymbol{z}}{\boldsymbol{z}'\boldsymbol{z}}$$

be the m-sparse minimal eigenvalue and m-sparse maximal eigenvalue of $\boldsymbol{S}$, respectively. $\boldsymbol{S}$ is said to satisfy the IDC if there exists a positive sequence e_n such that

$$\liminf_{n\to\infty} \frac{e_n \phi_{\min}(e_n^2 s_0)}{\phi_{\max}(s_0 + \min\{n, p\})} \geq 18, \tag{10.2}$$

where s_0 still denotes the number of non-zero elements in $\boldsymbol{\beta}$. Meinshausen and Yu [25] showed that if $\boldsymbol{S}$ satisfies the IDC, the true $\boldsymbol{\beta}$ is sparse, and $\boldsymbol{\beta}$'s minimal non-zero component does not converge to 0 too fast, then LASSO followed by a thresholding

could achieve selection-consistency.

However, when $p \gg n$, the IDC could still fail for many $\boldsymbol{S}$. The IDC essentially requires $\lambda_{\min}(\boldsymbol{S}_0)$ cannot converge to zero too fast, for any submatrix $\boldsymbol{S}_0$ of $\boldsymbol{S}$ with certain rank. When $p \gg n$, however, many submatrices of $\boldsymbol{S}$ could be singular or close to singular. In deed, any $m \times m$ submatrix of $\boldsymbol{S}$ is singular if $m > n$. Even for $m < n$, it is still very hard to check if the IDC holds for a particular $\boldsymbol{S}$.

All the difficulty is caused by the fact that $\boldsymbol{S}$ is not a good estimate of $\boldsymbol{\Sigma}$ in a high-dimensional setting. Again, instead of imposing strong conditions on $\boldsymbol{S}$, our approach replaces $\boldsymbol{S}$ with a regularized estimator $\hat{\boldsymbol{\Sigma}}$. When $\boldsymbol{\Sigma}$ is sparse, many regularized estimators $\hat{\boldsymbol{\Sigma}}$ in the literature, such as the adaptive thresholding estimator in Cai and Liu [9], satisfy $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \xrightarrow{P} 0$. If $\lambda_{\min}(\boldsymbol{\Sigma})$ is bounded away from 0, then $\lambda_{\min}(\hat{\boldsymbol{\Sigma}})$ is also asymptotically bounded away from 0. Hence, for any submatrix $\hat{\boldsymbol{\Sigma}}_0$ of $\hat{\boldsymbol{\Sigma}}$, $\lambda_{\min}(\hat{\boldsymbol{\Sigma}}_0) > 0$. That is, the IDC always holds for $\hat{\boldsymbol{\Sigma}}$. The same conclusion can be made when $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is sparse, $\lambda_{\max}(\boldsymbol{\Omega})$ is bounded away from ∞, and an L_2-consistent estimator $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ is adopted.

11 NUMERICAL RESULTS

In this section, we conduct several simulation studies to compare the following variable selection methods.

1. RLASSO(AT): RLASSO with $\hat{\mathbf{\Sigma}}$ in (8.4) being the adaptive thresholding estimator of $\mathbf{\Sigma}$.
2. RLASSO(CLIME): RLASSO with $\hat{\mathbf{\Sigma}}$ in (8.4) being the inverse of CLIME.
3. RLASSO(GLASSO): RLASSO with $\hat{\mathbf{\Sigma}}$ in (8.4) being the inverse of Graphical LASSO estimator.
4. LASSO: the ordinary LASSO method.
5. LASSO+T: the ordinary LASSO followed by a thresholding step.
6. Scout(1, 1): the Scout(1, 1) method in Witten and Tibshirani [35].
7. SLSE+T: the Sparse Least Square Estimator as described in (8.1) followed by a thresholding step.
8. SIS: the Sure Independence Screening method in Fan and Lv [17].

For tuning parameters in the above methods, we apply a 5-fold cross validation to choose the optimal ones that minimize the mean square error. For example, there are three tuning parameters in RLASSO(AT), namely δ, λ_n and t_n. We choose their optimal values $(\hat{\delta}, \hat{\lambda}_n, \hat{t}_n)$ by

$$(\hat{\delta}, \hat{\lambda}_n, \hat{t}_n) = \underset{\delta, \lambda_n, t_n}{\operatorname{argmin}} \frac{1}{5} \sum_{j=1}^{5} \sum_{i \in \text{fold } j} \{y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}^{-j}(\delta, \lambda_n, t_n)\}^2,$$

where $\hat{\boldsymbol{\beta}}^{-j}(\delta, \lambda_n, t_n)$ is the estimate of $\boldsymbol{\beta}$ without using jth fold data and a fixed set (δ, λ_n, t_n). Tuning parameters of other methods are chosen by the same 5-fold cross

validation scheme. In particular, for SIS, we determine the number of kept variables by 5-fold cross validation.

We examine the aforementioned variable selection methods under the following models.

