Variable Selection in High-Dimensional Classification

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Qing Mai

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Hui Zou, Advisor

June, 2013

Acknowledgements

When I look back at the years I spent as a Ph.D student at University of Minnesota, my heart is filled with happy memories and sincere gratitude. The journey has turned out more amazing than I could ever imagined, which could not have happened without all the great people I have met.

I am very grateful to my advisor, Dr. Hui Zou. His patience and encouragement and vision has greatly helped me through the doctoral program. His vision and knowledge has enlightened me in my four years working with him. He will always be my role model. My gratitude also goes to Dr. Birgit Grund, Dr. Douglas M. Hawkins, Dr. Glen Meeden and Dr. Wei Pan for being my dissertation committee. Their questions and suggestions have given me deeper understanding of many statistical problems.

I also would like to thank many faculty and staff members. Thank Dr. Galin Jones and Dr. Yuhong Yang for writing recommendation letters for my job application. Thank Dr. Peihua Qiu for his generous help. Thank Dr. Bradley P. Carlin, Dr. R. Dennis Cook, Dr. Charles Geyer, Dr. Tiefeng Jiang, Dr. Nicolai V. Krylov, Dr. Lan Wang, Dr. Sanford Weisberg for their wonderful classes. Thank Jane Sell and Megan Schlick for sending my recommendation letters.

Many friends also deserve my great gratitude. Thank Xin Zhang for being with me through the good and bad times. Special thanks go to Xin Chen and Zhihua Su for selflessly sharing their professional experience, and Yi Yang for all the technical support. And I sincerely thank all my friends who have helped me, and brought fun into my life. Although I cannot fit all their names into this one-page acknowledgement, I cherish everyday I spent with them.

The last but not the least, I would like to thank my parents for bringing me up, and caring about me every second.

Dedication

To the memory of my grandfather, Yunong Mai.

Abstract

Classification has long been an important research topic for statisticians. Nowadays, scientists are further challenged by classification problems for high-dimensional datasets in various fields, ranging from genomics, economics to machine learning. For such massive datasets, classical classification techniques may be inefficient or even infeasible, while new techniques are highly sought-after.

My dissertation work tackles high-dimensional classification problems by utilizing variable selection. In particular, three methods are proposed and studied: direct sparse discriminant analysis, semiparametric sparse discriminant analysis and the Kolmogorov filter.

In the proposal of direct sparse discriminant analysis (DSDA), I first point out the disadvantage in many current methods that they ignore the correlation structure between predictors. Then DSDA is proposed to extend the well-known linear discriminant analysis to high dimensions, fully respecting the correlation structure. The proposal is efficient and consistent, with excellent numerical performance. In addition to the proposal of DSDA, I also study its connection to many popular proposals of linear discriminant analysis in high dimensions, including the ℓ_1-Fisher's discriminant analysis and the sparse optimal scoring.

Semiparametric sparse discriminant analysis (SeSDA) extends DSDA by relaxing the normality assumption, which is fundamental for any method requiring the linear discriminant analysis model. SeSDA is more robust than DSDA, while it preserves the good properties of DSDA. Along with the development of SeSDA, a new concentration inequality is obtained that can provide theoretical justifications for methods based on Gaussian copulas.

Moreover, the Kolmogorov filter is proposed as a fully nonparametric method that performs variable selection for high-dimensional classification. It requires minimal assumptions on the distribution of the predictors, and is supported by both theoretical and numerical examples.

Also, some potential future work is discussed on variable selection in classification.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Modern scientists frequently encounter classification problems for high-dimensional datasets in fields such as genomics, economics and machine learning. For example, in cancer diagnosis, it is of vital importance to accurately determine whether a person is a patient or not. Now that biologists are able to measure the expression levels of tens of thousands of genes at a low cost for this purpose, a critical problem remains how to fit an efficient classification rule with observations on some many predictors. Another such example is the identification of spam emails. Automatic spam filters usually require the frequencies of a long list of words as predictors to classify an email. Therefore, high-dimensional classifiers are in need.

However, it is very hard to estimate accurate classifiers when many predictors are present for three reasons in general. First, it may be impossible to fit a classifier when the number of predictors exceeds the number of samples. One such example is the linear discriminant analysis. Second, even when fitting is possible, the classifier may be no better than random guessing [1], because of the noise accumulation. When too many predictors and hence parameters are present, the noise in estimating the parameters will ultimately swamp the signal, and the classification accuracy will deteriorate to 1/2. Third, the computation cost may be extremely high, such as for random forest [2].

Variable selection is a popular technique that has been employed to handle the impact of high dimensionality in regression [3, 4, 5]. It relies on the sparsity assumption

that, among the large number of predictors, only a few are useful. Then variable selection identifies these predictors and the final model only includes them. Because only a few predictors are involved, variable selection usually leads to efficient and accurate model estimation. Some examples of well-known variable selection techniques for regression are Lasso [3], the smoothly clipped absolute deviation (SCAD) penalty [4] and SURE independence screening [6], among others. These methods can effective reduce the dimension and achieve high prediction accuracy.

Unfortunately, only very preliminary results exist on how variable selection should be performed in classification. This motivates my thesis work. I will propose three methods for variable selection in high-dimensional classification: direct sparse discriminant analysis (DSDA) [7], semiparametric sparse discriminant analysis (SeSDA) and the Kolmogorov filter [8]. Also, I am going to point out some connection between the work of mine and other researchers' [9].

1.2 Organization of the Dissertation

The rest of this dissertation is as follows. The rest of Chapter 1 will rigorously define the classification problem and briefly review variable selection techniques in regression. Chapter 2 introduces DSDA, which generalizes the classical linear discriminant analysis (LDA) to high dimensions. This chapter starts with revealing a fundamental drawback of some early attempts in extending LDA that these methods ignore the correlations between predictors. Then DSDA is proposed as a method that takes the correlations into full consideration. More importantly, since it recasts LDA as a regression problem, DSDA is able to easily incorporate the variable selection techniques and can be efficiently solved. Also, DSDA is consistent in theory and outperforms its competitors in numerical examples.

Chapter 3 studies sparse LDA methods proposed by other researchers. This chapter begins with a brief review of some popular proposals of sparse LDA. Then it is shown that DSDA is deeply connected to two of them, the ℓ_1-Fisher's discriminant analysis (SFDA) and sparse optimal scoring (SOS). With properly chosen tuning parameters, DSDA can give the same discriminant direction as the other two.

In Chapter 4, semiparametric sparse discriminant analysis (SeSDA) is proposed as

a relaxation of DSDA. DSDA or any methods derived from LDA requires the normality assumption to be theoretically justified. However, this assumption is usually too stringent in practice. When this assumption is severely violated, both the variable selection and the final classification will be unreliable. SeSDA can be an alternative when this is the case. Under weaker assumptions, SeSDA preserves the efficient computation, excellent performance and theoretical consistency of DSDA.

The Kolmogorov filter in Chapter 5 is another variable selection method for high-dimensional classification problems. It belongs to the fast-developing field of screening techniques; yet it is distinguished from most of other screening methods by its fully nonparametric nature. With almost no model assumptions, it is able to consistently select all the useful predictors.

Finally, Chapter 6 discusses some potential future work.

1.3 Classification Problems

Now we mathematically define the classification problem and variable selection.

Classification problems are concerned with a pair of random variables $(Y, \mathbf{X})$, where Y is the class label and $\mathbf{X}$ is a p-dimensional predictor. A large part of this thesis considers the binary case, where Y is coded as $\{+1, -1\}$. The performance of a classifier $\delta(\mathbf{X})$ is measured by the misclassification rate $R(\delta(\mathbf{X})) = \Pr(Y \neq \delta(\mathbf{X}))$. Therefore, we hope to estimate the classification rule with the lowest misclassification rate, or, in other words, the Bayes rule

$$\delta^{\text{Bayes}}(\mathbf{X}) = \text{sign}\left(\log(\frac{\pi_+}{\pi_-}) + \log\frac{f_+(\mathbf{X})}{f_-(\mathbf{X})}\right) \tag{1.1}$$

where $\pi_y = \Pr(Y = y)$ and $f_y(\mathbf{X})$ is the density function of $\mathbf{X}$ conditional on $Y = y$. In practice, $\delta^{\text{Bayes}}(\mathbf{X})$ is estimated through the samples $(Y^i, \mathbf{X}^i)_{i=1}^n$, where n is the sample size. However, fully nonparametric model fitting is generally difficult, and hence researchers propose various assumptions along with classification techniques, such as linear discriminant analysis, logistic regression, boosting [10, 11], random forest [2] and support vector machine [12]; see [13]. Such methods are usually powerful when the sample size n is large comparing to the dimension p.

When p is large, however, traditional methods are usually incompetent. Instead,

[1] proved that variable selection is critical. To perform variable selection, we assume that there exists a set $\mathbf{D} \subset \{1, \ldots, p\}$ such that the cardinality of $\mathbf{D}$, $|\mathbf{D}| \ll p$ and $\delta^{\text{Bayes}}(\mathbf{X}) = \delta^{\text{Bayes}}(\mathbf{X}_{\mathbf{D}})$. This assumption is commonly referred to as the sparsity assumption. Then variable selection then aims to detect the set $\mathbf{D}$. It is in general unclear how to achieve this task in the classification problem, while many methods exist for regression problems.

1.4 Variable Selection in Regression

Variable selection techniques have been studied intensively for regression problems. Consider the linear regression model,

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{1.2}$$

where $Y \in \mathbb{R}$, $\epsilon \sim N(0, \sigma^2)$ and $\epsilon \perp \mathbf{X}$. Under this model, the maximum likelihood estimator is the so-called least squares estimator:

$$\hat{\boldsymbol{\beta}}^{ols} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \|Y^i - \mathbf{X}^i\boldsymbol{\beta}\|_2^2. \tag{1.3}$$

Traditional variable selection techniques for this model include the Akaike information criterion (AIC) [14], Bayesian information criterion (BIC) [15], C_p statistic [16], cross validation and generalized cross validation. These methods compares all the subsets of $\{1, \ldots, p\}$ and choose the subset that minimizes a particular criterion. For example, AIC picks the set $\hat{\mathbf{D}}$ such that

$$\hat{\mathbf{D}} = \arg\min_{\mathbf{D}} n \log \frac{\sum_i (Y^i - \mathbf{X}_{\mathbf{D}}\hat{\boldsymbol{\beta}}_{\mathbf{D}})^2}{n} + 2\|\boldsymbol{\beta}_{\mathbf{D}}\|_0 \tag{1.4}$$

However, these methods are computationally infeasible for high-dimensional datasets because they have to evaluate 2^p subsets. To resolve this issue, researchers have developed penalized likelihood methods and screening methods as computationally efficient variable selection techniques.

1.5 Penalized Likelihood Methods

Penalized likelihood methods estimate $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \|Y^i - \beta_0 - \mathbf{X}^i\boldsymbol{\beta}\|_2^2 + \sum_{i=1}^{p} P_\lambda(|\beta_j|), \tag{1.5}$$

where $P_\lambda(|\beta_j|)$ is a penalty function depending on the tuning parameter λ. The first term in (1.5) quantifies how well the model fits into data, and is usually smaller with more predictors included in the model. The second term, however, increases as more predictors enter the model. For example, the well-known Lasso penalty [3] sets $P_\lambda(t) = \lambda t$. Therefore, the penalty function prohibits including too many predictors. By adding up these two terms, we strike a balance between bias and variance and obtain a sparse estimate $\hat{\boldsymbol{\beta}}$ with many zero entries. Numerous papers have been devoted into studying the penalty function. Some successful candidates for $P_\lambda(\cdot)$ are Lasso [3], the smoothly clipped absolute deviation (SCAD) penalty [4], the elastic net penalty [17], adaptive lasso [18], the minimax concave penalty (MCP) [19] and group Lasso [20], among others. Now we review these methods individually. We start with Lasso and then proceed to one of its famous modifications, the adaptive Lasso. Two nonconvex penalties, SCAD and MCP, are then discussed. Finally, we talk about elastic net and group Lasso as two important examples when the predictors have some special structures.

1.5.1 The Lasso Penalty

Lasso is one of the most widely used penalty functions. It replaces the ℓ_0 norm in (1.4) with ℓ_1 norm. Then $\boldsymbol{\beta}$ is estimated by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \|Y^i - \beta_0 - \mathbf{X}^i\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{1.6}$$

Lasso can also be written as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \|Y^i - \beta_0 - \mathbf{X}^i\boldsymbol{\beta}\|_2^2, \text{ s.t. } \|\boldsymbol{\beta}\|_1 \leq \tau \tag{1.7}$$

Lasso yields sparse estimates of $\boldsymbol{\beta}$ because it shrinks all $\hat{\beta}_j^{ols}$'s toward zero. Therefore, if $\hat{\beta}_j^{ols}$ given by lease squares is close to zero, Lasso will estimate it with exactly zero.

Lasso is closely related to the nonnegative garrote [21] and the bridge regression [22]. In particular, the bridge regression defines $P_\lambda(t) = \lambda t^q$, and Lasso is an important special case of the bridge regression with $q = 1$. This choice enables variable selection and results in a convex problem. To solve this problem, many efficient algorithms have been proposed. For example, [23] proposed the shooting algorithm, while [24] rewrote Lasso as a quadratic programming problem. [25] took note of that the solution path of Lasso is piecewise linear and invented the efficient algorithm of `LARS`, implemented in the `R` package `lars`. Finally, [26] developed the extremely fast coordinate descent algorithm, as implemented in the `R` package `glmnet`.

Many papers discuss the theoretical properties of Lasso. An incomplete list includes [27, 28, 29, 30, 31, 32, 33]. In general, these works showed that Lasso identifies the important variables with high probabilities. In order to achieve such results, the irrepresentable condition is usually required, that

$$\|(\mathbf{X}_\mathbf{D}^{\mathrm{T}}\mathbf{X}_\mathbf{D})^{-1}\mathbf{X}_\mathbf{D}^{\mathrm{T}}\mathbf{X}_{\mathbf{D}^C}\|_\infty \leq 1 - \alpha, \text{ for some } 0 < \alpha < 1. \tag{1.8}$$

Efforts have been devoted to relaxing this irrepresentable condition. One well-known example is the adaptive Lasso.

1.5.2 The Adaptive Lasso

The adaptive Lasso [18] is an elegant modification of Lasso. Instead of penalizing each coefficient equally, it assigns a different weight to each X_j, estimating $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i (Y^i - \beta_0 - \mathbf{X}^i\boldsymbol{\beta})^2 + \lambda \sum_j \omega_j |\beta_j|. \tag{1.9}$$

In particular, [18] proposed to set $\omega_j = \dfrac{1}{|\hat{\beta}_j^0|^\gamma}$ with $\gamma > 0$, where $\hat{\boldsymbol{\beta}}^0$ is a $\sqrt{n}$ consistent estimate of $\boldsymbol{\beta}$, such as $\hat{\boldsymbol{\beta}}^{ols}$. It is easy to see that, if $\hat{\beta}_j^0$ is close to zero, then β_j is penalized heavily, while β_j is only slightly penalized if $\hat{\beta}_j^0$ is large. This modification enables the adaptive lasso to enjoy the oracle property even when the irreperesentable condition in (1.8) does not hold. Also, the adaptive Lasso is still convex as Lasso, so its solution is also very efficient.

1.5.3 The Smoothly Clipped Absolute Deviation Penalty and the Minimax Concave Penalty

The smoothly clipped absolute deviation (SCAD) penalty [4] is proposed to eliminate the biasedness of Lasso. Note that Lasso shrinks all $\hat{\beta}_j^{ols}$'s toward zero. Therefore, the large elements in $\boldsymbol{\beta}$ tend to be underestimated. On the other hand, SCAD does not have this drawback. The derivative of the SCAD penalty is

$$P'_\lambda(t) = \lambda\{I(t < \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\} \text{ for some } a > 2. \tag{1.10}$$

[4] suggested $a = 3.7$ in practice. Because $P'_\lambda(t) = 0$ for large t, SCAD gives asymptotically unbiased estimators. In fact, [4] showed that the SCAD penalty enjoys the oracle property that, with probability tending to 1,

1. $\hat{\boldsymbol{\beta}}_{\mathbf{D}^C} = 0$;
2. $\hat{\boldsymbol{\beta}}_{\mathbf{D}}$ is $\sqrt{n}$ consistent.

As the adaptive Lasso, SCAD does not require the irrepresentable condition to be consistent. Another such penalty is the minimax concave penalty with derivative

$$P'_\lambda(t) = 1 - \frac{t}{a\lambda}, \text{ where } a > 0. \tag{1.11}$$

Although SCAD is not convex, algorithms are available to compute its solution. [4] locally approximates the SCAD penalty with quadratic functions, while [34] employs the local approximation with linear functions. After the approximation, the problem becomes convex and can be easily solved.

1.5.4 The Elastic Net Penalty

The elastic net penalty [17] is proposed to improve the performance of Lasso when predictors are highly correlated. It connects the Lasso penalty and the ridge penalty

$$P_\lambda(t) = \lambda_1 t + \lambda_2 t^2 \tag{1.12}$$

Or, equivalently, regression with the elastic net penalty can be written as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i (Y^i - \beta_0 - \mathbf{X}^i\boldsymbol{\beta})^2 \text{ s.t. } \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2 \leq \tau. \tag{1.13}$$

With the constraint on $\|\boldsymbol{\beta}\|_1$, $\hat{\boldsymbol{\beta}}$ is typically sparse, while the constraint on $\|\boldsymbol{\beta}\|_2^2$ shrinks $\hat{\beta}_j$'s toward each other when X_j's are highly correlated. Hence, the behavior of elastic net is different from Lasso when strong correlations are present. Lasso picks one X_j, while elastic net preserves all the X_j's that are highly correlated. This property is usually known as the grouping effect.