Model 1: $n = 50$, $p = 100$, $\boldsymbol{\beta} = (2\boldsymbol{e}_8, \mathbf{0}_{p-8})$, where $\boldsymbol{e}_k$ is the k-dimensional vector with all components equal to 1 and $\mathbf{0}_k$ is the k-dimensional vector with all components equal to 0. $\boldsymbol{x}_i$'s are iid from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{B}_{10\times 10} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-10} \end{pmatrix} \quad \text{and} \quad \boldsymbol{B}_{10\times 10} = \begin{pmatrix} \boldsymbol{I}_8 & 0.2\boldsymbol{e}_{8\times 2} \\ 0.2\boldsymbol{e}_{2\times 8} & \boldsymbol{I}_2 \end{pmatrix},$$

$\boldsymbol{e}_{m\times n}$ is the $m\times n$ matrix with all elements equal to 1, and $\boldsymbol{I}_k$ is the k-dimensional identity matrix. $\boldsymbol{y}$ is generated from (7.1) with $\mu = 0$ and $\sigma_i = 1$ for $i = 1, \ldots, n$.

Model 2: $n = 50$, $p = 100$, $\boldsymbol{\beta} = (4, -1.2, 2.5, 1.5, 4.6, \mathbf{0}_{p-5})$. $\boldsymbol{x}_i$'s are iid from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \text{toeplitz}(1, 0.49, 0.44, 0.40, 0.36, 0.32, 0.29, 0.26, 0.23, \mathbf{0}_{p-9})$. $\boldsymbol{y}$ is generated from (7.1) with $\mu = 0$ and $\sigma_i = 1$ for $i = 1, \ldots, n$.

Model 3: $n = 50$, $p = 100$, $\boldsymbol{\beta} = (4, -1.2, 2.5, 1.5, 4.6, \mathbf{0}_{p-5})$. $\boldsymbol{x}_i$'s are iid from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the inverse of $\boldsymbol{\Omega} = \text{toeplitz}(1, 0.5, \mathbf{0}_{p-2})$. $\boldsymbol{y}$ is generated from (7.1) with $\mu = 0$ and $\sigma_i = 1$ for $i = 1, \ldots, n$.

Model 4: $n = 50$, $p = 100$, $\boldsymbol{\beta} = (\mathbf{1}_9, -7.2, \mathbf{0}_{p-10})$. $\boldsymbol{x}_i$'s are iid from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.2\boldsymbol{I}_{10} + 0.8\boldsymbol{e}_{10}\boldsymbol{e}_{10}' & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{p-10} \end{pmatrix}.$$

$\boldsymbol{y}$ is generated from (7.1) with $\mu = 0$ and $\sigma_i = 1$ for $i = 1, \ldots, n$.

In Model 1 and Model 4, both the covariance matrix $\boldsymbol{\Sigma}$ and its inverse $\boldsymbol{\Omega}$ are sparse. Model 2 and Model 3 represent two different sparsity structures. In Model 2, only $\boldsymbol{\Sigma}$ is sparse, while in Model 3, only $\boldsymbol{\Omega}$ is sparse. Model 1 represents a setting in which the SIC in (10.1) can hardly hold, since (10.1) fails with $\boldsymbol{S}$ replaced by $\boldsymbol{\Sigma}$. Model 4 is motivated by a "false negative" example in Fan and Lv [18], where condition (7.3) is violated. $\beta_{10} = -7.2$ in this model has a marginal effect of 0, but indeed has the largest effect. Also, condition (7.4) does not hold in any of Models 1-3, although it holds under Model 4.

We measure the performance of each variable selection method by the following criteria.

1. Sensitivity (SENS): the proportion of true non-zero β_j being estimated as non-zero.
2. Specificity (SPEC): the proportion of true zero β_j being estimated as zero.
3. Coverage probability (CR): the probability that the selected model covers the true model, i.e. $\mathcal{M}_{\boldsymbol{\beta}} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}}$.
4. Hit probability (HP): the probability that the selected model is identical to the true model, i.e. $\mathcal{M}_{\boldsymbol{\beta}} = \widehat{\mathcal{M}}_{\boldsymbol{\beta}}$.
5. Model size (SIZE): the size of selected model.

For each model, we carry out 200 simulation runs. The mean of above measurements, with simulation standard error in parenthesis, are reported in Tables 11.1-11.4.

It is clear that, for each model, at least one version of RLASSO performs much better than the other methods in terms of hit probability. In Tables 11.1 and

11.2, the hit probability of RLASSO(AT) is much larger than that of LASSO+T, which supports our idea of replacing the unstable covariance estimator $\boldsymbol{S}$ in (8.3) with a regularized estimator $\hat{\boldsymbol{\Sigma}}$ as in (8.4), since this is the only difference between RLASSO(AT) and LASSO+T. The performance of various versions of RLASSO depends on sparsity of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$. When $\boldsymbol{\Sigma}$ is sparse as in Model 2, RLASSO(AT) is better than RLASSO(CLIME) and RLASSO(GLASSO). When $\boldsymbol{\Omega}$ is sparse as in Model 3, RLASSO(CLIME) and RLASSO(GLASSO) are better than RLASSO(AT). In general, RLASSO(CLIME) and RLASSO(GLASSO) performs similarly, as they only differ in the estimation of $\boldsymbol{\Omega}$.