1.5.5 The Group Lasso

The group Lasso is especially suitable when there are predefined groups among the predictors. In such cases, practitioners often hope that a group of predictors should either be all included or excluded. This is the motivation of the group Lasso. Suppose that there are G groups and $\boldsymbol{\beta}_g$ denotes the coefficients of X_j's in gth group. Then the group Lasso solves for

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i (Y^i - \beta_0 - \mathbf{X}^i\boldsymbol{\beta})^2 + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2. \tag{1.14}$$

With properly chosen λ, the group Lasso may estimate all the elements in some $\boldsymbol{\beta}_g$'s as zero.

The group Lasso is not only useful in the linear regression context. For example, [35] applies it to fit sparse additive models.

1.6 Screening Methods

Screening methods is another family of variable screening methods that are extremely efficient in computation. Instead of modeling $\mathbf{X}$ together as in penalized likelihood methods, screening methods individually ranks the importance of each variable X_j by some measure m_j. The the X_j's are selected with m_j among the d_n's largest, where d_n is a predefined integer. Some examples of screening methods are the marginal correlation screening (MCS) [36], rank correlation screening (RCS) [37] and distance correlation screening (DCS) [38]. The major difference between these three methods is that they employ different m_j's. Now we briefly review these methods.

MCS is developed under the linear regression model in (1.1). [36] proposed to rank the importance of X_j by the absolute value of the Pearson correlation $r_j = \text{cor}(X_j, Y)$.

Then MCS selects $\hat{\mathbf{D}} = \{j : |r_j| \text{ is amongst the } d_n\text{-th largest.}\}$. In practice, d_n could be set to $\frac{n}{\log n}$ or n. RCS, on the other hand, uses the Kendall τ correlation

$$\tau_j = \frac{1}{n(n-1)} \sum_{i \neq k}^{n} 1(X_j^i < X_j^k) 1(Y^i < Y^k) - \frac{1}{4}.$$

Since the Kendall τ correlation is more robust than the the Pearson correlation, RCS is more robust than MCS. In particular, it can deal with the transformation linear regression model:

$$h(Y) = \mathbf{X}^{\mathrm{T}} \boldsymbol{\beta} + \epsilon,$$

where h is an unknown function.

DCS further improves the robustness of RCS, because it utilizes the distance correlation [39] to perform screening. The distance correlation is defined as follows. For two random vectors $\mathbf{u} \in \mathbb{R}^{d_\mathbf{u}}, \mathbf{v} \in \mathbb{R}^{d_\mathbf{v}}$, define

$$\mathrm{dcov}^2(\mathbf{u}, \mathbf{v}) = \int_{\mathbb{R}^{d_\mathbf{u}+d_\mathbf{v}}} \|\phi_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s}) - \phi_\mathbf{u}(\mathbf{t})\phi_\mathbf{v}(\mathbf{s})\|^2 w(\mathbf{t}, \mathbf{s}) \, \mathrm{d}\,\mathbf{t} \, \mathrm{d}\,\mathbf{s},$$

where $\phi_{\mathbf{u},\mathbf{v}}$ is the joint characteristic function of $\mathbf{u}$, $\mathbf{v}$, $\phi_\mathbf{u}$ and $\phi_\mathbf{v}$ are the characteristic functions of $\mathbf{u}$ and $\mathbf{v}$, respectively, and $w(\mathbf{t}, \mathbf{s}) = \{c_{d_\mathbf{u}} c_{d_\mathbf{v}} \|\mathbf{t}\|_{d_\mathbf{u}}^{1+d_\mathbf{u}} \|\mathbf{s}\|_{d_\mathbf{v}}^{1+d_\mathbf{v}}\}^{-1}$ with $c_d = \frac{\pi^{(1+d)/2}}{\Gamma\{(1+d)/2\}}$. Then the distance correlation between $\mathbf{u}$ and $\mathbf{v}$ is

$$\mathrm{dcorr}(\mathbf{u}, \mathbf{v}) = \frac{\mathrm{dcov}(\mathbf{u}, \mathbf{v})}{\sqrt{\mathrm{dcov}(\mathbf{u}) \, \mathrm{dcov}(\mathbf{v})}}.$$

It can be shown that $\mathrm{dcorr}(\mathbf{u}, \mathbf{v}) = 0$ if and only if $\mathbf{u} \perp \mathbf{v}$. Moreover, when $u, v \in \mathbb{R}$ are jointly normal, it can be shown that $\mathrm{dcorr}(u, v)$ is a monotone function in the correlation between u and v. Therefore, $\mathrm{dcorr}(Y, X_j)$ can be a measurement of the dependence. Indeed, DCS ranks X_j by $\mathrm{dcorr}(Y, X_j)$ and estimates $\mathbf{D}$ by $\hat{\mathbf{D}} = \{j : |\,\mathrm{dcorr}(Y, X_j)| \text{ is amongst the } d_n\text{-th largest.}\}$ DCS is applicable without any model specification.

Note that all screening methods require $d_n \ll p$, while it is not necessary that $d_n = d$. This is because screening methods usually serve as a crude first-stage analysis. When the dimension decreases to d_n, one could apply more refined methods to $(Y, \mathbf{X}_{\hat{\mathbf{D}}})$. For example, [36] suggested applying SCAD or the Dantzig selector [5] after MCS. Therefore,

an especially desirable property for screening methods is the SURE screening property that $\Pr(\hat{\mathbf{D}} \subset \mathbf{D}) \to 1$. Under suitable conditions, MCS, RCS and DCS all enjoy this property.

Chapter 2

Sparse Discriminant Analysis

2.1 Chapter Overview

Sparse discriminant methods based on independence rules, such as the nearest shrunken centroids classifier (Tibshirani et al. 2002) and features annealed independent rules (Fan & Fan, 2008), have been proposed as computationally attractive tools for feature selection and classification with high-dimensional data. A fundamental drawback of these rules is that they ignore correlations among features and thus could produce misleading feature selections and inferior classifications. We propose a new recipe for sparse discriminant analysis, motivated by least squares formulation of linear discriminant analysis. To demonstrate our proposal, we study the numerical and theoretical properties of discriminant analysis constructed via Lasso/SCAD penalized least squares. Our theory shows that both the proposed methods can consistently identify the subset of discriminative features contributing to the Bayes rule and at the same time consistently estimate the Bayes classification direction, even when the dimension can grow faster than any polynomial order of the sample size. The theory allows for general dependence among features. Simulated and real data examples show that our methods compare favorably with other popular sparse discriminant proposals in the literature.

2.2 Background

Consider a binary classification problem where $\mathbf{X} = (X_1, \ldots, X_p)$ represents the predictor vector and $Y = +1, -1$ denotes the class label. Linear discriminant analysis (LDA) is perhaps the oldest classification technique that is still being used routinely in real world applications. The linear discriminant analysis model assumes $\mathrm{x} \mid (Y = y) \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}), \Pr(Y = +1) = \pi_+, \Pr(Y = -1) = \pi_-$. Then, the Bayes rule, which is the theoretically optimal classifier minimizing the 0–1 loss, classifies a data point to class -1 if and only if

$$\left(\mathbf{X} - \frac{\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-}{2}\right)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+) + \log\frac{\pi_-}{\pi_+} > 0. \tag{2.1}$$

Let $\hat{\boldsymbol{\mu}}_+$, n_+ and $\hat{\boldsymbol{\mu}}_-$, n_- be the sample mean vector and sample size within class $+1$ and class -1, respectively. Let $\hat{\boldsymbol{\Sigma}}$ be the sample estimate of $\boldsymbol{\Sigma}$. To implement the Bayes rule, linear discriminant analysis substitutes $\boldsymbol{\mu}_+ = \hat{\boldsymbol{\mu}}_+$, $\boldsymbol{\mu}_- = \hat{\boldsymbol{\mu}}_-$, $\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}}$, $\pi_+ = n_+/n, \pi_- = n_-/n$ in (2.1). Despite its simplicity, linear discriminant analysis has been proven to be a reasonably good classifier in many applications. For example, [40] and [41] have shown that linear discriminant analysis has very competitive performance for many real world benchmark datasets.

With rapid advance of technology, high dimensional data appear more and more frequently in contemporary statistical problems, such as tumor classification using microarray data. In such data the dimension (p) can be much larger than the sample size (n). It has been empirically observed by many that for classification problems with high-dimension-and-low-sample-size data some simple linear classifiers perform as well as much more sophisticated classification algorithms such as the support vector machine and boosting. See, e.g., the comparison study by [42]. [43] provides some geometric insight into this interesting phenomenon. In recent years, many papers have considered ways to modify the usual LDA such that the modified discriminant analysis method is suitable for high dimensional classification. A seemingly obvious choice is by using more sophisticated estimates of the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ to replace the naive sample estimate. Under some sparsity assumption, one can obtain good estimators of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ even when p is much larger than n [44, 45, 46]. However, a better estimate of $\boldsymbol{\Sigma}^{-1}$ does not necessarily lead to a better classifier. Consider an ideal scenario where

we know $\mathbf{\Sigma}$ is an identity matrix. Even so, [47] showed that this classifier performs no better than random guessing when p is sufficiently large, due to noise accumulation in estimating $\boldsymbol{\mu}_+, \boldsymbol{\mu}_-$. Therefore, effectively exploiting sparsity is critically important for high-dimensional classification.

[48] proposed the nearest shrunken centroid classifier (NSC) for tumor classification and gene selection using microarray data. The shrunken centroid classifier is defined as follows. For each variable X_j, we compute $d_{+j} = \dfrac{n}{n_+ n_-}\dfrac{\hat{\mu}_{+j} - \hat{\mu}_{-j}}{s_j + s_0}$ and $d_{-j} = -d_{+j}$, where $\hat{\mu}_{+j}$ and $\hat{\mu}_{-j}$ are the within-class sample mean, s_j^2 is the sample estimate of Σ_{jj} and s_0 is a small positive constant added for robustness consideration. For simplicity, we can think $s_0 = 0$. Define the shrunken centroid mean by

$$\hat{\boldsymbol{\mu}}'_{yj} = \bar{X}_j + \sqrt{\frac{1}{n_y} - \frac{1}{n}} s_j d^{\lambda}_{yj}, \quad y = +1, -1$$

where $\bar{X}_j = \dfrac{n_+\hat{\mu}_{+j} + n_-\hat{\mu}_{-j}}{n}$ is the marginal sample mean of X_j, λ is a pre-chosen positive constant and d^{λ}_{yj} is computed by soft-thresholding d_{yj}: $d^{\lambda}_{yj} = \text{sign}(d_{yj})(|d_{yj}| - \lambda)_+, y = +1, -1$. The NSC classifies $\mathbf{X}$ to class -1 if

$$\sum_{j=1}^{p}\left(X_j - \frac{(\hat{\mu}'_{-j} + \hat{\mu}'_{+j})}{2}\right)\frac{(\hat{\mu}'_{-j} - \hat{\mu}'_{+j})}{s_j^2} + \log\frac{n_-}{n_+} > 0. \tag{2.2}$$

Comparing (2.2) and (2.1), we see that the nearest shrunken centroid classifier modifies the usual LDA in two directions. First, it only uses the diagonal sample covariance matrix to estimate $\mathbf{\Sigma}$. If $\lambda = 0$, NSC reduces to the so-called diagonal linear discriminant analysis. As shown in [49], the diagonal linear discriminant analysis may work much better than the usual LDA in high dimensions. Second, NSC classifier uses the shrunken centroid mean to estimate $\boldsymbol{\mu}_+$, $\boldsymbol{\mu}_-$ in order to perform feature selection. Note that if we use a sufficiently large λ, then the soft-thresholding operation will force $\hat{\mu}'_{+j} = \hat{\mu}'_{-j} = \bar{X}_j$ for some variables and those variables have no contribution to the classifier defined in (2.2). NSC is implemented in the `R` package `pamr` written by Hastie, Tibshirani, Narasimhan and Chu. See

`http://cran.r-project.org/web/packages/pamr/index.html`.

Many empirical experiments have shown that NSC is very competitive for high-dimensional classification. Variants of the shrunken centroid idea have been considered in other

sparse discriminant analysis proposals [50, 51]. More recently, [47] proposed features annealed independence rules in which gene selection is done by hard-thresholding marginal t-statistics for testing whether $\mu_{+j} = \mu_{-j}$.

Since the goal of sparse discriminant analysis is to find genes/features that contribute most to classification, the target of an ideal feature selection should be the discriminative set which contains all "discriminative genes" that contributes to the Bayes rule. This is a very natural argument because we would use the Bayes rule for classification if it was available. Feature selection is needed when the cardinality of the discriminative set is much smaller than the total number of genes/features. The performance of feature selection by a sparse discriminative method is measured by its probability of discovering the discriminative set. There is little theoretical work for justifying NSC and its variants. To our knowledge, only [47] provided some detailed theoretical analysis of features annealed independent rules (FAIR), where the fundamental assumption is that $\mathbf{\Sigma}$ is a diagonal matrix. However, such an assumption is too restrictive to hold in real applications, because strong correlations exist in microarrays and other types of high-dimensional data. It is not hard to see that ignoring the important correlation structure may lead to misleading feature selection results. In fact, we argue that both NSC and FAIR aim to discover the so-called signal set whose definition is given explicitly in Section 2.3. We further provide a necessary and sufficient condition under which the signal set is identical to the discriminative set. The necessary and sufficient condition can be easily violated and hence independent rules could select wrong features.

In this chapter we propose a new recipe for sparse discriminant analysis in high dimensions. Our proposal is motivated by the well-known fact that in the traditional low dimension setting the LDA classifier can be reconstructed exactly via least squares [13]. We suggest using penalized sparse least squares methods to derive sparse discriminant methods. Our proposal is computationally efficient in high dimensions with the help of efficient algorithms for computing penalized least squares. We further provide theoretical justifications for our proposal. Suppose the Bayes rule has a sparse representation. Our theoretical results show that the proposed sparse discriminant method can simultaneously identify the discriminative set and estimate the Bayes classification direction consistently. The theory is valid even when the dimension can grow faster than any polynomial order of the sample size and does not impose strong assumptions

on the correlation structure among predictors.

The rest of this section is organized as follows. In Section 2.3 we discuss the differences between the signal set and the discriminative set. In Section 2.4 we introduce the penalized least squares formulation of sparse discriminant analysis. In Section 2.5 we establish the theoretical properties of Lasso-DSDA and SCAD-DSDA, where the Lasso penalty and the SCAD penalty are used to do feature selection. Numerical results are presented in Section 2.6. Technical proofs are relegated to Appendix A.

2.3 The signal set and the discriminative set

Consider the problem of tumor classification with gene expression arrays. It is an intuitively sound claim that differentially expressed genes should be responsible for the tumor classification and equally expressed genes can be safely discarded. However, we show in this section that a differentially expressed gene can have no role in classification and at the same time an equally expressed gene can significantly influence classification.

We begin with some necessary notation. By definition, the discriminative set is equal to $A = \{j : \left(\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)\right)_j \neq 0\}$, since the Bayes classification direction is $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)$. Variables in A are called informative or discriminative variables. Define $\tilde{A} = \{j : \mu_{+j} \neq \mu_{-j}\}$ which is referred to as the signal set and variables in $\tilde{A}$ are called signals. Independent rules select genes by comparing their within-class means. In an ideal situation, $\tilde{A}$ is the gene selection outcome of an independent rule.

Finding $\tilde{A}$ is of course an interesting and valid statistical inference problem, which is often formulated as a multiple hypothesis testing problem [52, 53]. Various new methods and theories have been developed for doing thousands of hypotheses testing at the same time. See [54, 55, 56, 57, 58, 59, 60], among others. [61] discussed the connection between large-scale classification and large-scale testing under a special LDA model assuming a diagonal covariance matrix. When $\mathbf{\Sigma}$ is diagonal $A = \tilde{A}$. For a general covariance matrix, however, the informative set and the signal set can be very different, as shown in the following proposition.

Proposition 2.3.1 *Let us decompose* $\mathbf{\Sigma}$ *as*

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{A,A} & \mathbf{\Sigma}_{A,A^c} \\ \mathbf{\Sigma}_{A^c,A} & \mathbf{\Sigma}_{A^c,A^c} \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\tilde{A},\tilde{A}} & \mathbf{\Sigma}_{\tilde{A},\tilde{A}^c} \\ \mathbf{\Sigma}_{\tilde{A}^c,\tilde{A}} & \mathbf{\Sigma}_{\tilde{A}^c,\tilde{A}^c} \end{pmatrix}.$$

1. $A \subseteq \tilde{A}$ *if and only if* $\boldsymbol{\Sigma}_{\tilde{A}^c,\tilde{A}}\boldsymbol{\Sigma}_{\tilde{A},\tilde{A}}^{-1}(\boldsymbol{\mu}_{-,\tilde{A}} - \boldsymbol{\mu}_{+,\tilde{A}}) = 0$.

2. $\tilde{A} \subseteq A$ *if and only if* $\boldsymbol{\mu}_{+,A^c} = \boldsymbol{\mu}_{-,A^c}$ *or* $\boldsymbol{\Sigma}_{A^c,A}\boldsymbol{\Sigma}_{A,A}^{-1}(\boldsymbol{\mu}_{-,A} - \boldsymbol{\mu}_{+,A}) = 0$.

Based on Proposition 2.3.1 it is very easy to construct concrete examples to show that a non-signal can be informative, and vice versa. Here are two examples. Consider a LDA model with $\boldsymbol{\mu}_+ = 0_p$, $\Sigma_{i,i} = 1$ and $\Sigma_{i,j} = 0.5$, $1 \le i, j \le 25$ and $i \ne j$, where $p = 25$. If $\boldsymbol{\mu}_- = (1, 1, 1, 1, 1, 0, \cdots, 0)^T$, then $\tilde{A} = \{1, 2, 3, 4, 5\}$ and $A = \{j : j = 1, \cdots, 25\}$, i.e., all variables are informative. Similarly, if let $\boldsymbol{\mu}_- = (3, 3, 3, 3, 3, 2.5, \cdots, 2.5)^T$, then all variables are signals but $A = \{1, 2, 3, 4, 5\}$.

The above arguments warn us that sparse discriminant analysis using independent rules could end up with a wrong set of features. A different sparse discriminant analysis method was recently proposed by [62]. Their proposal starts with Fisher's view of LDA, that is, the LDA direction is obtained by maximizing $\boldsymbol{\beta}^{\mathrm{T}}\hat{\mathbf{B}}\boldsymbol{\beta}/\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta}$ where $\hat{\boldsymbol{\Sigma}}$ is the within covariance matrix and $\hat{\mathbf{B}} = (\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)$ is the between covariance matrix. Note that $\boldsymbol{\beta}^{\mathrm{T}}\hat{\mathbf{B}}\boldsymbol{\beta} = \|(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\boldsymbol{\beta}\|_2^2$. [62] proposed the following SFDA:

$$\min \boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} \quad \text{s.t.} \quad (\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\boldsymbol{\beta} = 1 \text{ and } \|\boldsymbol{\beta}\|_1 \le \tau. \tag{2.3}$$

[63] proposed another ℓ_1-penalized linear discriminant analysis:

$$\max\{\boldsymbol{\beta}^{\mathrm{T}}\hat{\mathbf{B}}\boldsymbol{\beta} - \lambda\sum_{j=1}^{p}|s_j\beta_j|\}, \text{subject to } \boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} \le 1. \tag{2.4}$$

Little is known about the theoretical properties of the estimators defined in (2.3) and (2.4). However, it is not our interest in this paper to prove or disprove these two methods, although we do include them in our numerical experiments.

2.4 Methodology

2.4.1 Sparse LDA via penalized least squares

Our approach to sparse LDA is motivated by the intimate connection between linear discriminant analysis and least squares in the classical $p < n$ setting [13]. Suppose we numerically code the class labels as $Y = -n/n_+$ if $\mathbf{X}$ belongs to the positive class and

$Y = n/n_-$ if $\mathbf{X}$ belongs to the negative class, where $n = n_+ + n_-$. Let

$$(\hat{\boldsymbol{\beta}}^{\text{ols}}, \hat{\beta}_0^{\text{ols}}) = \arg\min_{\boldsymbol{\beta},\beta_0} \sum_{i=1}^{n} (Y^i - \beta_0 - (\mathbf{X}^i)^{\mathrm{T}}\boldsymbol{\beta})^2 \tag{2.5}$$

Then $\hat{\boldsymbol{\beta}}^{\text{ols}} = c\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)$ for some positive constant c. In other words, the least square formulation in (2.5) exactly derives the usual LDA direction.

The connection is lost in high dimensional problems because the sample covariance estimate is no longer invertible and the LDA direction is not well defined in its original form. However, we may consider a penalized least squares formulation to produce a classification direction. Let $P_\lambda(\cdot)$ be a generic sparsity-inducing penalty. Specific choices of $P_\lambda(\cdot)$ are given in Section 2.2. We first compute the solution to a penalized least squares problem

$$(\hat{\boldsymbol{\beta}}^{\lambda}, \hat{\beta}_0^{\lambda}) = \arg\min_{\boldsymbol{\beta},\beta_0} \frac{1}{n} \sum_{i=1}^{n} (Y^i - \beta_0 - (\mathbf{X}^i)^{\mathrm{T}}\boldsymbol{\beta})^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|). \tag{2.6}$$

Then our classification rule is to assign x to class -1 if

$$\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{\lambda} + \hat{\beta}_0 > 0. \tag{2.7}$$

It is important to note that $\hat{\beta}_0$ in (2.7) differs from $\hat{\boldsymbol{\beta}}_0^{\lambda}$ in (2.6). In the $p \ll n$ case, consider the OLS estimator and the usual LDA. Let us write $\hat{\boldsymbol{\beta}}^{\text{ols}} = c\hat{\boldsymbol{\beta}}^{\text{LDA}}$, $\hat{\boldsymbol{\beta}}^{\text{LDA}} = \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)$. We should use $\hat{\beta}_0 = c\hat{\beta}_0^{\text{LDA}}$ in (2.7) where

$$\hat{\boldsymbol{\beta}}_0^{\text{LDA}} = \log\left(\frac{n_+}{n_-}\right) - \left(\frac{\hat{\boldsymbol{\mu}}_+ + \hat{\boldsymbol{\mu}}_-}{2}\right)^{\mathrm{T}} \hat{\boldsymbol{\beta}}^{\text{LDA}}$$

such that the OLS classifier and the LDA rule yield identical classification. If we use $\hat{\beta}_0^{\text{ols}}$ in (2.7), the the OLS classifier is in general not identical to the LDA rule.

Finding the right intercept is critical for classification but receives little attention in the literature. [13] mentioned that one could choose the intercept $\hat{\beta}_0$ empirically by minimizing the training error. We show here that for a given classification direction, there is a nice closed-form formula for the optimal intercept.

Proposition 2.4.1 *Suppose a linear classifier assigns* x *to class* -1 *if* $\mathbf{X}^{\mathrm{T}}\tilde{\boldsymbol{\beta}} + \tilde{\beta}_0 > 0$. *Given* $\tilde{\beta}$, *if* $(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}} > 0$, *then the optimal intercept* $\tilde{\beta}_0$ *is*

$$\tilde{\beta}_0^{opt.} = -\frac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)^{\mathrm{T}}\tilde{\boldsymbol{\beta}} + \frac{\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}}{(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}}} \log\frac{\pi_-}{\pi_+}, \tag{2.8}$$

which can be estimated by

$$\widehat{\tilde{\beta}_0^{opt.}} = -\frac{1}{2}(\hat{\boldsymbol{\mu}}_+ + \hat{\boldsymbol{\mu}}_-)^{\mathrm{T}}\tilde{\boldsymbol{\beta}} + \frac{\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\beta}}}{(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}}}\log\frac{n_-}{n_+}. \tag{2.9}$$

Without sparsity condition on $\tilde{\beta}$, the estimator given in (2.9) would not work well when $p > n$. However, when $\tilde{\beta}$ is sparse, we have

$$\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}} = \sum_{i,j:\tilde{\beta}_i\neq 0,\tilde{\beta}_j\neq 0}\Sigma_{ij}\tilde{\beta}_i\tilde{\beta}_j, \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\beta}} = \sum_{i,j:\tilde{\beta}_i\neq 0,\tilde{\beta}_j\neq 0}\hat{\Sigma}_{ij}\tilde{\beta}_i\tilde{\beta}_j.$$

Even when $p \gg n$, as long as $\|\tilde{\boldsymbol{\beta}}\|_0 \ll n$, $\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\beta}}$ is a good estimator for $\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}$. Using a regularized estimate of $\boldsymbol{\Sigma}$ could provide some further improvement. For example, for banded covariance matrices, the banding estimator [44] and the tapering estimator [45] are better estimators for $\boldsymbol{\Sigma}$ than the sample covariance. However, in this work our primary focus is $\hat{\boldsymbol{\beta}}^{\lambda}$ and we do not want to entangle the issue of estimating large covariance matrices with the problem of feature selection.

The condition $(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}} > 0$ in proposition 2.4.1 is very mild. Suppose the linear classifier actually yields $(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}} < 0$, then it is easy to show that such a classifier is dominated by the other linear classifier using direction $\tilde{\boldsymbol{\beta}}_{new} = -\tilde{\boldsymbol{\beta}}$ that obeys $(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_{new} > 0$.

By proposition 2.4.1 , the sparse LDA classifier is defined as follows: assigning x to class -1 if

$$\left(\mathbf{X} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_+ + \hat{\boldsymbol{\mu}}_-)\right)^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{\lambda} + \frac{(\hat{\boldsymbol{\beta}}^{\lambda})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}}^{\lambda}}{(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{\lambda}}\log\frac{n_-}{n_+} > 0. \tag{2.10}$$

Because this sparse LDA classifier directly regularizes the discriminant direction, we refer to this classifier as direct sparse discriminant analysis (DSDA). DSDA depends on the regularization parameter λ. In practice we need to select a good regularization parameter such that the generalization error is as small as possible. Cross-validation is a popular method for tuning. In this paper we use cross-validation to select λ under the 0-1 loss.

2.4.2 Choice of penalty and computing algorithm

We now discuss the choice of penalty functions in (2.6). We note first that our DSDA approach can work with any sparsity-inducing penalty function. In recent years, many

papers have been devoted to designing nice penalty functions for sparse regression. Some well-known examples are lasso [3], SCAD [4], elastic net [17], fused lasso [64], grouped lasso [20], adaptive lasso [18], MCP [19] and SICA [65], among others. [66] provided a good review on feature selection and penalized regression models. Roughly speaking, these penalty functions can be classified into two categories: the convex family and the concave family, with lasso and SCAD being the representing examples. To fix idea, we focus on the lasso and SCAD penalties when constructing sparse LDA classifiers. The lasso penalty function is $P_\lambda(t) = \lambda t$ for $t \geq 0$. The SCAD penalty function is defined by $P_{\lambda,a}(0) = 0$ and $P'_{\lambda,a}(t) = \lambda I(t \leq \lambda) + \frac{(a\lambda - t)_+}{a-1} I(t > \lambda)$ for $t > 0$ where $a > 2$. Following Fan and Lv (2001), we used $a = 3.7$ in our numerical experiments. If the lasso penalty is used in (2.6), we call the resulting classifier Lasso-DSDA. Likewise, if the SCAD penalty is used, we call the resulting classifier SCAD-DSDA.

There has been considerable progress in developing efficient algorithms for computing sparse regularized regression models in high dimensions. The LARS algorithm [25], implemented in the `R` package `lars`, computes the entire solution path for the lasso regression with the same order of computational cost as a single ordinary least squares fit. [26] implemented the coordinate descent algorithm for computing the lasso regression in the `R` package `glmnet` and showed that `glmnet` can be even faster than `lars`. [34] showed that using the LLA algorithm one can solve the concave penalized regression problem via an iterative weighted-lasso regression procedure. One could combine the LLA algorithm and `glmnet` to solve any concave penalized least squares for each fixed λ. For the SCAD penalty, it turns out that an even faster algorithm is possible by directly applying the coordinate descent principle. The coordinate decent algorithm works well for the lasso because the univariate lasso regression solution is given by the soft-thresholding rule [3, 26]. Likewise, the univariate SCAD solution also has a closed-form formula given by the SCAD thresholding rule [4]. Therefore, with some proper modification, `glmnet` can be used to compute the SCAD penalized regression. In a word, both Lasso-DSDA and SCAD-DSDA can be computed very efficiently even when $p \gg n$. Hence they are practically useful for high dimensional classification problems.

2.5 Statistical Theory

In this section we study the theoretical properties of the sparse LDA classifiers based on lasso/SCAD penalized least squares. In the literature there are many results on sparse penalized least squares [4, 18, 32, 67, 19, 36, 65]. But they cannot be directly applied to our setting although we borrow the least squares criterion to derive the sparse LDA classifier, because the linear model assumption ($y = \sum_j x_j\beta_j + error$), the foundation for these existing theoretical work, does not hold for the LDA model. Furthermore, if we regard the predictor matrix as the "design" matrix, then our theory always deals with the random design case, whereas the fixed design theory is common in the existing work on high-dimensional penalized least squares regression.

2.5.1 Notation and definitions

We first introduce some necessary notation to be used in the theoretical analysis. For a general $m \times n$ matrix $\boldsymbol{M}$, define $\|\boldsymbol{M}\|_\infty = \max_{i=1,..,m} \sum_{j=1}^n |M_{ij}|$. For any vector $\boldsymbol{b}$, let $\|\boldsymbol{b}\|_\infty = \max_j |b_j|$ and $|\boldsymbol{b}|_{\min} = \min_j |b_j|$. We let $\boldsymbol{\beta}(\text{Bayes}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)$ represent the Bayes classifier coefficient vector. So $A = \{j : \beta(\text{Bayes})_j \neq 0\}$ and let $s = |A|$. We use $\mathbf{C} = \text{Cov}(\mathbf{X})$ to represent the marginal covariance matrix of the predictors and partition $\mathbf{C}$ as $\mathbf{C} = \begin{pmatrix} \mathbf{C}_{AA} & \mathbf{C}_{AA^c} \\ \mathbf{C}_{A^cA} & \mathbf{C}_{A^cA^c} \end{pmatrix}$. We define three quantities that frequently appear in our analysis:

$$\kappa = \|\mathbf{C}_{A^cA}(\mathbf{C}_{AA})^{-1}\|_\infty, \tag{2.11}$$

$$\varphi = \|(\mathbf{C}_{AA})^{-1}\|_\infty, \tag{2.12}$$

$$\Delta = \|\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A}\|_\infty. \tag{2.13}$$

Suppose X is the predictor matrix and let $\tilde{X}$ be the centered predictor matrix such that the column-wise mean is zero. Obviously, $C^{(n)} = \frac{1}{n}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}$ is an empirical version of C. Likewise, we can write $\frac{1}{n}\tilde{\mathbf{X}}_A^{\mathrm{T}}\tilde{\mathbf{X}}_A = \mathbf{C}_{AA}^{(n)}$ and $\frac{1}{n}\tilde{\mathbf{X}}_{A^c}^{\mathrm{T}}\tilde{\mathbf{X}}_A = \mathbf{C}_{A^cA}^{(n)}$.

Denote $\boldsymbol{\beta}^* = (\mathbf{C}_{AA})^{-1}(\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})$. Now we can define $\tilde{\boldsymbol{\beta}}(\text{Bayes})$ by letting $\tilde{\boldsymbol{\beta}}(\text{Bayes})_A = \boldsymbol{\beta}^*$ and $\tilde{\boldsymbol{\beta}}(\text{Bayes})_{A^c} = 0$. The following is a simple but very useful result, showing the equivalence between $\tilde{\boldsymbol{\beta}}(\text{Bayes})$ and $\boldsymbol{\beta}(\text{Bayes})$ in the context of LDA model.

Proposition 2.5.1 *$\tilde{\boldsymbol{\beta}}$(Bayes) and $\boldsymbol{\beta}$(Bayes) are equivalent in the sense that $\tilde{\boldsymbol{\beta}}$(Bayes) = $c\boldsymbol{\beta}$(Bayes) for some positive constant c and the Bayes classifier is also equivalent to assigning $\mathbf{X}$ to class -1 if*

$$(\mathbf{X} - \frac{\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-}{2})^{\mathrm{T}}\tilde{\boldsymbol{\beta}}(\text{Bayes}) + \frac{(\tilde{\boldsymbol{\beta}}(\text{Bayes}))^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}(\text{Bayes})}{(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}}(\text{Bayes})}\log\frac{\pi_-}{\pi_+} > 0. \tag{2.14}$$

Proposition 2.5.1 tells us that it suffices to show the proposed sparse LDA can consistently recover the support of $\tilde{\boldsymbol{\beta}}$(Bayes) and estimate $\boldsymbol{\beta}^*$.

2.5.2 Main results

We now present the main theoretical results. In our analysis we assume the variance of each variables is bounded by a finite constant. This regularity condition usually holds. In practice, one often standardizes the data beforehand. Then the finite constant can be taken as one. In this subsection, ϵ_0 and c_1, c_2 are some positive constants.

Suppose the Lasso-DSDA estimator does find the support of the Bayes rule, A, then we have $\hat{\beta}(\text{lasso})_{A^c} = 0$ and $\hat{\boldsymbol{\beta}}(\text{lasso})_A$ should be identical to $\hat{\boldsymbol{\beta}}_A$, where

$$\hat{\boldsymbol{\beta}}_A = \arg\min_{\boldsymbol{\beta},\beta_0}\frac{1}{n}\sum_{i=1}^{n}(Y^i - \beta_0 - \sum_{j\in A}X_{ij}\beta_j)^2 + \sum_{j\in A}\lambda|\beta_j|. \tag{2.15}$$

We introduce $\hat{\boldsymbol{\beta}}_A$ only for mathematical analysis. It is not a real estimator, because its definition depends on knowing A.

To ensure the Lasso-DSDA classifier has the variable selection consistency property, we impose a condition on the covariance matrix of the predictors:

$$\kappa = \|\mathbf{C}_{A^cA}(\mathbf{C}_{AA})^{-1}\|_\infty < 1. \tag{2.16}$$

The above condition is an analogue of the irrepresentable condition for the Lasso regression estimator [32, 18].

Theorem 2.5.2 (Analysis of Lasso-DSDA) *Pick any λ such that $\lambda < \min(\frac{1}{2}|\boldsymbol{\beta}^*|_{min}/\varphi, \Delta)$.*

1. *Assuming the condition in (4.14), with probability at least $1 - \delta_1$, $\hat{\boldsymbol{\beta}}_A(lasso) = \hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_{A^c}(lasso) = 0$, where*

$$\delta_1 = 2ps\exp(-\frac{n}{s^2}\epsilon^2 c_1) + 2p\exp(-nc_2(\frac{\lambda}{4}\frac{1-\kappa-2\epsilon\varphi}{1+\kappa})^2) \tag{2.17}$$

and ϵ is any positive constant satisfying $\epsilon < \min(\epsilon_0, \frac{\frac{\lambda}{4\varphi}(1-\kappa)}{\frac{\lambda}{2}+(1+\kappa)\Delta})$.