Table 11.1: Simulation Results of Eight Methods under Model 1

	SENS(%)	SPEC(%)	CP(%)	HP(%)	SIZE
RLASSO(AT)	95.62(1.10)	98.70(0.18)	91.00(2.03)	64.00(3.40)	8.85(1.99)
RLASSO(CLIME)	61.56(1.33)	91.05(0.46)	3.00(1.21)	0.00(0.00)	13.15(6.87)
RLASSO(GLASSO)	65.75(1.53)	92.41(0.40)	9.50(2.08)	0.00(0.00)	12.24(6.11)
Scout(1,1)	81.62(1.16)	81.39(1.05)	26.00(3.11)	0.00(0.00)	23.65(14.3)
LASSO	92.38(0.91)	80.36(0.51)	65.00(3.38)	0.00(0.00)	25.45(6.81)
LASSO+T	82.50(1.63)	94.59(0.28)	51.00(3.54)	4.50(1.47)	11.57(3.47)
SLSE+T	78.12(1.47)	68.58(1.27)	26.50(3.13)	0.00(0.00)	35.16(17.7)
SIS	93.56(0.88)	59.07(0.76)	67.00(3.33)	0.00(0.00)	45.15(0.74)

In comparison of LASSO and LASSO+T, Tables 11.1, 11.2 and 11.4 indicate that a second step of thresholding is useful to improve hit probability. Even though LASSO can shrink variables toward zero to reach a sparse model, the shrinkage may not be enough to remove all unimportant variables.

Since the marginal effect $\boldsymbol{\beta}_M$ differs very much from $\boldsymbol{\beta}$, the SIS can hardly select the true model, which is reflected by an almost zero hit probability in Tables 11.1-11.4.

Table 11.2: Simulation Results of Eight Methods under Model 2

	SENS(%)	SPEC(%)	CP(%)	HP(%)	SIZE
RLASSO(AT)	89.40(0.84)	99.98(0.01)	52.00(3.54)	51.00(3.54)	4.49(0.61)
RLASSO(CLIME)	75.50(0.90)	99.24(0.07)	4.00(1.39)	1.00(0.71)	4.50(1.35)
RLASSO(GLASSO)	76.60(0.53)	99.98(0.01)	0.00(0.00)	0.00(0.00)	3.85(0.40)
Scout(1,1)	80.10(0.17)	98.24(0.18)	1.00(0.71)	0.00(0.00)	5.67(2.38)
LASSO	93.70(0.66)	90.81(0.55)	68.50(3.29)	0.00(0.00)	13.42(7.66)
LASSO+T	82.90(0.71)	99.99(0.01)	21.00(2.89)	21.00(2.89)	4.15(0.51)
SLSE+T	77.30(0.81)	85.23(0.79)	6.50(1.75)	0.00(0.00)	17.90(10.9)
SIS	98.50(0.37)	62.04(1.21)	92.50(1.87)	0.50(0.50)	40.98(1.16)

Table 11.3: Simulation Results of Eight Methods under Model 3

	SENS(%)	SPEC(%)	CP(%)	HP(%)	SIZE
RLASSO(AT)	80.70(0.71)	99.18(0.10)	8.50(1.98)	2.50(1.11)	4.82(1.49)
RLASSO(CLIME)	88.00(0.71)	98.58(0.15)	40.50(3.48)	10.00(2.13)	5.75(2.18)
RLASSO(GLASSO)	84.20(0.61)	97.35(0.31)	22.00(2.94)	5.00(1.54)	6.72(4.27)
Scout(1,1)	87.40(0.68)	88.73(0.39)	37.00(3.42)	0.00(0.00)	15.07(5.39)
LASSO	80.80(0.28)	97.95(0.16)	4.00(1.39)	0.00(0.00)	5.99(2.29)
LASSO+T	80.50(0.22)	99.12(0.10)	2.50(1.11)	0.00(0.00)	4.87(1.40)
SLSE+T	97.10(0.54)	85.05(0.63)	86.50(2.42)	0.00(0.00)	35.06(8.62)
SIS	99.20(0.28)	62.76(1.28)	96.00(1.39)	0.00(0.00)	40.34(1.22)

Table 11.4: Simulation Results of Eight Methods under Model 4

	SENS(%)	SPEC(%)	CP(%)	HP(%)	SIZE
RLASSO(AT)	89.80(1.06)	98.16(0.21)	54.00(3.53)	38.00(3.44)	10.63(2.67)
RLASSO(CLIME)	97.80(0.29)	89.73(0.43)	78.00(2.94)	1.00(0.71)	19.02(5.54)
RLASSO(GLASSO)	96.80(0.36)	88.27(0.42)	70.00(3.25)	0.00(0.00)	20.24(5.43)
Scout(1,1)	99.70(0.12)	33.32(0.32)	97.00(1.21)	0.00(0.00)	69.98(4.05)
LASSO	93.40(0.61)	76.92(0.78)	55.50(3.52)	0.00(0.00)	30.11(10.1)
LASSO+T	89.00(0.88)	98.22(0.17)	48.50(3.54)	33.00(3.33)	10.51(2.07)
SLSE+T	99.80(0.10)	73.41(0.75)	98.00(0.99)	0.00(0.00)	33.91(9.52)
SIS	83.70(1.91)	61.20(0.98)	30.00(3.25)	0.00(0.00)	43.29(1.07)

Since condition (7.3) holds but condition (7.4) does not hold in Models 1-3, the SIC has a high SENS but a low SPEC. In some parsimonies case, such as Model 4, even its SENS and coverage probability are not high.