2. *With probability at least $1-\delta_2$, none of the elements of $\hat{\beta}_A$ is zero, where*

$$\delta_2 = 2s^2 \exp(-\frac{nc_1}{s^2}\epsilon^2) + 2s\exp(-nc_2\epsilon^2), \tag{2.18}$$

where ϵ is any positive constant satisfying $\epsilon < \min(\epsilon_0, \frac{1}{\varphi}\frac{\zeta}{3+\zeta})$.

3.

$$\Pr(\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}^*\|_\infty \le 4\varphi\lambda) \quad \ge \quad 1 - 2s^2\exp(-\frac{nc_1}{s^2}\epsilon^2) - 2s\exp(-nc_2\epsilon^2), \tag{2.19}$$

where ϵ is any positive constant satisfying $\epsilon < \min(\epsilon_0, \frac{\lambda}{2\varphi\Delta}, \lambda)$.

In our analysis we compare SCAD-DSDA to an oracle estimator knowing the true feature set A. We first define

$$\tilde{\boldsymbol{\beta}}(\text{oracle})_A = (\mathbf{C}_{AA}^{(n)})^{-1}(\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A}). \tag{2.20}$$

It is easy to see that $\tilde{\boldsymbol{\beta}}(\text{oracle})_A$ is the solution to a least squares criterion $\min \sum_{i=1}^n (Y^i - \beta_0 - \sum_{j\in A} X_j^i\beta_j)^2$. Hence, in terms of classification $\tilde{\boldsymbol{\beta}}(\text{oracle})$ is equivalent to the oracle LDA only using the subset A. The oracle estimator is then defined as $\hat{\boldsymbol{\beta}}(\text{oracle})$ such that $\hat{\boldsymbol{\beta}}(\text{oracle})_A = \tilde{\boldsymbol{\beta}}(\text{oracle})_A$ and $\hat{\boldsymbol{\beta}}(\text{oracle})_{A^c} = 0$.

Theorem 2.5.3 (Analysis of SCAD-DSDA) *1. For any $\epsilon > 0$ satisfying*

$$\frac{\epsilon}{\epsilon + 2\Delta\varphi} \le \min(\epsilon_0\varphi, \epsilon_0\frac{1}{\Delta}),$$

we have

$$\begin{aligned} & \Pr(\|\hat{\boldsymbol{\beta}}(oracle) - \boldsymbol{\beta}^*\|_\infty \ge \epsilon) \\ \le \quad & 2s^2\exp(-\frac{nc_1}{4s^2}\frac{\epsilon^2}{\varphi^2(\epsilon+2\Delta\varphi)^2}) + 2s\exp(-\frac{nc_2}{4}\frac{\epsilon^2\Delta^2}{(\epsilon+2\Delta\varphi)^2}). \end{aligned} \tag{2.21}$$

2. *For any $\lambda < \frac{|\beta^*|_{min}}{a}$, with probability at least $1-\delta_3$, none of the elements of $\hat{\beta}(oracle)_A$ is zero and $\hat{\beta}(oracle)$ is a local solution to the SCAD-DSDA criterion, where*

$$\begin{aligned} \delta_3 \quad = \quad & 2p\exp(-nc_2\epsilon^2) + 2ps\exp(-\frac{nc_1}{s^2}\frac{1}{\varphi^2}(\frac{\epsilon}{\epsilon+\kappa+1})^2) \\ & +2s^2\exp(-\frac{nc_1}{4s^2}\frac{\epsilon^2}{\varphi^2(\epsilon+2\Delta\varphi)^2}) + 2s\exp(-\frac{nc_2}{4}\frac{\epsilon^2\Delta^2}{(\epsilon+2\Delta\varphi)^2}), \end{aligned} \tag{2.22}$$

where ϵ is any positive constant, $\epsilon < \epsilon_0$ and obeys the following constraints: $\frac{\epsilon}{\epsilon+\kappa+1} < \varphi\epsilon_0$, $\frac{\epsilon}{\epsilon+2\Delta\varphi} \leq \min(\epsilon_0\varphi, \epsilon_0\frac{1}{\Delta})$, *and* $\epsilon < \min(|\beta^*|_{min} - a\lambda, \frac{\lambda}{6\Delta}, \frac{\lambda}{6}, \frac{\lambda}{6\kappa+\frac{\lambda}{\Delta}})$.

The analysis of SCAD-DSDA does not require condition (4.14). Theorem 2 works for any positive κ.

The non-asymptotic results in Theorems 1 and 2 can be easily translated into some asymptotic arguments when considering the triple of (n, s, p) goes to infinity at some proper rates. To highlight the main points, we assume Δ, κ, φ are constants in the asymptotic arguments. In addition, we need the following two regularity conditions:

(C1). $n, p \to \infty$ and $\log(ps)s^2/n \to 0$,

(C2). $|\boldsymbol{\beta}^*|_{min} \gg \sqrt{\frac{\log(ps)s^2}{n}}$.

Condition (C1) puts some restriction on p. Clearly, we cannot expect the proposed method (or any sensible method) works for an arbitrarily large p. However, the restriction is rather loose. Consider the case where $s = o(n^{\frac{1}{2}-\gamma})$ for some $\gamma < \frac{1}{2}$. (C1) holds as long as $p \ll e^{n^{2\gamma}}$. Therefore, p is allowed to grow faster than any polynomial order of n. This implies the applicability of the Lasso-DSDA and SCAD-DSDA to real world problems such as gene expression classification.

Condition (C2) requires the non-zero elements of the Bayes rule to be large enough such that we could consistently separate them from zeros by using observed data. The lower bound actually converges to zero asymptotically under (C1), and hence condition (C2) is not a strong assumption.

Theorem 2.5.4 (Asymptotic properties of Lasso-DSDA and SCAD-DSDA) *Under conditions (C1) and (C2), if we choose some $\lambda = \lambda_n$ such that $\lambda_n \ll |\boldsymbol{\beta}^*|_{min}$ and $\lambda_n \gg \sqrt{\log(ps)s^2/n}$, then , with probability going to one, a SCAD-DSDA solution is identical to the oracle LDA that is consistent in feature selection and $\|\hat{\boldsymbol{\beta}}(oracle)_A - \boldsymbol{\beta}^*\|_\infty = o_P(\sqrt{\log(s)s^2/n})$. Moreover, if we further assume $\kappa < 1$, then the Lasso-DSDA is consistent in feature selection and $\|\hat{\boldsymbol{\beta}}(oracle)_A - \boldsymbol{\beta}^*\|_\infty = o_P(\lambda_n)$.*

Finally, it is important to point out that our theory does not require any structure assumption on the common covariance matrix $\boldsymbol{\Sigma}$, which clearly shows the fundamental

difference between our method and those based on high dimensional covariance estimation. In the current literature on covariance or inverse-covariance matrix estimation, a commonly used assumption is that the target matrix has some sparsity structure [44, 45, 46]. Such assumptions are not needed in our method.

2.6 Numerical Results

2.6.1 Simulation

We use simulated data to demonstrate the good performance of Lasso-DSDA and SCAD-DSDA. For comparison, we included NSC, FAIR, the sparse LDA proposed by [63] and the ℓ_1 Fisher's discriminant analysis (SFDA) proposed by [62]. NSC is implemented in the R package `pamr`. The sparse LDA proposed by [63] is implemented in the R package `penalizedLDA`. We used the code by Dr. Wu to implement their proposal of sparse LDA.

We randomly generated n class labels such that $\pi_+ = \pi_- = 0.5$. Conditioning on the class labels Y, we generated the p-dimensional predictor $\mathbf{X}$ from a multivariate normal distribution with mean vector $\boldsymbol{\mu}_y$ and covariance $\boldsymbol{\Sigma}$. Without loss of generality, we set $\boldsymbol{\mu}_+ = 0$ and $\boldsymbol{\mu}_- = \boldsymbol{\Sigma}\boldsymbol{\beta}^{\text{Bayes}}$. We considered six different simulation models. The choices of n, p, $\boldsymbol{\mu}_-$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}^{\text{Bayes}}$ are shown in Table 2.1. Models 1-4 are sparse discriminant models with different covariance and mean structure, while models 5 and 6 are "practically sparse" in the sense that their Bays rules depend on all variables in theory but can be well approximated by sparse discriminant functions. Table 2.2 summarizes the simulation results based on 2000 replications. For each measure we reported its median and the corresponding standard error in parentheses. Only Lasso-DSDA and SCAD-DSDA show consistently good performance in all six simulation settings. They closely mimic the Bayes rule, regardless of the Bayes error and covariance structure. NSC and FAIR have very comparable performance, but they are much worse than Lasso-DSDA and SCAD-DSDA except in model 1. By direct calculation one can see that the first five elements of $\boldsymbol{\mu}_- - \boldsymbol{\mu}_+$ are much larger than the rest, which implies that independence rules can include all three discriminative variables. On the other hand, although model 2 uses the same $\boldsymbol{\Sigma}$ as in model 1, it has very different mean structure: the first two elements of $\boldsymbol{\mu}_- - \boldsymbol{\mu}_+$ are dominating while the rest are much smaller. This means that

independence rules have difficulty in selecting variable three, resulting inferior classification. Wu's method has good classification accuracy overall, but it can often miss some important features. Witten's method has rather poor performance, which is somewhat surprising because the basic idea behind Witten's method is similar to Wu's. We do notice that Witten's formulation in (4) is nonconvex while Wu's formulation in (3) and is convex, which may help explain their different performance.

Table 2.1: Simulation settings.

Model	n	p	$\mathbf{\Sigma}$	β^{Bayes}
1	100	400	Σ_{ij} =$0.5^{\lvert i-j\rvert}$	$0.556(3,\ 1.5,\ 0,\ 0,\ 2,0_{p-5})^{\mathrm{T}}$
2	100	400	Σ_{ij} =$0.5^{\lvert i-j\rvert}$	$0.582(3,\ 2.5,\ -2.8,\ 0_{p-3})^{T}$
3	400	800	$\Sigma_{jj}=1, \Sigma_{ij}$=0.5, $i\neq j$.	$0.395(3,\ 1.7,\ -2.2,\ -2.1,2.55,0_{p-5})^{\mathrm{T}}$
4	300	800	$\mathbf{\Sigma}=\mathbf{I}\otimes\tilde{\mathbf{\Sigma}}$, $\tilde{\Sigma}_{jj}=1$, $\tilde{\Sigma}_{ij}=0.6, i\neq j$	$0.916(1.2,-1.4,1.15,-1.64,1.5,-1,2,0_{p-7})^{\mathrm{T}}$
5	400	800	$\Sigma_{jj}=1, \Sigma_{ij}=0.5$, $i\neq j$.	$0.551(3,\ 1.7,\ -2.2,-2.1,2.55,(p-5)^{-1}1_{p-5})^{\mathrm{T}}$
6	400	800	$\Sigma_{jj}=1, \Sigma_{ij}=0.5$, $i\neq j$.	$0.362(3, 1.7, -2.2, -2.1, 2.55, (p-5)^{-1}1_{p-5})^{\mathrm{T}}$

2.6.2 Real data

We further compare the methods on two benchmark datasets: Colon and Prostate cancer data. The basic task here is to predict whether an observation is tumor or normal tissue. We randomly split the datasets into the training and test sets with $2:1$ ratio. Model fitting was done on the training set and the classification accuracy was evaluated on the test set. This procedure was repeated 100 times. Shown in Table 2.4 are the (median) classification accuracy and the number of selected genes by each competitor.

Colon and Prostate data have been previously used to test classification and feature selection methods. See [68], [69] and [42]. [42] reported that BagBoost was the most accurate classifier for the Prostate data, with classification accuracy 92.5% and the nearest shrunken centroids classifier was the most accurate classifier for the Colon data. Table 2.3 shows that both Lasso-DSDA and SCAD-DSDA are as accurate as the nearest shrunken centroids classifier on the Colon data and Lasso-DSDA significantly outperforms BagBoost on the Prostate data. Since BagBoost does not do gene selection, we do not include it in Table 2.3. Witten's method works quite well on these two real

datasets.

To better understand the different performance, we list the five most frequently selected genes in Table 2.3. It can be seen that, in the Colon data, genes #625 and # 1772 are frequently used by Lasso-DSDA and SCAD-DSDA, while not by NSC and FAIR. In Prostate data, gene #5016 is frequently used by Lasso and SCAD DSDA, but not by NSC and FAIR. This further shows that the correlations will have an impact on variable selection.

Note that Lasso-DSDA and SCAD-DSDA select similar numbers of genes. If we force other methods to select similar numbers of genes, the results would be as listed in Table 2.4. Wu's and Witten's methods have improved performance, while NSC and FAIR have worse performance.

2.7 Discussion

Sparse discriminant analysis based on independence rules is computationally attractive for high dimensional classification. However, they may lead to misleading feature selection results and hence poor classification performance. Their limitation is due to the fundamental difference between discriminative and signal variables. When doing feature selection in classification, one should aim to recover the discriminative set not the signal set. Finding the signal set is the goal of large-scale hypothesis testing. We should point out that our arguments are not against the developments of theory and methodology for large-scale multiple hypothesis testing. Discovering "signals" is the fundamental question of research in many scientific studies. We only wish to warn the practitioners that the problem of identifying features for discrimination could be very different from identifying interesting signals, and hence the statistical tools for data analysis should be carefully chosen.

Built upon such insight, we have proposed a regularized least squares approach towards sparse LDA models. This approach is computationally efficient for handling high dimensional data. We have established some non-asymptotic theory for the Lasso/SCAD penalized LDA classifiers, from which NP-dimension asymptotic consistency results have

been shown to hold for the Lasso and SCAD DSDA classifiers. In addition, the numerical results are very promising, suggesting the great potential of the proposed sparse LDA classifiers for real world applications.

The regularized least squares can be flexibly modified to accommodate some specific goals. For instance, if we wish to conduct group-wise variable selection when the groups are clearly defined, then we could use the grouped lasso penalty [20]. If we wish to impose certain smoothness structure to the classification coefficients, we could apply the fused lasso penalty [64]. In some situations the predictors may have a natural ordering where the ordered variable selection is preferred. For that, we could apply the hierarchical LARS algorithm [70] or the nested lasso penalty [71] to obtain the regularized least squares fit.

Table 2.2: Simulation results. The standard errors are reported in parentheses. Lasso-DSDA and SCAD-DSDA consistently give good results in both classification and variable selection.

	Bayes rule	Lasso DSDA	SCAD DSDA	SFDA	Witten	NSC	FAIR
Model 1							
Error(%)	10	10.89	11.39	13.71	10.81	10.94	11.47
		(0.03)	(0.04)	(0.01)	(0.01)	(0.02)	(0.05)
TRUE Selection	3	3	3	3	1	3	3
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE Selection	0	2	0	0	26	6	7
		(0.16)	(0)	(0.49)	(0.11)	(0.61)	(0.66)
Model 2							
Error(%)	10	12.84	14.03	14.5	14.25	15.12	15.67
		(0.05)	(0.05)	(0.01)	(0.02)	(0.05)	(0.07)
TRUE Selection	3	3	3	1	2	2	2
		(0)	(0.48)	(0.14)	(0)	(0.34)	(0)
FALSE Selection	0	6	13	0	4	9	8
		(0.27)	(0.64)	(0)	(0.61)	(0.73)	(0.29)
Model 3							
Error(%)	20	21.93	21.37	22.37	33.69	27.48	25.69
		(0.03)	(0.03)	(0.05)	(0.01)	(0.07)	(0.02)
TRUE Selection	5	5	5	5	3	3	2
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE Selection	0	14	12	2	419.5	2	0
		(0.59)	(0.49)	(0)	(10.19)	(0.31)	(0)
Model 4							
Error(%)	10	12.50	12.12	13.99	23.90	19.25	18.56
		(0.02)	(0.03)	(0.03)	(0.01)	(0.04)	(0.00)
TRUE Selection	7	7	7	6	4	4	3
		(0)	(0.03)	(0)	(0)	(0)	(0)
FALSE Selection	0	18	15	2	35	1	0
		(0.70)	(0.42)	(0)	(4.43)	(0.48)	(0)
Model 5							
Error(%)	10	11.11	10.55	12.07	21.99	14.72	14.27
		(0.02)	(0.03)	(0.07)	(0.01)	(0.03)	(0.01)
Fitted model size	800	21	14	7	737	3	3
		(0.65)	(0.19)	(0.16)	(2.29)	(0.46)	(0)
Model 6							
Error(%)	20	22.22	21.62	23.34	30.43	26.13	24.14
		(0.03)	(0.03)	(0.05)	(0.01)	(0.07)	(0)
Fitted model size	800	20	16	5	592.5	8	3
		(0.53)	(0.31)	(0.49)	(7.46)	(0.51)	(0)

Table 2.3: Comparison on the Colon and the Prostate data.

		Lasso DSDA	SCAD DSDA	SFDA	Witten	NSC	FAIR
Colon	Error(%)	86.4	86.4	84.1	86.4	86.4	86.4
		(1.54)	(2.08)	(2.17)	(0.49)	(1.20)	(0.61)
	Fitted	5	6	1	10	89	11
	model size	(0.63)	(0.60)	(0)	(1.39)	(29.95)	(1.19)
Prostate	Error(%)	94.1	91.2	91.2	91.2	91.2	76.5
		(0.55)	(1.37)	(0.70)	(0.24)	(0.96)	(0.54)
	Fitte	10	8	1	18	10	4
	model size	(0.77)	(0.96)	(0)	(4.45)	(0.84)	(0.40)

Table 2.4: Gene selection results for Colon and Prostate Data. We report the five most frequently selected genes.