Note that SLSE+T is computationally much simpler than RLASSO, as it does not require to solve an optimization problem and has less parameters to tune. In some cases, e.g., Model 3 and Model 4, its coverage probability is large. In particular, for Model 4, since SLSE takes correlation into account, its coverage probability is much larger than that of SIS. However, the performance of SLSE+T is in general whose than that of RLASSO, indicating that the extra computation effort is worthwhile.

12 PROOFS

Proof of Lemma 9.1. From (7.1),

$$\begin{aligned}\hat{\beta}_{M_j} - \beta_{M_j} &= \frac{1}{n}\sum_{i=1}^{n} x_{ij}\left(\mu + \sum_{k=1}^{p}\beta_k x_{ik} + \sigma_i\epsilon_i\right) - \sum_{k=1}^{p}\mathrm{E}(x_j x_k)\beta_k \\ &= \sum_{k=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \mathrm{E}(x_j x_k)\right]\beta_k + \frac{1}{n}\sum_{i=1}^{n}\sigma_i x_{ij}\epsilon_i + \frac{1}{n}\sum_{i=1}^{n}\mu x_{ij}.\end{aligned}$$

Then,

$$\begin{aligned}P\left(|\hat{\beta}_{M_j} - \beta_{M_j}| > t\right) \le\ & P\left(\sum_{k=1}^{p}\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \mathrm{E}(x_j x_k)\right| |\beta_k| > \frac{t}{3}\right) \\ &+ P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i x_{ij}\epsilon_i\right| > \frac{t}{3}\right) \\ &+ P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mu x_{ij}\right| > \frac{t}{3}\right)\end{aligned} \tag{12.1}$$

Under condition (C1), applying Lemma 1 of Cai and Liu [9] to $\pm(x_{ij}x_{ik} - \mathrm{E}(x_{ij}x_{ik}))$ gives that there exist $D_1 > 0$ and $D_2 > 0$ such that

$$\max_{jk} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \mathrm{E}(x_j x_k)\right| > t\right) \le 2\exp(-D_1 n t^2),$$

for all $0 < t \leq D_2$. Then, under condition (C2), it follows by Bonferroni inequality that there exist $C_1 > 0$ and $C_3 > 0$ such that, for any $1 \leq j \leq p$,

$$\begin{aligned}
&P\left(\sum_{k=1}^{p}\left|\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ik} - \mathrm{E}(x_j x_k)\right| |\beta_k| > \frac{t}{3}\right) \\
&\leq P\left(M^{1-h}s_h \max_k \left|\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ik} - \mathrm{E}(x_{ij}x_{ik})\right| > \frac{t}{3}\right) \\
&\leq p \cdot \max_{jk} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ik} - \mathrm{E}(x_j x_k)\right| > \frac{t}{3M^{1-h}s_h}\right) \\
&\leq 2p\exp(-C_1 n t^2/s_h^2),
\end{aligned} \tag{12.2}$$

for all $0 < t \leq C_3 s_h$. Similarly,

$$\max_{1\leq j\leq p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i x_{ij}\epsilon_i\right| > \frac{t}{3}\right) \leq 2\exp(-C_2 n t^2), \tag{12.3}$$

$$\max_{1\leq j\leq p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mu x_{ij}\right| > \frac{t}{3}\right) \leq 2\exp(-C_2 n t^2), \tag{12.4}$$

for some $C_2 > 0$.

Therefore,

$$P\left(|\hat{\beta}_{M_j} - \beta_{M_j}| > t\right) \leq 2p\exp(-C_1 n t^2/s_h^2) + 4\exp(-C_2 n t^2). \tag{12.5}$$

Then, Lemma 9.1 follows by

$$P\left(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t\right) \leq \sum_{j=1}^{p} P\left(|\hat{\beta}_{M_j} - \beta_{M_j}| > t\right).$$

□

Proof of Lemma 9.2. By Karush-Kuhn-Tucker conditions, the solution $\tilde{\boldsymbol{\beta}}$ to (8.4) satisfies that

$$\hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_M = -\lambda_n \boldsymbol{Z}, \tag{12.6}$$

where $\boldsymbol{Z}$ has the form of

$$\boldsymbol{Z} = \begin{cases} 1, & \text{if } \tilde{\beta}_j > 0; \\ -1, & \text{if } \tilde{\beta}_j < 0; \\ \in [-1, 1], & \text{if } \tilde{\beta}_j = 0. \end{cases} \tag{12.7}$$

Simple algebra from (12.6) yields

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}[\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M - \lambda_n \boldsymbol{Z} - (\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta} - (\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})]. \tag{12.8}$$

Hence,

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty \leq v_p(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty + \|\lambda_n \boldsymbol{Z}\|_\infty + \|(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}\|_\infty + \|(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_\infty).$$

Equivalently,

$$\frac{1}{v_p}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty - \|(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_\infty \leq \|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty + \|\lambda_n \boldsymbol{Z}\|_\infty + \|(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}\|_\infty$$