Colon

Lasso-DSDA		SCAD-DSDA		NSC		FAIR		SFDA	
Gene	Freq.	Gene	Freq.	Gene	Freq.	Gene	Freq.	Gene	Freq.
377	0.71	377	0.69	249	1	493	0.89	249	0.41
493	0.56	493	0.51	493	1	1635	0.81	1423	0.25
625	0.33	249	0.34	765	1	377	0.77	493	0.23
249	0.31	625	0.26	1423	1	249	0.68	897	0.04
1772	0.31	1772	0.25	1635	1	1423	0.57	1325	0.03

Prostate

Lasso-DSDA		SCAD-DSDA		NSC		FAIR		SFDA	
Gene	Freq.	Gene	Freq.	Gene	Freq	Gene	Freq.	Gene	Freq.
1839	1.00	2619	0.96	1839	1	2619	1.00	2619	0.91
2619	1.00	1839	0.89	2619	1	1839	0.99	1839	0.15
3423	0.75	3423	0.61	5016	0.89	5016	0.90	4701	0.03
5016	0.75	5016	0.47	4155	0.87	4701	0.65	4155	0.02
4288	0.57	4288	0.43	2425	0.74	4155	0.63	5016	0.02

Table 2.5: Classification accuracies if we force all the methods to select similar numbers of genes as Lasso-DSDA and SCAD-DSDA.

		SFDA	Witten	NSC	FAIR
Colon	Error(%)	86.4(1.06)	86.4(0.51)	63.6(0.70)	77.3(2.16)
Prostate	Error(%)	91.2(1.24)	94.1(1.25)	91.2(1.39)	73.5(1.11)

Chapter 3

The Connection of Some Existing Sparse Linear Discriminant Analysis Methods

3.1 Chapter Overview

In this chapter we reveal the connection and equivalence of three sparse linear discriminant analysis methods: the ℓ_1-Fisher's discriminant analysis (SFDA) proposed in [62], the sparse optimal scoring proposed in [72] and the direct sparse discriminant analysis proposed in [7]. It is shown that, for any sequence of penalization parameters, the normalized solutions of direct sparse discriminant analysis equal the normalized solutions of the other two methods at different penalization parameters. A prostate cancer dataset is used to demonstrate the theory.

3.2 Background

Along with the sparse LDA methods in Chapter 2 [7], there has been a sharp rise of interest in developing sparse LDA methods in the last couple of years. An incomplete list includes [73, 62, 72, 63, 74, 75, 76]. A review of these methods is [77]. These methods respect the correlations between variables, which distinguishes them from FAIR and PAM, are able to estimate a sparse discriminant rule, and have impressive numerical

performance. Among these methods, rigorous theories have been established for the methods in [7, 74, 75, 76].

In this chapter we prove the equivalence of the ℓ_1-Fisher's discriminant analysis (SFDA) [62, 76], the direct sparse discriminant analysis (DSDA) [7], and SOS [72]. We then present the main theorems concerning the connection and the equivalence between the methods. Finally, a prostate cancer dataset is used to demonstrate the theory. A direct consequence of the theory is that we can directly apply the theoretical results in [7] to justify SOS and SFDA.

3.3 Linear Programming Discriminant

The linear programming discriminant (LPD) rule [75] aims to find a sparse approximation of $\boldsymbol{\beta}^{\text{Bayes}}$. It is derived from the observation that

$$\boldsymbol{\Sigma}\boldsymbol{\beta}^{\text{Bayes}} = \boldsymbol{\mu}_- - \boldsymbol{\mu}_+. \tag{3.1}$$

Hence, LPD estimates $\boldsymbol{\beta}^{\text{Bayes}}$ by

$$\hat{\boldsymbol{\beta}}^{\text{LPD}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1, \text{ s.t. } \|\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - (\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)\|_\infty \leq \lambda. \tag{3.2}$$

The resulting $\hat{\boldsymbol{\beta}}^{\text{LPD}}$ is typically sparse. Note that (3.2) is similar to Dantzig selector [5] in the regression context. Indeed, $\hat{\boldsymbol{\beta}}^{\text{LPD}}$ can be found by the primal-dual interior-point method [78] as the Dantzig selector.

By assuming $\pi_+ = \pi_-$, LPD classifies an observation to Class -1 if

$$\left(\mathbf{X} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_+ + \hat{\boldsymbol{\mu}}_-)\right)^{\text{T}} \hat{\boldsymbol{\beta}}^{\text{LPD}} > 0.$$

For theoretical consideration, LPD is consistent in the sense that the error rate of LPD will tend to the Bayes error rate under proper regularity conditions.

3.4 The ℓ_1-Fisher's Discriminant Analysis

the ℓ_1-Fisher's discriminant analysis (SFDA) generalizes LDA from a different angle from LPD. It intends to minimize the classification error rate. As in LPD, SFDA

assumes that $\pi_+ = \pi_- = 1/2$. Then for any $\tilde{\boldsymbol{\beta}}$, the expected error rate is

$$R(\tilde{\boldsymbol{\beta}}) = 1 - \Phi\left(\frac{\tilde{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)}{2(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}})^{1/2}}\right). \tag{3.3}$$

Therefore, by minimizaing $\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}$ subject to $\tilde{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)/2 = 1$, we can minimize the expected error rate.

In order to encourage sparsity in high-dimensional datasets, SFDA uses the following formula to estimate the direction.

$$\hat{\boldsymbol{\beta}}^{\mathrm{SFDA}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta}, \text{ s.t. } (\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\boldsymbol{\beta}/2 = 1 \text{ and } \sum_{j=1}^{p}|\beta_j| \le \tau. \tag{3.4}$$

The constraint $\sum_{j=1}^{p}|\beta_j| \le \tau$ usually leads to sparse $\hat{\boldsymbol{\beta}}^{\mathrm{SFDA}}$. SFDA is shown to be able to asymptotically achieve the Bayes error rate as n tends to infinity, without extra conditions on $\boldsymbol{\Sigma}$. In the case that we have prior knowledge of $\boldsymbol{\Sigma}$, we can directly substitute an appropriate estimator in (3.4). With $\hat{\boldsymbol{\beta}}^{\mathrm{SFDA}}$, the prediction is

$$\hat{Y} = \mathbf{1}\left((\mathbf{X} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_+ + \hat{\boldsymbol{\mu}}_-))^{\mathrm{T}}\hat{\boldsymbol{\beta}}^{\mathrm{SFDA}} > 0\right) + 1.$$

For implementation, SFDA approximates (3.4) by

$$\hat{\boldsymbol{\beta}}_\gamma^{\mathrm{SFDA}} = \arg\min \frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1 + \frac{1}{2}\gamma(\boldsymbol{\beta}^{\mathrm{T}}(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)/2 - 1)^2. \tag{3.5}$$

Ideally, if $\gamma \to \infty$, $\hat{\boldsymbol{\beta}}_\gamma^{\mathrm{SFDA}} \to \hat{\boldsymbol{\beta}}^{\mathrm{SFDA}}$. In practice, $\hat{\boldsymbol{\beta}}_\gamma^{\mathrm{SFDA}}$ is insensitive to the choice of γ as long as it is reasonably large. For a fixed γ, SFDA uses coordinate descent to compute $\hat{\boldsymbol{\beta}}_\gamma^{\mathrm{SFDA}}$ [79, 80]. The matlab code for SFDA can be found at

`http://www.mathworks.com/matlabcentral/fileexchange/40047`

Also, SFDA is theoretically justified. The original paper provides the rates of convergence for both $\hat{\boldsymbol{\beta}}^{\mathrm{SFDA}}$ and the error rate.

A slightly different formula was proposed by [62]:

$$\tilde{\boldsymbol{\beta}}^{\mathrm{SFDA}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta}, \text{ s.t. } (\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{T}\boldsymbol{\beta} = 1 \text{ and } \|\boldsymbol{\beta}\|_1 \le \tau, \tag{3.6}$$

for some tuning parameter $\tau > 0$. For the same purpose as in SFDA, [62] added an ℓ_1 constraint to (3.7). Their implementation is available at:

`http://www.bios.unc.edu/~mwu/software/sLDA/SLDAPathway.R`

3.5 Direct Sparse Discriminant Analysis

[7] developed the direct sparse discriminant analysis by taking advantage of a least squares formulation of linear discriminant analysis. Let $y_i = -n/n_+$ if $Y^i = +1$ and $y_i = n/n_-$ if $Y^i = -1$. Define the solution to DSDA as follows

$$\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{X}^i\boldsymbol{\beta})^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Note that DSDA is actually Lasso-DSDA in Chapter 2. For convenience, we always refer to Lasso-DSDA as DSDA from now on. [7] showed that DSDA can recover the support of the Bayes rule and estimate the Bayes classifier direction with an overwhelming probability, even when the dimension grows with the sample size at a non-polynomial rate. DSDA is computationally most efficient among the three methods. One can solve $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda)$ for all values of λ using the `lars` package [25] or solve $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda)$ for a fine grid values of λ using the `glmnet` package [26].

3.6 Witten's Method

All the methods discussed above are designed for binary classification except for PAM. Now we introduce two methods that deal with K-class problems, where $K \geq 2$. We start with Witten's proposal in [63]. Define $\mathbf{Y}^{\mathrm{dm}}$ as an $n \times K$ matrix of dummy variables with $\mathbf{Y}^{\mathrm{dm}}_{ik} = 1(Y_i = k)$. Witten's method regularizes a variant of Fisher's discriminant analysis:

$$\hat{\boldsymbol{\beta}}_k = \arg\max_{\boldsymbol{\beta}_k} \boldsymbol{\beta}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_b^k \boldsymbol{\beta}_k \text{ s.t. } \boldsymbol{\beta}_k^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}} \boldsymbol{\beta}_k \leq 1, \tag{3.7}$$

where

$$\hat{\boldsymbol{\Sigma}}_b^k = \mathbf{X}^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}}((\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}})^{-1/2}\boldsymbol{\Omega}_k((\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}})^{-1/2}(\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{X} \tag{3.8}$$

and $\boldsymbol{\Omega}_k$ is the identity matrix if $k = 1$ and otherwise an orthogonal projection matrix with column space orthogonal to $((\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{Y})^{-1/2}\mathbf{Y}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}}_l$ for all $l < k$.

To generalize (3.7) to high dimensions, Witten's method estimates the discriminant directions by

$$\hat{\boldsymbol{\beta}}_k = \arg\max_{\boldsymbol{\beta}_k} \boldsymbol{\beta}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_b^k \boldsymbol{\beta}_k \text{ s.t. } \boldsymbol{\beta}_k^{\mathrm{T}} \tilde{\boldsymbol{\Sigma}} \boldsymbol{\beta}_k \leq 1, \tag{3.9}$$

where

$$\hat{\mathbf{\Sigma}}_b^k = \mathbf{X}^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}}((\mathbf{Y}^{\mathrm{dm}})^T\mathbf{Y}^{\mathrm{dm}})^{-1/2}\mathbf{\Omega}_k((\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}})^{-1/2}(\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{X} \tag{3.10}$$

$$\hat{\boldsymbol{\beta}}_k^{\mathrm{FSDA}} = \arg\max_{\boldsymbol{\beta}_k} \boldsymbol{\beta}_k^{\mathrm{T}}\hat{\mathbf{\Sigma}}_b\boldsymbol{\beta}_k - P_{\lambda_k}(\boldsymbol{\beta}_k),\ \text{s.t. } \boldsymbol{\beta}_k^{\mathrm{T}}\tilde{\mathbf{\Sigma}}\boldsymbol{\beta}_k \le 1, \tag{3.11}$$

with $\tilde{\mathbf{\Sigma}}$ being a positive definite estimation of $\mathbf{\Sigma}$, such as $\hat{\mathbf{D}}$ in IR, and $P_{\lambda_k}(\boldsymbol{\beta}_k)$ a penalty function such as Lasso and fused Lasso.

The estimator $\hat{\boldsymbol{\beta}}^{\mathrm{Witten}}$ can be found by iteration. First, we initialize (3.11) with $\hat{\boldsymbol{\beta}}^{(0)}$ equal to the first eigenvector of $\tilde{\Sigma}^{-1}\hat{\Sigma}_b$. Then one can iteratively solve for $\boldsymbol{\beta}^{(m)}$ as follows

$$\boldsymbol{\beta}_k^{(m)} = \max_{\boldsymbol{\beta}_k}\{2\boldsymbol{\beta}_k\hat{\mathbf{\Sigma}}_b^k\boldsymbol{\beta}_k^{(m-1)} - P_{\lambda_k}(\boldsymbol{\beta}_k)\}, \text{s.t. } \boldsymbol{\beta}^{\mathrm{T}}\tilde{\mathbf{\Sigma}}\boldsymbol{\beta}_k \le 1. \tag{3.12}$$

This is indeed the algorithm implemented in the `R` package `penalizedLDA`.

3.7 Sparse Optimal Scoring

Sparse optimal scoring (SOS) [72] also tackles multi-class problems. It adds a penalty to the optimal scoring formulation [81]. Again, $\mathbf{Y}^{\mathrm{dm}}$ is a matrix of dummy variables. Further define a K-dimensional vector $\boldsymbol{\theta}_k$ of scores. The optimal scoring problem is formulated as

$$\begin{aligned}(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\beta}}_k) &= \arg\min_{\boldsymbol{\theta}_k,\boldsymbol{\beta}_k}\sum_{i=1}^{n}(\mathbf{Y}^{\mathrm{dm}}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)^2 \\ \text{s.t.} &\quad \frac{1}{n}\boldsymbol{\theta}_k^{\mathrm{T}}(\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}}\boldsymbol{\theta}_k = 1, \boldsymbol{\theta}_k^{\mathrm{T}}(\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}}\boldsymbol{\theta}_l = 0, l < k.\end{aligned} \tag{3.13}$$

To perform variable selection, SOS combines the optimal scoring with ℓ_1 penalty:

$$\begin{aligned}(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\beta}}_k^{\mathrm{SOS}}) &= \arg\min_{\boldsymbol{\theta}_k,\boldsymbol{\beta}_k}\sum_{i=1}^{n}(\mathbf{Y}^{\mathrm{dm}}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)^2 + \lambda\|\boldsymbol{\beta}_k\|_1 \\ &\quad \text{s.t. } \frac{1}{n}\boldsymbol{\theta}_k^{\mathrm{T}}(\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}}\boldsymbol{\theta}_k = 1, \boldsymbol{\theta}_k^{\mathrm{T}}(\mathbf{Y}^{\mathrm{dm}})^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}}\boldsymbol{\theta}_l = 0, l < k.\end{aligned} \tag{3.14}$$

Because SOS assumes that $\mathbf{X}$ is centered, (3.14) does not involve the intercept term. Then SOS applies LDA to $(\mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1^{\mathrm{SOS}}, \ldots, \mathbf{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{K-1}^{\mathrm{SOS}})$.

SOS can also be solved by iterative algorithms, because for fixed $\boldsymbol{\beta}_k$, $\boldsymbol{\theta}_k$ can be easily found, and vice versa. Such an algorithm is implemented in the `R` package `sparseLDA`.

3.8 Theory

We are going to provide some results on the connection of DSDA, SFDA and SOS. For SFDA, we use the formula in (3.6). We first study the connection between $\hat{\boldsymbol{\beta}}^{\mathrm{SFDA}}(\lambda)$ and $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda)$. Note that, by definition, $\hat{\boldsymbol{\beta}}^{\mathrm{SFDA}}(\lambda)$ always satisfies the equality constraint in (3.6). Thus we consider a properly normalized $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda)$ defined as follows

$$\tilde{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda) = \frac{\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda)}{c_1(\lambda)}, \quad \text{where } c_1(\lambda) = (\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}} \hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda).$$

Theorem 3.8.1 *Given any fixed $\lambda > 0$, we have*

$$\tilde{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda) = \hat{\boldsymbol{\beta}}^{\mathrm{SFDA}}(\tilde{\lambda})$$

with $\tilde{\lambda} = \dfrac{\lambda}{n|c_1(\lambda)|}$.

Next we study the equivalence between the sparse optimal scoring and the direct sparse discriminant analysis.

Theorem 3.8.2 *Given any $\lambda > 0$, we have*

$$\hat{\boldsymbol{\beta}}^{\mathrm{SOS}}(\lambda) = \sqrt{\hat{\pi}_+ \hat{\pi}_-}\, \hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}\left(\frac{\lambda}{\sqrt{\hat{\pi}_+ \hat{\pi}_-}} \right),$$

where $\hat{\pi}_+ = n_+/n, \hat{\pi}_- = n_-/n$.

Theorems 1 and 2 can be used to provide strong theoretical support to the ℓ_1-Fisher's discriminant analysis and the sparse optimal scoring. [62] and [72] provided numerical examples to demonstrate the efficacy of their proposals but there was no theoretical result to explain why their methods work well. In [7] it has been shown that, under certain regularity conditions, if let λ be some properly chosen λ_n, then $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda_n)$ consistently recovers the support of the Bayes rule and estimates the Bayes rule coefficient. By Theorems 1 and 2, the ℓ_1-Fisher's discriminant analysis with $\lambda = \lambda_n/(n|c_1(\lambda_n)|)$ and the sparse optimal scoring with $\lambda = \sqrt{\hat{\pi}_+ \hat{\pi}_-}\lambda_n$ work as well as the Bayes rule asymptotically.

We would like to make a remark here that the above theorems are established for the binary classification setting. Binary classification has been the center of attention in the modern machine learning literature. For example, both support vector machines and

boosting were first proposed for solving binary classification problems. On the other hand, multi-class classification problems can be very different than the binary case. SFDA and DSDA do not have a direct multi-class generalization. SOS was proposed to solve multi-class classification and cover binary classification as a special case. Right now we do not know a good multi-class generalization of SFDA or DSDA that allows us to prove results like Theorems 1 and 2 for the multi-class setting.

3.9 A Numerical Example

Both Theorems 1 and 2 are exact finite sample results that hold for each given dataset. In this section we use the prostate cancer dataset [69, 42] to illustrate Theorems 1 and 2. This dataset contains the expression levels of 6033 genes, measured on 50 normal tissues and 52 prostate cancer tumors. We normalized the predictors such that each predictor has zero sample mean and unit sample variance. We took a fine grid of λ values and computed the corresponding $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda)$. We then computed $\tilde{\lambda}$ from those λs using the formula $\tilde{\lambda} = \lambda/(n|c_1(\lambda)|)$. For each $\tilde{\lambda}$ we computed $\hat{\boldsymbol{\beta}}^{\mathrm{SFDA}}(\tilde{\lambda})$ using the code of Dr. Wu. Figure 3.1 compares these two solutions and gives a graphical illustration of the equivalence result in Theorem 3.1. Numeric calculation confirms that the differences between the two panels of Figure 1 are indeed zero. Similarly, Figure 3.2 demonstrates the equivalence between the sparse optimal scoring and the direct sparse discriminant analysis. We took a fine grid of λ values and computed the corresponding $\hat{\boldsymbol{\beta}}^{\mathrm{SOS}}(\lambda)$ by using the R package sparseLDA (available at http://cran.r-project.org/web/packages/sparseLDA/index.html). For each λ we then computed $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}$ at $\lambda/\sqrt{\hat{\pi}_+\hat{\pi}_-}$ and multiplied it by $\sqrt{\hat{\pi}_+\hat{\pi}_-}$. Once again, numeric calculations confirm exact equality of the two curves in Figure 3.2, demonstrating that Theorem 2 does indeed hold for this example.

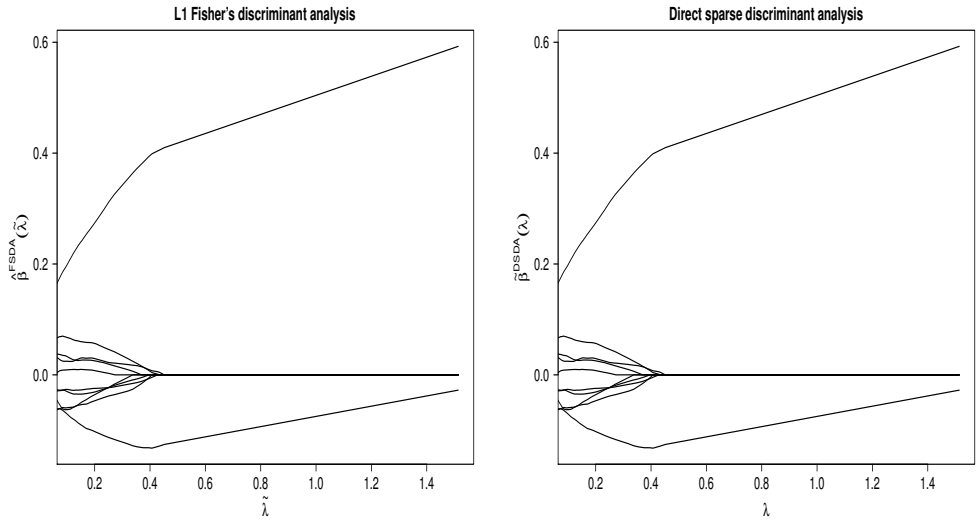

Figure 3.1: Using prostate cancer data to demonstrate Theorem 3.1. We have computed 6033 coefficient curves but only show 10 curves here for ease of presentation. $\tilde{\lambda} = \lambda/(n|c_1(\lambda)|)$.

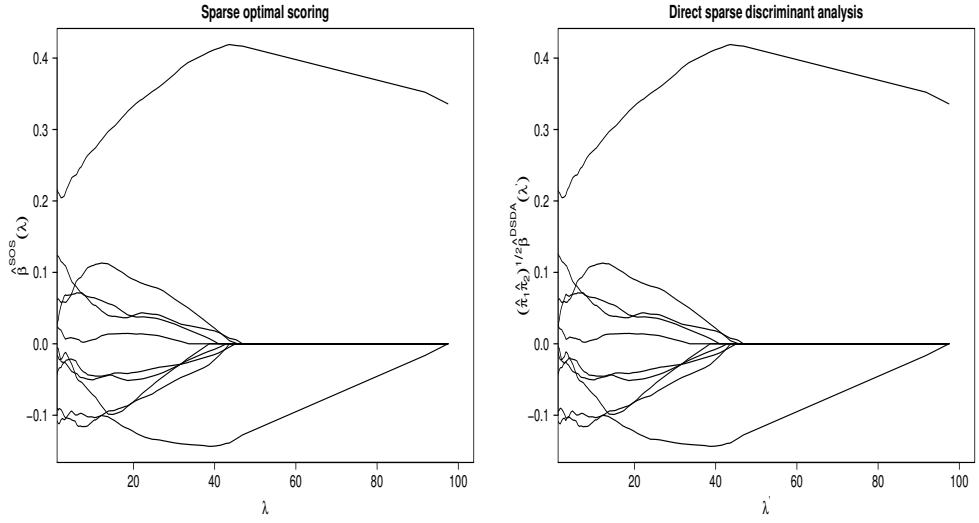

Figure 3.2: Using prostate cancer data to demonstrate Theorem 2. We have computed 6033 coefficient curves but only show 10 curves here for ease of presentation. $\lambda' = \lambda/\sqrt{\hat{\pi}_1\hat{\pi}_2}$.

Chapter 4

Semiparametric Sparse Discriminant Analysis

4.1 Chapter Overview

In this chapter, we develop high-dimensional semiparametric sparse discriminant analysis (SeSDA) that generalizes the normal-theory discriminant analysis in two ways: it relaxes the Gaussian assumptions and can handle non-polynomial (NP) dimension classification problems. If the underlying Bayes rule is sparse, SeSDA can estimate the Bayes rule and select the true features simultaneously with overwhelming probability, as long as the logarithm of dimension grows slower than the cube root of sample size. Simulated and real examples are used to demonstrate the finite sample performance of SeSDA. At the core of the theory is a new exponential concentration bound for semiparametric Gaussian copulas, which is of independent interest.

4.2 Introduction

In Chapters 2 and 3, we have introduced many proposals for sparse LDA methods that can perform variable selection for high-dimensional data. However, the existing sparse LDA methods still require the Gaussian data assumption, at least in theory. Empirical evidence given in Section 4.6.1 shows that sparse LDA methods become ineffective for non-normal data. In the lower dimensional classification problems, some researchers

have considered ways to relax the Gaussian distribution assumption. For example, [82] proposed the mixture discriminant analysis (MDA) that uses a mixture of Gaussian distributions to model the conditional densities of variables given the class label. MDA is estimated by the Expectation-Maximization algorithm. MDA is a nonparametric generalization of LDA, but it is not clear how to further extend MDA to the high-dimensional classification setting. [83] proposed another interesting approach to relaxing the Gaussian data assumption in LDA. Their approach starts with the assumption that, through a set of unknown monotone univariate transformations, the observed data follow the LDA model, and hence the new model is called the semiparametric LDA model (SeLDA). [83] further showed that the unknown transformations can be accurately estimated and thus the SeLDA model can be estimated when p is fixed and n goes to infinity. With the consistently estimated transformation, one can transform the data and fit a LDA model. However, the estimator in [83] cannot handle high-dimensional classification problems, especially when p exceeds n.

In this chapter, we develop high-dimensional semiparametric sparse discriminant analysis (SeSDA), a generalization of SeLDA for high-dimensional classification and variable selection. In particular, we propose a new estimator for the transformation function and establish its uniform consistency property as long as the logarithm of p is smaller than the cube root of n. With the new transformation estimator, we can transform the data and fit a sparse LDA classifier. In this work we use the direct sparse discriminant analysis (DSDA) developed by [7]. SeSDA enjoys great computational efficiency: its computational complexity grows linearly with p. We show that, if the Bayes rule of the SeLDA model is sparse, then SeSDA can consistently select the important variables and estimate the Bayes rule. At the core of the theory is an exponential concentration bound for semiparametric Gaussian copulas, which is of independent interest.

The rest of this chapter is organized as follows. The semiparametric LDA model is introduced in Section 4.3, and the methodological details of SeSDA are introduced in Section 4.4. Statistical theory is presented in Section 4.5. Numerical examples are shown in Section 4.6 to demonstrate the finite sample performance of SeSDA. Technical proofs are relegated to the appendix.

4.3 Semiparametric LDA Model

Consider the binary classification problem where we have observed n random pairs $(Y^i, X^i), 1 \leq i \leq n$ and wish to classify Y using a function of X. [83] proposed the following semiparametric LDA (SeLDA) model that assumes that

$$(h_1(X_1), \cdots, h_p(X_p)) \mid Y \sim N(\mu_Y, \mathbf{\Sigma}),$$

where $h = (h_1, \cdots, h_p)$ is a set of strictly monotone univariate transformations. It is important to note that the SeLDA model does not assume that these univariate transformations are known or have any parametric forms. Because of this nonparametric flavor, [84] used this model to discuss the sure property of a nonparametric marginal screening method, the so-called Kolmogorov filter. By properties of the Gaussian distribution, h is only unique up to location and scale shifts. Therefore, for identifiability, assume that $\mu_+ = 0$, $\mathbf{\Sigma}_{jj} = 1, 1 \leq j \leq p$. The Bayes rule of the SeLDA model is

$$\hat{Y}^{\text{Bayes}} = \text{sign}\left((h(X) - \frac{1}{2}(\mu_+ + \mu_-))^{\text{T}} \mathbf{\Sigma}^{-1}(\mu_+ - \mu_-) + \log \frac{\pi_+}{\pi_-} \right).$$

The SeLDA model is a very natural generalization of the LDA model. It is equivalent to modeling the within-group distributions with semiparametric Gaussian copulas. For any continuous univariate random variable, W, we have

$$\Phi^{-1} \circ F(W) \sim N(0, 1), \tag{4.1}$$

where F is the cumulative probability function (CDF) of W and Φ is the CDF of the standard normal distribution. For multivariate data, the SeLDA model can be viewed as an additive model by adding semiparametrically specified two-way interactions [83]. The additive model [85] assumes that the log-likelihood takes the following form:

$$\log f_y(X) = \psi_{y0} + \sum_{j=1}^{p} \psi_{yj}(X_j). \tag{4.2}$$

When interactions are not negligible, the additive model (4.2) may not be sufficient. SeLDA tries to model the interaction effects. Define $\Omega = (\omega_{jk}) = \mathbf{\Sigma}^{-1}$ and then the SeLDA model can be written as

$$\log f_y(X) \quad = \quad \psi_{y0} + \sum_{j=1}^{p} \psi_{yj}(X_j) + \sum_{j \neq k} \psi_{y,jk}(X_j, X_k), \tag{4.3}$$

where

$$\begin{aligned}\psi_{yj}(X_j) &= -\frac{\omega_{jj}(h_j(X_j)-\mu_{yj})^2}{2}+\log|h_j'(X_j)|,\\ \psi_{y,jk}(X_j,X_k) &= -\omega_{jk}(h_j(X_j)-\mu_{yj})(h_k(X_k)-\mu_{yk}).\end{aligned}$$

In the SeLDA model, the main effects in (4.3) are as general as those in the additive model (4.2), because any univariate continuous random variable follows the semiparametric normal distribution, while the two-way interactions are semiparametrically specified to strike a balance between flexibility and computation cost. For more detailed discussion on the connection between the semiparametric model and nonparametric models, the readers are referred to [83].

In light of (4.1), the SeLDA model can be estimated in the low-dimensional setting. The basic idea is straightforward: we first find $\hat{h}_j(\cdot)$ as good estimates of these univariate transformation functions and then fit the LDA model on the "pseudo data" $\left(Y^i, \hat{h}(X^i)\right), 1 \le i \le n$. To be more specific, in seek of $\hat{h}_j$, we let F_{+j}, F_{-j} be the CDF of X_j conditional on $Y=+1$ and $Y=-1$, respectively, and then we have

$$h_j = \Phi^{-1} \circ F_{+j} = \Phi^{-1} \circ F_{-j} + \mu_-.$$

It can be seen that we only need an estimate of F_{+j}. For convenience, denote X_{yj} as the jth entry of an observation X belonging to the group $Y=y$, and $\tilde{F}_{+j}$ as the empirical CDF of X_{+j}. Note that, we cannot directly plug in $\tilde{F}_{+j}$ so that $\hat{h}_j = \Phi^{-1} \circ \tilde{F}_{+j}$, because infinite values would occur at tails. Instead, $\tilde{F}_{+j}$ is Winsorized at a predefined pair of numbers (a,b) to obtain $\hat{F}_{+j}^{a,b}$

$$\hat{F}_{+j}^{a,b}(x) = \begin{cases} b & \text{if } \tilde{F}_{+j}(x) > b; \\ \tilde{F}_{+j}(x) & \text{if } a \le \tilde{F}_{+j}(x) \le b; \\ a & \text{if } \tilde{F}_{+j}(x) < a. \end{cases} \tag{4.4}$$

Then

$$\hat{h}_j = \Phi^{-1} \circ \hat{F}_{+j}^{a,b}. \tag{4.5}$$

The Winsorization can be viewed as a bias-variance trade-off. If necessary, one could interchange the group labels so that $n_+ > n_-$ to obtain more efficient estimates, where n_y is with-in-group sample size.

With $\hat{h}_j$, the covariance matrix $\mathbf{\Sigma}$ is estimated by the pooled sample covariance matrix of $\hat{h}(X^i)$ and μ_{-j} is estimated by

$$\begin{aligned}\hat{\mu}_{-j} &= q^{-1}\Big(\frac{1}{n_-}\sum_{i=1}^{n_-}\hat{h}(X^i_{-j})1_{\tilde{F}(X^i_{-j})\in(a,b)} + \phi(\Phi^{-1}\circ\tilde{F}_{-j}\circ\tilde{F}^{-1}_{+j}(b)) \\ &\quad -\phi(\Phi^{-1}\circ\tilde{F}_{-j}\circ\tilde{F}^{-1}_{+j}(a))\Big)\end{aligned}$$

where ϕ is the density function for a standard normal random variable and

$$q = \frac{1}{n_-}\sum_{i=1}^{n_-}1_{\tilde{F}_{+j}(X^i)\in(a,b)}.$$

$\hat{\mu}_{-j}$ has this complicated form because of the Winsorization. [83] showed that when p is fixed and n tends to infinity, $\hat{\mathbf{\Sigma}}$, $\hat{\mu}_-$ are consistent.

4.4 Estimation of The High-dimensional Semiparametric LDA Model

There are two fundamental difficulties in applying SeLDA to high-dimensional classification. First, [83] justified their estimates of the transformation functions but their asymptotic theory only works for the fixed p setting. We show later that, in order to obtain good estimators of p transformation functions uniformly, we shall modify the estimators defined in (4.4) and (4.5). Second, the second stage of SeLDA estimation is just the ordinary LDA, which is infeasible for high-dimensional problems, even when we know the true transformation functions. To overcome this difficulty, we propose to fit a sparse SeLDA model by exploiting sparsity assumption on the Bayes rule. For the sake of presentation, we first discuss how to fit a sparse SeLDA model, provided that good estimators of $h_j(\cdot), 1 \le j \le p$, are already obtained. After introducing the sparse SeLDA, we focus on a new strategy to estimate $h_j(\cdot), 1 \le j \le p$.

4.4.1 Exploiting sparsity

We assume that the Bayes rule of the SeLDA model only involves a small number of predictors. To be more specific, let $\boldsymbol{\beta}^{\text{Bayes}} = \mathbf{\Sigma}^{-1}(\mu_+ - \mu_-)$ and define $A = \{j :$

$\boldsymbol{\beta}_j^{\text{Bayes}} \neq 0\}$. Sparsity means that $|A| \ll p$. An elegant feature of SeLDA is that it keeps the interpretation of LDA, that is, variable j is irrelevant if and only if $\beta_j^{\text{Bayes}} = 0$.

Suppose that we have obtained $\hat{h}_j(\cdot)$ as a good estimate of $h_j(\cdot), 1 \leq j \leq p$, we focus on estimating the sparse LDA model using the "pseudo data" $(Y^i, \hat{h}(X^i)), 1 \leq i \leq n$. As discussed in Chapter 3, there have been several sparse LDA proposals in the literature. We do not want to assume $\boldsymbol{\Sigma}$ in the SeLDA model is a diagonal matrix. Thus we do not use NSC or FAIR in this work, although they may perform quite well in some applications. The sparse LDA in [74] can take care of the correlations only under some strong sparsity assumptions on the covariance matrix $\boldsymbol{\Sigma}$. Another sparse LDA algorithm was proposed by [73], but their paper focused on the $p < n$ senario. Sparse LDA algorithms that can handle general correlation structures in high dimensions include [75], [72], [76], [7], [62] and [63].

Also as pointed out in Chapter 3, many of these proposals are deeply connected. For example, [9] showed that SOS, sLDA and DSDA are equivalent in the sense that for proper sequences of λ, they give the same set of directions. [63] proved that a critical point for SFDA must also be a critical point for SOS, if SOS uses the same $\tilde{\boldsymbol{\Sigma}}$ as in SFDA. By [86], DSDA with lasso penalty can be equivalently written as

$$\hat{\boldsymbol{\beta}}^{\text{DSDA}} = \arg\min_{\boldsymbol{\beta}} \sum_{j=1}^{p} |\beta_j|, \text{ s.t. } \|\hat{\boldsymbol{\Sigma}}_{h(X)}\boldsymbol{\beta} - (\hat{\mu}_+ - \hat{\mu}_-)\|_\infty \leq \lambda, \tag{4.6}$$

where $\hat{\boldsymbol{\Sigma}}_{h(X)}$ is the sample covariance of $h(X)$, which closely resembles LPD. Therefore, we can stick with one sparse LDA method to illustrate SeSDA to fix ideas. In this dissertation, we use DSDA.