Then, by $\|(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_\infty \le \|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_\infty$, it holds that

$$
\begin{aligned}
&P(\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_\infty > t) \\
&= P\left(\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_\infty > t, \|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1 \le \frac{1}{2v_p}\right) + P\left(\|\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_\infty > t, \|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1 > \frac{1}{2v_p}\right) \\
&\le P\left(\|\hat{\boldsymbol{\beta}}_M-\boldsymbol{\beta}_M\|_\infty + \|\lambda_n \boldsymbol{Z}\|_\infty + \|(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\beta}\|_\infty > \frac{t}{2v_p}, \|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1 \le \frac{1}{2v_p}\right) \\
&\quad + P\left(\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1 > \frac{1}{2v_p}\right) \\
&\le P\left(\|\hat{\boldsymbol{\beta}}_M-\boldsymbol{\beta}_M\|_\infty > \frac{t}{6v_p}\right) + P\left(\|\lambda_n \boldsymbol{Z}\|_\infty > \frac{t}{6v_p}\right) + P\left(\|(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\beta}\|_\infty > \frac{t}{6v_p}\right) \\
&\quad + P\left(\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1 > \frac{1}{2v_p}\right) \\
&:= I + II + III + IV
\end{aligned}
$$

By Lemma 9.1, there exist positive constants C_4 and C_5 such that

$$
I \le 2p^2 \exp(-C_4 n t^2/(s_h v_p)^2) + 4p \exp(-C_5 n t^2/v_p^2). \tag{12.9}
$$

By the choice of λ_n, $II = 0$, when n is sufficiently large. For III, under condition (C2), $\|(\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma})\boldsymbol{\beta}\|_\infty \le \|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1\|\boldsymbol{\beta}\|_\infty \le M\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1$. Then, it follows from Theorem 1(i) of Cai and Liu [9] that,

$$
III + IV \le 2P\left(\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_1 > \frac{t}{6Mv_p}\right) \le C_6 n^{-1/2} p^{-(\delta-2)} (r_q v_p/t)^{1/(1-q)}, \tag{12.10}
$$

for some $C_6 > 0$. This completes the proof of the lemma. □

Proof of Theorem 9.1. Note that $\widehat{\mathcal{M}}_{\boldsymbol{\beta}} = \mathcal{M}_{\tilde{\boldsymbol{\beta}},t_n}$ and

$$
\begin{aligned}
&P\left(\mathcal{M}_{\boldsymbol{\beta},a_nt_n} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}}\right) \\
&= 1 - P\left(\cup_{j:|\beta_j|>a_nt_n}\left\{|\tilde{\beta}_j| \le t_n\right\}\right) \\
&\ge 1 - P\left(\cup_{j:|\beta_j|>a_nt_n}\left\{|\tilde{\beta}_j - \beta_j| > (a_n-1)t_n\right\}\right) \\
&\ge 1 - P\left(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > (a_n-1)t_n\right) \\
&\ge 1 - O\left[\exp\left(-C_7(\log n)^{-2}n^{1-2\alpha_1-2\alpha_3-2\eta}\right) + (1/p)^{\delta-2}\left((\log n)(1/n)^{\frac{1-q}{2}-\alpha_2-\alpha_3-\eta}\right)^{1/(1-q)}\right]
\end{aligned}
$$

by Lemma 9.2 and the choice of a_n and t_n. Similarly,

$$
\begin{aligned}
&P(\widehat{\mathcal{M}_{\boldsymbol{\beta}}} \subset \mathcal{M}_{\boldsymbol{\beta},t_n/a_n}) \\
&= P\left(\cap_{j:|\beta_j|\le t_n/a_n}\left\{|\tilde{\beta}_j| \le t_n\right\}\right) \\
&\ge 1 - P\left(\cup_{j:|\beta_j|\le t_n/a_n}\left\{|\tilde{\beta}_j - \beta_j| > (1-a_n^{-1})t_n\right\}\right) \\
&\ge 1 - P\left(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > (1-a_n^{-1})t_n\right) \\
&= 1 - O\left[\exp\left(-C_7(\log n)^{-2}n^{1-2\alpha_1-2\alpha_3-2\eta}\right) + (1/p)^{\delta-2}\left((\log n)(1/n)^{\frac{1-q}{2}-\alpha_2-\alpha_3-\eta}\right)^{1/(1-q)}\right].
\end{aligned}
$$

This completes the proof of the first part of Theorem 9.1. In particular, if we choose

$h = 0$,

$$
\begin{aligned}
& P(\mathcal{M}_{\boldsymbol{\beta}} \subset \widehat{\mathcal{M}}_{\boldsymbol{\beta}}) \\
&= 1 - P\left(\cup_{j \in \mathcal{M}_{\boldsymbol{\beta}}} \left\{ |\tilde{\beta}_j| \leq t_n \right\}\right) \\
&\geq 1 - P\left(\cup_{j \in \mathcal{M}_{\boldsymbol{\beta}}} \left\{ |\beta_j| - |\tilde{\beta}_j - \beta_j| \leq t_n \right\}\right) \\
&\geq 1 - P\left(\cup_{j \in \mathcal{M}_{\boldsymbol{\beta}}} \left\{ |\tilde{\beta}_j - \beta_j| \geq t_n/2 \right\}\right) \\
&= 1 - O\left[\exp\left(-C_8 n^{1-2\alpha_1-2\alpha_3-2\eta}\right) + (1/p)^{\delta-2}(1/n)^{(\frac{1-q}{2}-\alpha_2-\alpha_3-\eta)/(1-q)}\right],
\end{aligned}
\tag{12.11}
$$

since under (C5), by the choice of t_n, $\min_{j \in \mathcal{M}_{\boldsymbol{\beta}}} |\beta_j| > \frac{3}{2} t_n$ for large enough n.

On the other hand,

$$
\begin{aligned}
& P(\widehat{\mathcal{M}}_{\boldsymbol{\beta}} \subset \mathcal{M}_{\boldsymbol{\beta}}) \\
&= 1 - P\left(\cup_{j \notin \mathcal{M}_{\boldsymbol{\beta}}} \left\{ |\tilde{\beta}_j| > t_n \right\}\right) \\
&= 1 - P\left(\cup_{j \notin \mathcal{M}_{\boldsymbol{\beta}}} \left\{ |\tilde{\beta}_j - \beta_j| > t_n \right\}\right) \\
&= 1 - O\left[\exp\left(-C_8 n^{1-2\alpha_1-2\alpha_3-2\eta}\right) + (1/p)^{\delta-2}(1/n)^{(\frac{1-q}{2}-\alpha_2-\alpha_3-\eta)/(1-q)}\right].
\end{aligned}
\tag{12.12}
$$

(12.11) and (12.12) together prove the theorem. □

Proof of Lemma 9.3. The proof is analogous to that of Lemma 9.1. Under condition (C1′), it holds that

$$
\mathrm{E}\,|x_{ij}x_{ik} - \mathrm{E}(x_j x_k)|^{2l} \leq 2^{2l-1}\left[\mathrm{E}|x_{ij}x_{ik}|^{2l} + (\mathrm{E}|x_{ij}x_{ik}|)^{2l}\right] = O(1).
$$

Then, by Chebyshev Inequality and Theorem 2 in Whittle [34].

$$\begin{aligned}
&\max_{jk} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \mathrm{E}(x_j x_k)\right| > t\right) \\
&\leq t^{-2l}\mathrm{E}\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \mathrm{E}(x_j x_k)\right|^{2l} \\
&= O(t^{-2l}n^{-l}).
\end{aligned}$$

Therefore, by replacing (12.2) with

$$\begin{aligned}
&P\left(\sum_{k=1}^{p}\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \mathrm{E}(x_j x_k)\right| |\beta_k| > \frac{t}{3}\right) \\
&\leq p \cdot \max_{jk} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik} - \mathrm{E}(x_j x_k)\right| > \frac{t}{3M^{1-h}s_h}\right) \\
&\leq C_9 p s_h^{2l} t^{-2l} n^{-l},
\end{aligned}$$

for some $C_9 > 0$ and replacing (12.3) and (12.4) with

$$\begin{aligned}
\max_{1\leq j\leq p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \sigma_i x_{ij}\epsilon_i\right| > \frac{t}{3}\right) &\leq \frac{C_{10}}{2} t^{-2l} n^{-l}, \\
\max_{1\leq j\leq p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mu x_{ij}\right| > \frac{t}{3}\right) &\leq \frac{C_{10}}{2} t^{-2l} n^{-l},
\end{aligned}$$

for some $C_{10} > 0$, the rest of proof follows from (12.1). □

Proof of Lemma 9.4. From Lemma 9.3, it holds that

$$P\left(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > \frac{t}{6v_p}\right) \leq C_{11} p^2 s_h^{2l} v_p^{2l} t^{-2l} n^{-l} + C_{12} p v_p^{2l} t^{-2l} n^{-l}, \tag{12.13}$$

for some $C_{11} > 0$ and $C_{12} > 0$.

Under condition (C3′), it follows from Theorem 1(ii) of Cai and Liu [9] that, there exists $C_{13} > 0$ that

$$P\left(\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_1 > \frac{t}{6Mv_p}\right) \leq C_{13}\left(n^{-1/2}p^{-(\delta-2)}(r_q v_p/t)^{1/(1-q)} + n^{-\frac{l-1-\tau}{2}}\right). \quad (12.14)$$

Replacing (12.9) and (12.10) with (12.13) and (12.14) and observing that by the choice of λ_n, $P(\lambda_n \boldsymbol{Z} > \frac{t}{6v_p}) = 0$, when n is sufficiently large, the rest of the proof resembles the proof of Lemma 9.2. □

Proof of Theorem 9.2. The proof is the same as the proof of Theorem 9.1 by replacing results in Lemma 9.2 with results in Lemma 9.4. □

Proof of Lemma 9.5. From (12.6),

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\beta}}_M - \lambda_n \hat{\boldsymbol{\Omega}}\boldsymbol{Z}.$$

Recall that, $\boldsymbol{\beta} = \boldsymbol{\Omega}\boldsymbol{\beta}_M$. Hence,

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\beta}}_M - \boldsymbol{\Omega}\boldsymbol{\beta}_M - \lambda_n \hat{\boldsymbol{\Omega}}\boldsymbol{Z}.$$

Then,

$$P(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > t) \leq P(\|\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\beta}}_M - \boldsymbol{\Omega}\boldsymbol{\beta}_M\|_\infty > t/2) + P(\|\lambda_n \hat{\boldsymbol{\Omega}}\boldsymbol{Z}\|_\infty > t/2). \quad (12.15)$$