If we knew these transformation functions h in the SeLDA model, (**??**) could be directly used to estimate the Bayes rule of SeLDA. In SeSDA we substitute h_j with its estimator $\hat{h}_j$ and apply sparse LDA methods to $(Y, \hat{h}(X))$. For example, to use DSDA in the SeLDA model, we solve for

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \{n^{-1} \sum_{i=1}^{n} \left(Y^i - \boldsymbol{\beta}_0 - \hat{h}(X^i)^{\mathrm{T}}\boldsymbol{\beta}\right)^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|)\}, \\ \hat{\boldsymbol{\beta}}_0 &= -\frac{(\hat{\mu}_+ + \hat{\mu}_-)^{\mathrm{T}}\hat{\boldsymbol{\beta}}}{2} + \frac{\hat{\boldsymbol{\beta}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\beta}}}{(\hat{\mu}_+ - \hat{\mu}_-)^{\mathrm{T}}\hat{\boldsymbol{\beta}}} \log\frac{\hat{\pi}_+}{\hat{\pi}_-}. \end{aligned} \tag{4.7}$$

Then (4.7) yields the SeSDA classification rule: sign $\left(\hat{\boldsymbol{\beta}}_0 + \hat{h}(X)^{\mathrm{T}}\hat{\boldsymbol{\beta}}\right)$.

4.4.2 Uniform estimation of transformation functions

We propose a high-quality estimator of the monotone transformation function. In order to establish the theoretical property of SeSDA, we need all p estimators of the transformation function to uniformly converge to the truth at a certain fast rate, even when p is much larger than n. Our estimator is defined as

$$\hat{F}_{+j}(x) = \begin{cases} 1 - \frac{1}{n_+^2} & \text{if } \tilde{F}_{+j}(x) > 1 - \frac{1}{n_+^2} \\ \tilde{F}_{+j}(x) & \text{if } \frac{1}{n_+^2} \leq \tilde{F}_{+j}(x) \leq 1 - \frac{1}{n_+^2} \\ \frac{1}{n_+^2} & \text{if } \tilde{F}_{+j}(x) < \frac{1}{n_+^2} \end{cases} \tag{4.8}$$

and then

$$\hat{h}_j = \Phi^{-1} \circ \hat{F}_{+j}.$$

In other words, instead of fixing the Winsorization parameters a, b as in (4.4), we let

$$(a, b) = (a_n, b_n) = (\frac{1}{n_+^2}, 1 - \frac{1}{n_+^2}). \tag{4.9}$$

With the presence of Φ^{-1}, it is necessary to choose $a_n > 0, b_n < 1$ to avoid extreme values at tails. On the other hand, $a_n \to 0, b_n \to 1$ so that the bias will automatically vanish as $n \to \infty$. To further see that (4.9) are proper choices of a_n, b_n, see the theory developed in Section 3 for mathematical justification.

Other estimations have been proposed. For example, [87] considered a one-class problem with Gaussian copulas, which essentially states $h(X) \sim N(0, \mathbf{\Sigma})$, and aims to estimate $\mathbf{\Sigma}^{-1}$. In their paper, h_j is estimated by $\hat{h}_j = \Phi^{-1} \circ \hat{F}^{a_n, b_n}$, where $a_n = 1 - b_n = \frac{1}{4n^{1/4}\sqrt{\pi \log n}}$. [87] showed that this estimator is consistent when p is smaller than any polynomial order of n, but it is not clear whether the final SeSDA can handle non-polynomial high dimensions. Another rank-based estimator was independently proposed by [88, 89] that can handle non-polynomial dimensions in estimating $\mathbf{\Sigma}$. However, the difficulty still remains in estimating the mean vectors.

4.5 Theoretical Results

4.5.1 Estimation of transformation functions

To explore the consistency property of SeSDA, we first study the estimation accuracy of semiparametric Gaussian copulas. The results in this subsection are applicable to any statistical model using semiparametric Gaussian copulas, which is of independent interest itself. Consider the one-class estimation case first. Assume that X is a p-dimensional random variable such that $h(X) \sim N(0_p, \mathbf{\Sigma})$ with $h_j = \Phi^{-1} \circ F_j$ and $\hat{h}_j = \Phi^{-1} \circ \hat{F}_j$, where $\hat{F}_j$ is defined as in (4.8). Denote $\hat{\mu}_j$ and $\hat{\mathbf{\Sigma}}_{jk}$ as the sample mean and sample covariance for corresponding features. We establish exponential concentration bounds for $\hat{\mu}_j$ and $\hat{\Sigma}_{jk}$.

Theorem 4.5.1 *Define*

$$\begin{aligned}
\zeta_1^*(\epsilon) &= 2\exp(-cn\epsilon^2) + 4\exp(-cn^{1-\rho}\frac{\epsilon^2}{\rho}) + 4\exp(-cn^{\frac{1}{2}-\rho}) \\
\zeta_2^*(\epsilon) &= c\exp(-cn\epsilon^2) + c\exp(-cn^{\frac{1}{3}-\rho}) + c\exp(-cn^{1-\rho}) \\
&\quad + c\exp(-c\frac{n^{1-\rho}\epsilon^2}{\rho^2\log^2 n})
\end{aligned}$$

where c is a generic positive constant. For sufficiently large n and any $0 < \rho < \frac{1}{3}$, there exists a positive constant ϵ_0 such that, for any $0 < \epsilon < \epsilon_0$, we have

$$\Pr(|\hat{\mu}_j - \mu_j| > \epsilon) \leq \zeta_1^*(\epsilon) \tag{4.10}$$

$$\Pr(|\hat{\Sigma}_{jk} - \Sigma_{jk}| > \epsilon) \leq \zeta_2^*(\epsilon) \tag{4.11}$$

Remark 4.5.2 *Semiparametric Gaussian copulas have been used by [87] to develop a semiparametric graphical model. They derived some probability bounds concerning $\hat{\mu}_j$ and $\hat{\mathbf{\Sigma}}_{jk}$ as well, but they required that p should grow at a polynomial order of n. Our results are much stronger because p can be as large as $\exp(n^{\frac{1}{3}-\rho})$ for any $0 < \rho < \frac{1}{3}$. Theorem 4.5.1 and its proof can be used for other high-dimensional statistical problems involving semiparametric Gaussian copulas, including the penalized estimation of the aforementioned semiparametric graphical model by the graphical lasso [87].*

For the two-class SeLDA model, we can easily obtain the following corollary from Theorem 4.5.1.

Corollary 4.5.3 *Define*

$$\zeta_1(\epsilon) = \zeta_1^*(\frac{\sqrt{\pi_+}\epsilon}{2}) + \zeta_1^*(\frac{\sqrt{\pi_-}\epsilon}{2}) + 4\exp(-cn) \tag{4.12}$$

$$\zeta_2(\epsilon) = \zeta_2^*(\frac{\sqrt{\pi_+}\epsilon}{2}) + \zeta_2^*(\frac{\sqrt{\pi_-}\epsilon}{2}) + 4\exp(-cn) + 2\zeta_1(\epsilon) \tag{4.13}$$

Then there exists a positive constant ϵ_0 such that, for any $0 < \epsilon < \epsilon_0$, we have

$$\Pr(|(\hat{\mu}_{+j} - \hat{\mu}_{-j}) - (\mu_{+j} - \mu_{-j})| > \epsilon) \leq \zeta_1(\epsilon)$$
$$\Pr(|\hat{\mathbf{\Sigma}}_{jk} - \mathbf{\Sigma}_{jk}| > \epsilon) \leq \zeta_2(\epsilon)$$

Corollary 1 is the fundamental result for establishing the rate of convergence of SeSDA.

4.5.2 Consistency of SeSDA

With the results in Section 4.1, we are ready to prove the rate of convergence of SeSDA. We first define necessary notation. Define $\boldsymbol{\beta}^\star = \mathbf{C}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$. Recall that $\boldsymbol{\beta}^\star$ is equal to $c\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-) = c\boldsymbol{\beta}^{\text{Bayes}}$ for some positive constant [7]. Then we can write $A = \{j : \beta_j^\star \neq 0\}$. Let s be the cardinality of A. In addition, for an $m_1 \times m_2$ matrix M, denote $\|M\|_\infty = \max_{i=1,..,m_1} \sum_{j=1}^{m_2} |M_{ij}|$, and, for a vector u, $\|u\|_\infty = \max |u_j|$. Throughout the proof, we assume that $s \ll n^{1/4}$. Define the following quantities that are repeatedly used:

$$\kappa = \|\mathbf{C}_{A^cA}(\mathbf{C}_{AA})^{-1}\|_\infty, \quad \varphi = \|(\mathbf{C}_{AA})^{-1}\|_\infty, \Delta = \|\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A}\|_\infty,$$
$$\Delta_1 = \|\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A}\|_1, \quad \Delta_2 = \|\boldsymbol{\mu}_{1A} + \boldsymbol{\mu}_{2A}\|_\infty \quad \nu = \frac{\min_{j\in A}|\beta_j|}{\Delta\varphi}.$$

Suppose that the lasso estimator correctly shrinks $\hat{\boldsymbol{\beta}}_{A^c}$ to zero, then SeSDA should be equivalent to performing SeLDA on X_A. Therefore, define the hypothetical estimator

$$\hat{\boldsymbol{\beta}}_A^{\text{hyp}} = \arg\min_{\boldsymbol{\beta},\beta_0}\{n^{-1}\sum_{i=1}^{n}\left(Y^i - \beta_0 - \sum_{j\in A}\hat{h}_j(X_j^i)\beta_j\right)^2 + \sum_{j\in A}\lambda|\beta_j|\}.$$

Then, we wish that $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_A^{\text{hyp}}, 0_{A^c})$ with $\hat{\beta}_j \neq 0$ for $j \in A$. To ensure the consistency of SeSDA, we further require the following condition:

$$\kappa = \|\mathbf{C}_{A^cA}(\mathbf{C}_{AA})^{-1}\|_\infty < 1. \tag{4.14}$$

The condition in (4.14) is an analogue of the ir-representable condition for the lasso penalized linear regression model [90, 18, 32, 33].

Theorem 4.5.4 *Define ζ_1, ζ_2 as in Corollary 4.5.3. Pick any λ such that $\lambda < \min\{\frac{\min_{j\in A}|\beta_j|}{2\varphi}, \Delta\}$. Then for any $\epsilon > 0$ and sufficiently large n such that $\epsilon > Csn^{-1/4}$, where C does not depend on (n, p, s), we have*

1. Assuming the condition in (4.14), with probability at least $1 - \psi_1$, $\hat{\boldsymbol{\beta}}_A = \hat{\boldsymbol{\beta}}_A^{\text{hyp}}$ and $\hat{\boldsymbol{\beta}}_{A^c} = 0$, where

$$\psi_1 = 2ps\zeta_2(\frac{\epsilon}{s}) + 2p\zeta_1(\frac{\lambda(1-\kappa-2\epsilon\varphi)}{4(1+\kappa)})$$

and ϵ is any positive constant less than $\min\left\{\epsilon_0, \frac{\lambda(1-\kappa)}{4\varphi\,(\lambda/2+(1+\kappa)\Delta)}\right\}$.

2. With probability at least $1 - \psi_2$, none of the elements of $\hat{\boldsymbol{\beta}}_A$ is zero, where

$$\psi_2 = 2s^2\zeta_2(\frac{\epsilon}{s}) + 2s\zeta_1(\epsilon)$$

and ϵ is any positive constant less than $\min\left\{\epsilon_0, \frac{\nu}{(3+\nu)\varphi}, \frac{\Delta\nu}{6+2\nu}\right\}$.

3. For any positive ϵ satisfying $\epsilon < \min\left\{\epsilon_0, \frac{\lambda}{2\varphi\Delta}, \lambda\right\}$, we have

$$\Pr(\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A\|_\infty \le 4\varphi\lambda) \quge 1 - 2s^2\zeta_2(\frac{\epsilon}{s}) - 2s\zeta_1(\epsilon).$$

Theorem 4.5.4 provides the foundation for asymptotic results. Assume the following two regularity conditions.

(C1). $n, p \to \infty$ and $\frac{s^2\log(ps)}{n^{\frac{1}{3}-\rho}} \to 0$, for some ρ in $(0, 1/3)$;

(C2). $\min_{j\in A}|\boldsymbol{\beta}_j| \gg \max\{sn^{-1/4}, \sqrt{\log(ps)\frac{s^2}{n^{\frac{1}{3}-\rho}}}\}$ for for some ρ in $(0, 1/3)$.

Condition (C1) restricts that p, s should not grow too fast comparing to n. However, p is allowed to grow faster than any polynomial order of n. Condition (C2) states that the important features should be sufficiently large such that we can separate them from the noises, which is a standard assumption in the literature of sparse recovery. The next theorem shows that SeSDA consistently recovers the Bayes rule of the SeLDA model.

Theorem 4.5.5 *Let* $\widehat{A} = \{j : \hat{\beta}_j \neq 0\}$. *Under conditions (C1) and (C2), if we choose* $\lambda = \lambda_n$ *such that* $\lambda_n \ll \min_{j\in A} |\boldsymbol{\beta}_j|$ *and* $\lambda_n \gg \sqrt{\log(ps)\frac{s^2}{n^{\frac{1}{3}-\rho}}}$, *and further assume* $\kappa < 1$, *then* $\Pr(\widehat{A} = A) \to 1$ *and* $\Pr\left(\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A\|_\infty \leq 4\varphi\lambda_n\right) \to 1$.

Remark 4.5.6 *Although penalized least squares is used for feature selection in SeSDA, Theorems 2 and 3 are fundamentally different from the theoretical results of lasso penalized regression [32, 33], because the previous work is built on the linear regression model* $Y = X\boldsymbol{\beta} + \epsilon$ *with* ϵ *being independent normal or sub-Gaussian and this model is obviously not true for* $(Y^i, h_j(X_j^i))$ *or* $(Y^i, \hat{h}_j(X_j^i))$.

Further, we prove that SeSDA is asymptotically equivalent to the Bayes rule in terms of error rate as n tends to infinity. Define the Bayes error rate $R = \Pr(Y \neq \text{sign}(h(X)^{\mathrm{T}}\boldsymbol{\beta}^* + \boldsymbol{\beta}_0))$ and $R_n = \Pr(Y \neq \text{sign}(\hat{h}(X)^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}_0))$. Then we have the following theorem.

Theorem 4.5.7 *Define* ζ_1, ζ_2 *as in Corollary 4.5.3. Pick any* λ *such that* $\lambda < \min\{\frac{\min_{j\in A}|\beta_j|}{2\varphi}, \Delta\}$. *Then for a sufficiently small constant* $\epsilon > 0$ *and sufficiently large* n *such that* $\epsilon > Csn^{-1/4}$, *where* C *does not depend on* (n, p, s), *with probability no smaller than* $1 - \psi_3$, *we have* $R_n - R < \epsilon$, *where*

$$\psi_3 = Cs\zeta_1\left(\frac{\epsilon}{s(\phi\Delta_1 + \Delta_2)}\right) + Cp\zeta_1\left(\frac{\lambda(1-\kappa+2\epsilon\phi)}{4(1+\kappa)}\right) + 2ps\zeta_2\left(\frac{C\epsilon}{s}\right) + o(1). \tag{4.15}$$

Corollary 4.5.8 *Under conditions (C1) and (C2), if we choose* $\lambda = \lambda_n$ *such that* $\lambda_n \ll \min_{j\in A}|\boldsymbol{\beta}_j|$ *and* $\lambda_n \gg \sqrt{\log(ps)\frac{s^2}{n^{\frac{1}{3}-\rho}}}$, *and further assume* $\kappa < 1$, *then*

$$R_n - R \to 0 \quad \text{in probability} \tag{4.16}$$

Remark 4.5.9 *Our results concerning the error rate of SeSDA are much more involved than those for sparse LDA algorithms in [75, 76], because of the semiparametric assumptions. Under the parametric LDA model, the error rate tends to the Bayes error as long as the discriminant direction* $\boldsymbol{\beta}$ *is estimated consistently. However, under the SeLDA model, we deal with the extra uncertainty in estimating* h *and need some uniform convergence results on* $\hat{h}(X)$.

4.6 Numerical Results

4.6.1 Simulation

We examine the finite sample performance of SeSDA by simulation. For comparison, in the simulation study we also include DSDA and the sparse LDA algorithm [63] denoted by Witten for presentation purpose. After we apply the estimated transformation to the data, we use Witten's sparse LDA algorithm to fit the classifier. This gives us Se-Witten, another competitor in the simulation study.

Four types of SeLDA models were considered in the study. In each model, we first generated Y with $\pi_+ = \pi_- = 0.5$. We fixed $\mu_- = 0$ and $\mu_+ = \mathbf{\Sigma}\boldsymbol{\beta}^{\text{Bayes}}$.

Model 1: $n = 150$, $p = 400$. $\mathbf{\Sigma}$ has AR(0.5) structure.

$$\boldsymbol{\beta}^{\text{Bayes}} = 0.556(3, 1.5, 0, 0, 2, 0_{p-5})^{\text{T}}.$$

Model 2: $n = 200$, $p = 400$. $\mathbf{\Sigma}$ has AR(0.5) structure.

$$\boldsymbol{\beta}^{\text{Bayes}} = 0.582(3, 2.5, -2.8, 0_{p-3})^{\text{T}}.$$

Model 3: $n = 400$, $p = 800$. $\mathbf{\Sigma}$ has CS(0.5) structure.

$$\boldsymbol{\beta}^{\text{Bayes}} = 0.395(3, 1.7, -2.2, -2.1, 2.55, 0_{p-5})^{\text{T}}.$$

Model 4: $n = 300$, $p = 800$. $\mathbf{\Sigma}$ is block diagonal with 5 blocks of dimension 160×160. Each block has CS(0.6) structure.

$$\boldsymbol{\beta}^{\text{Bayes}} = 0.916(1.2, -1.4, 1.15, -1.64, 1.5, -1, 2, 0_{p-7})^{\text{T}}.$$

We transform V to X by $X = g(V)$ and the final data to be used are (X, Y). In each type of model, we consider two sets of g. We call the resulting models series a and b. In series a, $X = V$ so that the SeLDA model becomes the LDA model. In series b, we considered some commonly used transformations such that that some features become heavily skewed, some heavy-tailed and some bounded. The choices of g are listed in Table 4.1. In the simulation study we also considered the oracle sparse discriminant classifiers including oracle DSDA and oracle Witten. The idea is to apply the true

Table 4.1: Choices of g_j in Models 1b–4b.