Since

$$\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\beta}}_M - \boldsymbol{\Omega}\boldsymbol{\beta}_M = (\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\hat{\boldsymbol{\beta}}_M + \boldsymbol{\Omega}(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M),$$

it holds that

$$P(\|\hat{\boldsymbol{\Omega}}\hat{\boldsymbol{\beta}}_M - \boldsymbol{\Omega}\boldsymbol{\beta}_M\|_\infty > t/2) \leq P(\|(\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\hat{\boldsymbol{\beta}}_M\|_\infty > t/4) + P(\|\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M)\|_\infty > t/4). \tag{12.16}$$

The first item in (12.16) is bounded by

$$\begin{aligned}
&P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 \|\hat{\boldsymbol{\beta}}_M\|_\infty > t/4\right) \\
&\leq P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 \|\hat{\boldsymbol{\beta}}_M\|_\infty > t/4 \cap \|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty \leq t/8\right) + P(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t/8) \\
&\leq P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 \|\boldsymbol{\beta}_M\|_\infty > t/8 \cap \|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty \leq t/8\right) \\
&\quad + P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 \|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t/8 \cap \|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty \leq t/8\right) \\
&\quad + P\left(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t/8\right) \\
&\leq P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 > t/[8M^{1-h}s_h]\right) + P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 > 1\right) + P\left(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t/8\right).
\end{aligned}$$

For the second item in (12.16), it follows from the assumption $\|\boldsymbol{\Omega}\|_1 \leq M$ that

$$\begin{aligned}
P(\|\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M)\|_\infty > t/4) &\leq P\left(\|\boldsymbol{\Omega}\|_1 \|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t/4\right) \\
&\leq P\left(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t/[4M]\right).
\end{aligned}$$

Without loss of generality, assume $M \geq 2$. Then, for any $0 < t < 8M^{1-h}s_h$,

$$\begin{aligned} P\left(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\infty > t\right) \leq\ & 2P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 > t/[8M^{1-h}s_h]\right) \\ & + 2P\left(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t/[4M]\right), \end{aligned} \tag{12.17}$$

since $\|\lambda_n \hat{\boldsymbol{\Omega}}\boldsymbol{Z}\|_\infty \leq |\lambda_n| \|\hat{\boldsymbol{\Omega}}\|_1 \|\boldsymbol{Z}\|_\infty \leq |\lambda_n| \|\hat{\boldsymbol{\Omega}}\|_1 \leq 2|\lambda_n| \|\boldsymbol{\Omega}\|_1 \leq 2M|\lambda_n|$. By the choice of λ_n, when n is sufficiently large, $P(\|\lambda_n \hat{\boldsymbol{\Omega}}\boldsymbol{Z}\|_\infty > t/2) = 0$.

Under (C1), it follows by Theorem 1(a) of Cai, Liu, and Luo [7] that

$$P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 > t/[8M^{1-h}s_h]\right) \leq 4\exp\left(-C_{14}n[t/s_h r_q]^{2/(1-q)}\right), \tag{12.18}$$

for some $C_{14} > 0$.

From Lemma 9.1, it holds that

$$P(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t/(4M)) \leq 2p^2 \exp\left(-C_{15}nt^2/s_h^2\right) + 4p\exp\left(-C_{16}nt^2\right). \tag{12.19}$$

(12.17), (12.18) together with (12.19) proves the lemma. □

Proof of Lemma 9.6. Under conditions (C1′) and (C3‴), by Theorem 1(ii) of Cai, Liu, and Luo [7].

$$P\left(\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}\|_1 > \frac{t}{8M^{1-h}s_h}\right) \leq \frac{C_{21}}{2}\left[\exp\left(-C_{22}n\left[t/s_h r_q\right]^{\frac{2}{1-q}}\right) + n^{-\frac{l-1-\tau}{2}}\right].$$

From Lemma 9.3, it holds that

$$P(\|\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_M\|_\infty > t/(4M)) \leq \frac{C_{23}}{2} p^2 s_h^{2l} t^{-2l} n^{-l} + \frac{C_{24}}{2} p t^{-2l} n^{-l}.$$

The rest of proof follows by (12.17). □

Proof of Theorem 9.3 and Theorem 9.4. By using results in Lemma 9.5 and Lemma 9.6, the proof is the same as that of Theorem 9.1. □

Part IV

Conclusion

Recent boom of high dimensional data in various disciplines necessitates the development of new statistical tools. To this end, I have worked on two aspects of high dimensional data inference: (a) classification of high dimensional data; (b) variable selection in high dimensional linear regression model. Both of them could have important applications in real world, e.g. the applications in cancer studies as seen in Section 5.2.

For the high dimensional classification problem, I extend the classical Quadratic Discriminant Analysis (QDA) to accommodate high dimensional data. The most significant change is to use sparse estimators in the discriminant function, instead of Maximum Likelihood Estimators. Theoretical properties of my proposed method is thoroughly studied. It has been shown that the new method is Fisher consistent to the optimal Bayes rule, namely the misclassification rate of the new method asymptotically converges to that of the optimal Bayes rule. A by-product of such analysis is the establishment of Fisher consistency of classical QDA when the dimension of data tends to infinity, but in a much slower rate than the sample size. Such theoretical results are also new to the literature.

Moreover, I have compared the new method (SQDA) with the Sparse LDA (SLDA) method proposed by Shao, Wang, Deng, and Wang [30]. I have shown that when the difference between the two classes' covariance matrices is big enough, then SQDA could outperform SLDA. An open problem is to find the how big the difference needs to be in order for SQDA to outperform SLDA. This is an interesting topic for my future research.

For the problem of variable selection of high dimensional linear regression, I develop a Regularized LASSO(RLASSO) method, which applies two additional

regularizations to the well-known LASSO method. The first regularization is to replace sample covariance with a sparse estimator of covariance and the second regularization is to add another thresholding step after solving the LASSO type problem. The main contribution of the first regularization is to stabilize the item related to sample covariance in the optimization problem of LASSO. The main contribution of the second regularization is to remove the small estimated coefficients left over by the solution to LASSO problem. With the aid the these two new regularizations, RLASSO relaxes the strong conditions that LASSO needs to have good variable selection performance. It has shown that as long as the regression coefficients and the covariance of covariates (or its inverse) are sparse, RLASSO can asymptotically choose the correct model.

RLASSO also reveals the connection between LASSO and some other popular variable selection methods, such as Sure Independence Screening and scout method. Actually, I have shown that all of them can be regarded as special cases of RLASSO, which makes RLASSO a very general method. Several simulation studies have been performed to compare RLASSO with other variable selection methods. The simulation results have shown that RLASSO in general performs much better than those methods.

At this stage, RLASSO only applies to linear regression model. It will be an interesting task to extend it to high dimensional generalized linear regression model, such as logistic regression, etc. This could be a topic for my future research.

BIBLIOGRAPHY

[1] Alon, U., N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. 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.

[2] Anderson, Theodore Wilbur. 2003. *An introduction to multivariate statistical analysis*, vol. 2. Wiley Series in Probability and Statistics.

[3] Bickel, Peter J., and Elizaveta Levina. 2008. Covariance regularization by thresholding. *The Annals of Statistics* 36(6):2577–2604.

[4] Bickel, P.J., and E. Levina. 2004. Some theory for fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 989–1010.

[5] ———. 2008. Covariance regularization by thresholding. *The Annals of Statistics* 36(6):2577–2604.

[6] Bühlmann, P., and S. Van De Geer. 2011. *Statistics for high-dimensional data: Methods, theory and applications*. Springer-Verlag New York Inc.

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

[8] Cai, T. Tony, Cun-Hui Zhang, and Harrison H. Zhou. 2010. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics* 38(4): 2118–2144.

[9] Cai, Tony, and Weidong Liu. 2011. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* 106(494): 672–684.

[10] ———. 2011. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association* 106(496):1566–1577.

[11] Cheng, Y. 2004. Asymptotic probabilities of misclassification of two discriminant functions in cases of high dimensional data. *Statistics & probability letters* 67(1): 9–17.

[12] Clemmensen, L.K.H., T. Hastie, and B.K. Ersbøll. 2008. Sparse discriminant analysis. Tech. Rep., DTU Informatics Lyngby.

[13] Dettling, M. 2004. Bagboosting for tumor classification with gene expression data. *Bioinformatics* 20(18):3583–3593.

[14] Fan, J., and Y. Fan. 2008. High dimensional classification using features annealed independence rules. *Annals of statistics* 36(6):2605–2637.

[15] Fan, J., and R. Song. 2010. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics* 38(6):3567–3604.

[16] Fan, Jianqing, Yang Feng, and Xin Tong. 2012. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

[17] Fan, Jianqing, and Jinchi Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5):849–911.

[18] ———. 2010. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20(1):101–148.

[19] Fan, Jianqing, and Heng Peng. 2004. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* 32(3):928–961.

[20] Fisher, Ronald A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics* 7(2):179–188.

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

[22] Guo, Y., T. Hastie, and R. Tibshirani. 2007. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* 8(1):86–100.

[23] Hunter, David R., and Runze Li. 2005. Variable selection using MM algorithms. *Annals of statistics* 33(4):1617–1642.

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

[25] Meinshausen, Nicolai, and Bin Yu. 2009. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* 37(1):246–270.

[26] Petrov, V.V. 1995. *Limit theorems of probability theory.* Oxford Science Publications.

[27] Qiao, Z., L. Zhou, and J.Z. Huang. 2009. Sparse linear discriminant analysis with applications to high dimensional low sample size data. *IAENG International Journal of Applied Mathematics* 39(1):1–13.

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

[29] Saulis, L., and VA Statulevicius. 1991. *Limit theorems for large deviations*, vol. 73. Springer.

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

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

[32] Wang, Hansheng. 2009. Forward regression for Ultra-High dimensional variable screening. *Journal of the American Statistical Association* 104(488):1512–1524.

[33] Wang, Hansheng, Runze Li, and Chih-Ling Tsai. 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94(3): 553–568.

[34] Whittle, P. 1960. Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and its Applications* 5(3):302.

[35] Witten, Daniela M, and Robert Tibshirani. 2009. Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(3):615–636.

[36] Zhang, Cun-Hui, and Jian Huang. 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* 36(4): 1567–1594.

[37] Zhang, Q., and H. Wang. 2011. On bic's selection consistency for discriminant analysis. *Statistica Sinica* 21:731–740.

[38] Zhao, Peng, and Bin Yu. 2006. On model selection consistency of lasso. *J. Mach. Learn. Res.* 7:2541–2563.

[39] Zou, Hui. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101(476):1418–1429.