$g_j(v)$	Models 1b,2b j	Model 3b j	Model 4b j
v^3	$1, 101, \ldots, 150$	$1, 201, \ldots, 300$	$3, 201, \ldots, 300$
$\exp(v)$	$2, 151, \ldots, 200$	$2, 301, \ldots, 400$	$4, 301, \ldots, 400$
$\arctan(v)$	$3, 201, \ldots, 300$	$3, 401, \ldots, 500$	$5, 401, \ldots, 500$
v^3	$4, \ldots, 50$	$4, 6, \ldots, 100$	$1, 8, \ldots, 100$
$\Phi(v)$	$51, \ldots, 100$	$5, 101, \ldots, 200$	$2, 101, \ldots, 200$
$(v+1)^3$	$301, \ldots, 350$	$501, \ldots, 600$	$6, 501, \ldots, 600$
$\arctan(2v)$	$351, \ldots, 400$	$601, \ldots, 800$	$7, 601, \ldots, 800$

transformation to variables and then fit a sparse LDA classifier using DSDA or Witten and Tibshirani's method.

The simulation results for Models 1a–4a and Models 1b–4b are reported in Table 4.2 and Table 4.3, respectively. Note that in Table 4.2 DSDA and Witten are the oracle DSDA and the oracle Witten. We can draw the following conclusions from Tables 4.2 and 4.3.

- Models 1a–4a are actually LDA models. SeSDA performs very similarly to DSDA. Although SeSDA has slightly higher error rates, this is expected because SeSDA does not use the parametric assumption. On the other hand, in Models 1b–4b, SeSDA performs much better than DSDA. These results jointly show that SeSDA is a much more robust sparse discriminant analysis algorithm than those based on the LDA model.

- In both tables, SeSDA is very close to the oracle DSDA, which empirically shows the high quality of the proposed transformation estimator in Section 4.3. In all eight cases, SeSDA is a good approximation to the Bayes rule, which is consistent with the theoretical results.

- Se-Witten is a different SeSDA classifier in which Witten and Tibshirani's method is used to fit the SeLDA model after estimating the transformation functions. Se-Witten performs very well in Models 1a,2a,1b,2b but it performs very poorly in

Models 3a,4a,3b,4b. The same is true for the oracle Witten method. By comparing SeSDA and Se-Witten, we see that DSDA works better than Witten and Tibshirani's method. In addition to the theory in Section 4.5, the simulation also supports the use of DSDA in fitting the high-dimensional sparse semiparametric LDA model.

4.6.2 Malaria data

We further demonstrate SeSDA by using the malaria data [91]. This dataset is available at

`http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2362.`

Out of 71 samples in the dataset, 49 have been infected with malaria, while 22 are healthy people. The predictors are the expression levels of 22283 genes. The 71 samples were split with 2:1 ratio to form training and testing sets. We report the median of 100 replicates in Table 4.4. Besides DSDA, the ℓ_1 logistic regression [26] was also considered because it is an obvious choice for sparse high-dimensional classification. From Table 4.4, it can be seen that SeSDA is slightly more accurate than DSDA and the ℓ_1 logistic regression. In addition, SeSDA selects 6 genes, while the other two methods select about 22 genes.

To gain more insight, we compared the selected genes by SeSDA and those by DSDA or ℓ_1 logistic regression. In those 100 tries the 2059th gene is most frequently selected by SeSDA, but seldom by DSDA or ℓ_1 logistic regression. This gene is encoded by `IRF1`, as it is the first identified interferon regulatory transcription factor (http://en.wikipedia.org/wiki/IRF1). Discovering the role of `IRF1` was a major finding in [91]. Previous studies show that `IRF1` influences the immune response. Therefore, healthy and sick people may have different expression levels on this gene. It is very interesting that we can use a pure statistical method like SeSDA to select `IRF1`. We plot in Figure 4.1 the within-group density functions of gene `IRF1` (the 2059th gene). It can be seen that the raw expression levels of `IRF1` are skewed, making linear rules unreliable on this gene. After applying the estimated transformation, the distributions of both groups become close to normal, with similar variances. The transformation

separates the two groups farther apart from each other, which helps explain the more accurate classification by SeSDA.

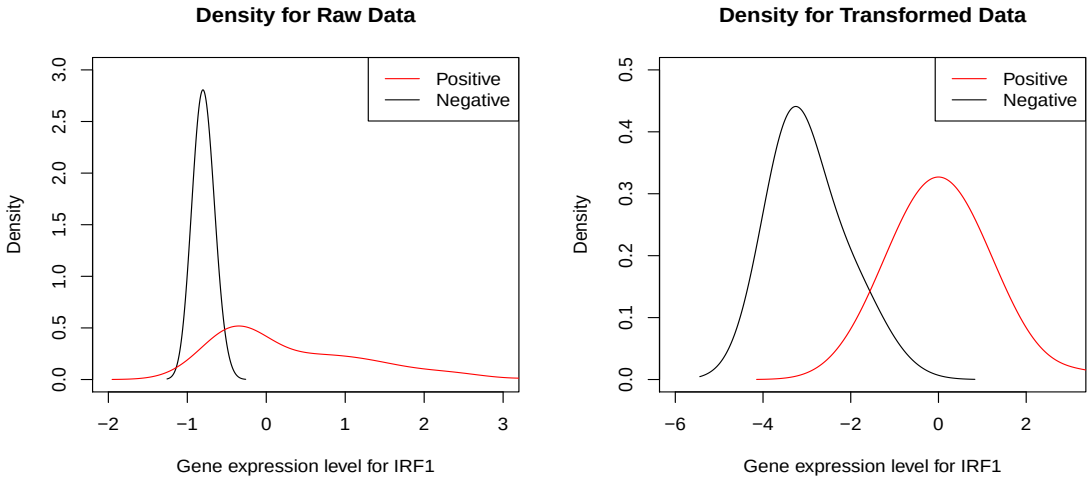

Figure 4.1: Density functions of gene `IRF1` (the 2059th gene) in the malaria data. The plot on the left displays the density function of the normalized raw data, while the one on the right is of the transformed data.

4.6.3 The Celiac dataset

The data celiac [92] has 132 samples and 18981 dimensions. In this dataset, 110 of the samples have celiac diseases, while the rest 22 do not. Again, we apply DSDA, SeSDA and ℓ_1-logistic regression on this dataset. The classification accuracies are as in Table 4.5. It can be seen that SeSDA is comparable to ℓ_1 logistic regression and is better than DSDA.

Moreover, the 2407th gene is frequently used by both methods. We plot its density in Figure 4.2. It is obvious that the raw expression level for this gene is not normal, since within each group, the distribution of the expression level is bimodal. However, after the transformation in SeSDA, the LDA model seems to roughly hold. Hence, it is evident that SeSDA is very useful in practice.

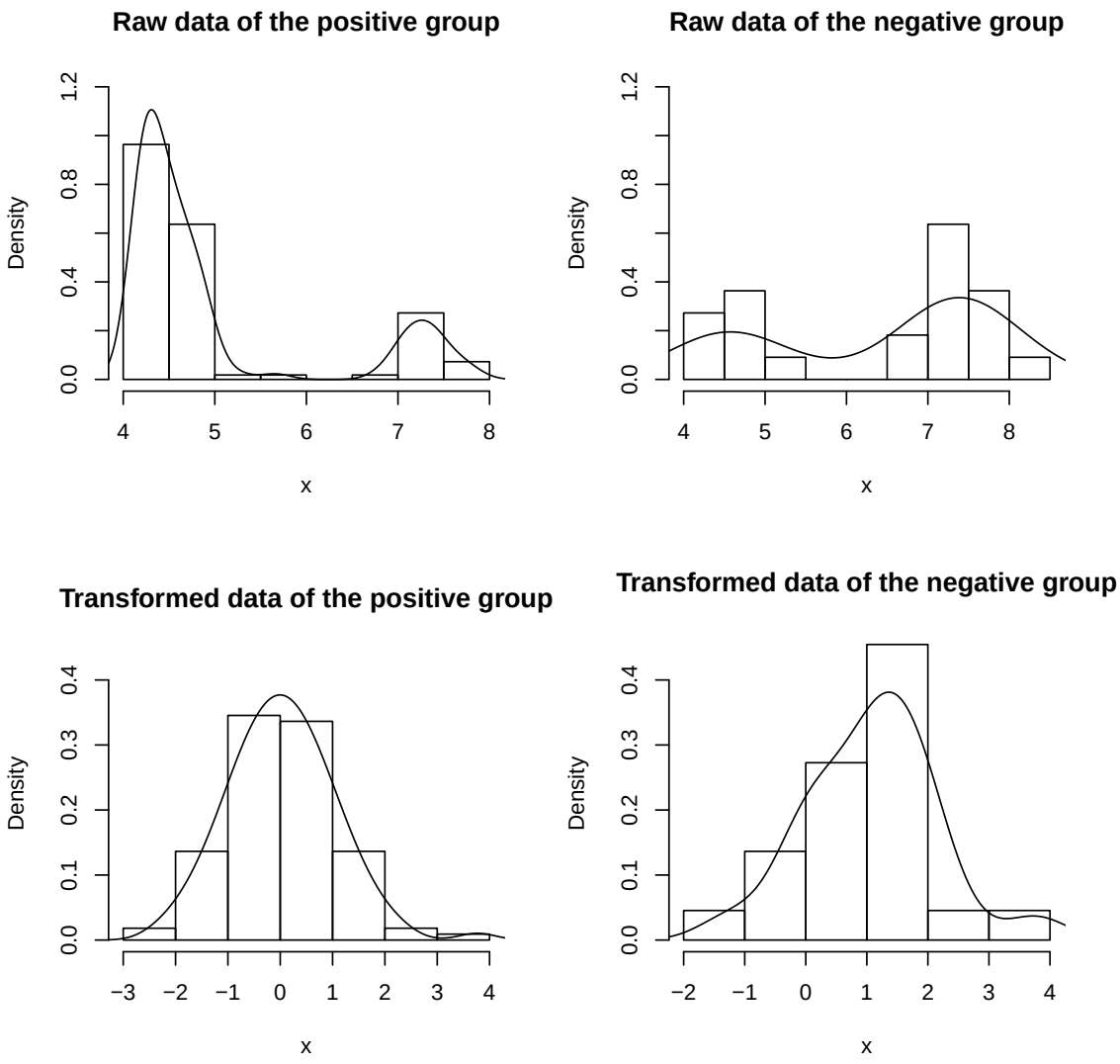

Figure 4.2: Density functions of the 2047th gene in the celiac data. The transformation in SeSDA makes the LDA roughly hold for this gene.

4.7 Discussion

It has been a hot subject of research in recent years to develop sparse discriminant analysis for high-dimensional classification and feature selection, rejuvenating the traditional discriminant analysis. However, sparse discriminant algorithms based on the LDA model can be very ineffective for non-normal data, as shown in the simulation study. To overcome the normality limitation, we consider the semiparametric discriminant analysis model and propose the SeSDA, a high-dimensional semiparametric sparse discriminant classifier. We have justified SeSDA both theoretically and empirically. For high-dimensional classification and feature selection, SeSDA is more appropriate than the existing sparse discriminant analysis proposals in the literature.

In order to demonstrate the main idea, we have only considered using the lasso penalty to do feature selection in SeSDA. In some specific applications, we wish to perform specific feature selection. For instance, we can use the group-lasso penalty [20] in DSDA to conduct group-wise variable selection when the groups are clearly defined; or we can apply the fused-lasso penalty [64] in DSDA in order to do fused feature selection.

Table 4.2: Simulation results for Models 1a–4a. The reported numbers are medians based on 2000 replications. Their standard errors obtained by bootstrap are in parentheses. TRUE selection and FALSE selection denote the numbers of selected important variables and unimportant variables, respectively.

	Bayes	Oracle DSDA	SeSDA	DSDA	Oracle Witten	Se-Witten	Witten
Model 1 (a)							
Error(%)	10	10.71	11.5	10.71	11.39	11.56	11.39
		(0.02)	(0.03)	(0.02)	(0.02)	(0.01)	(0.02)
TRUE selection	3	3	3	3	3	3	3
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	1	2	1	26	26	26
		(0.14)	(0.38)	(0.14)	(0.42)	(0.09)	(0.42)
Model 2 (a)							
Error(%)	10	11.09	11.66	11.09	13.36	13.46	13.36
		(0.02)	(0.03)	(0.02)	(0.03)	(0.04)	(0.03)
TRUE selection	3	3	3	3	3	3	3
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	5	6	5	24	24	24
		(0.37)	(0.51)	(0.37)	(0)	(0)	(0)
Model 3 (a)							
Error(%)	20	21.93	22.13	21.93	33.69	34.18	33.69
		(0.03)	(0.03)	(0.03)	(0.01)	(0)	(0.01)
TRUE selection	5	5	5	5	3	5	3
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	14	13	14	419.5	795	419.5
		(0.59)	(0.57)	(0.59)	(10.19)	(0)	(10.19)
Model 4 (a)							
Error(%)	10	12.50	13.20	12.50	23.90	26.14	23.90
		(0.02)	(0.05)	(0.02)	(0.01)	(0.01)	(0.01)
TRUE selection	7	7	7	7	4	5	4
		(0)	(0)	(0)	(0)	(0.02)	(0)
FALSE selection	0	18	17	18	35	153	35
		(0.70)	(0.54)	(0.70)	(4.43)	(0)	(4.43)

Table 4.3: Simulation results for Models 1a–4a. The reported numbers are medians based on 2000 replications. Their standard errors obtained by bootstrap are in parentheses. TRUE selection and FALSE selection denote the numbers of selected important variables and unimportant variables, respectively.

	Bayes	Oracle DSDA	SeSDA	DSDA	Oracle Witten	Se-Witten	Witten
Model 1 (b)							
Error(%)	10	10.71	11.42	18.24	11.39	11.56	16.19
		(0.02)	(0.04)	(0.10)	(0.01)	(0.02)	(0.05)
TRUE selection	3	3	3	3	3	3	3
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	1	2	2	26	26	25
		(0.14)	(0.42)	(0)	(0.42)	(0.09)	(0.50)
Model 2 (b)							
Error(%)	10	11.09	11.66	19.47	13.36	13.46	20.16
		(0.02)	(0.03)	(0.09)	(0.03)	(0.03)	(0.04)
TRUE selection	3	3	3	3	3	3	2
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	5	6	2	24	24	20
		(0.37)	(0.51)	(0)	(0)	(0.32)	(0.17)
Model 3 (b)							
Error(%)	20	21.93	22.13	26.76	33.69	34.18	34.25
		(0.03)	(0.03)	(0.03)	(0.01)	(0)	(0)
TRUE selection	5	5	5	5	3	5	5
		(0)	(0)	(0)	(0)	(0)	(0)
FALSE selection	0	14	13	15	419.5	795	795
		(0.59)	(0.57)	(0.67)	(10.19)	(0)	(0)
Model 4 (b)							
Error(%)	10	12.50	13.4	19.88	23.90	26.14	26.83
		(0.02)	(0.03)	(0.04)	(0.01)	(0.01)	(0.01)
TRUE selection	7	7	7	6	4	5	6
		(0)	(0)	(0)	(0)	(0.02)	(0.23)
FALSE selection	0	18	17	25	35	153	153
		(0.70)	(0.54)	(0.83)	(4.43)	(0)	(0.09)

Table 4.4: Comparison of SeSDA, DSDA and ℓ_1 logistic regression on the malaria dataset. The reported numbers are medians of 100 replicates, with standard errors obtained by bootstrap in parentheses.

	SeSDA	DSDA	Logistic
Testing Error	1/23(0%)	2/23(2.06%)	2/23(0.17%)
Fitted Model Size	6(0.86)	22.5(1.76)	23(0.83)

Table 4.5: Comparison of SeSDA, DSDA and ℓ_1 logistic regression on the celiac dataset. The reported numbers are medians of 100 replicates, with standard errors obtained by bootstrap in parentheses.

	SeSDA	DSDA	Logistic
Error	7/44(1.04%)	8/44(1.04%)	7/44(0)
Fitted Model Size	24(3.95)	18(3.47)	23(0.83)

Chapter 5

The Kolmogorov Filter

5.1 Chapter Overview

Variable screening techniques have been proposed to mitigate the impact of high dimensionality in classification problems, including t-test marginal screening [1] and maximum marginal likelihood screening [93]. However, these methods rely on strong modeling assumptions that are easily violated in real applications. To circumvent the parametric modeling assumptions, we propose a new variable screening technique for binary classification based on the Kolmogorov–Smirnov statistic. We prove that this so-called Kolmogorov filter enjoys the sure screening property under much weakened model assumptions. We supplement our theoretical study by a simulation study.

5.2 Motivation

5.2.1 Background

Binary classification problems concern predicting a class label based on the predictors. Let $Y = +1, -1$ be the class label. We use X to denote the predictor vector and $X \in \mathbb{R}^p$. The optimal classifier is the Bayes rule, predicting Y by

$$\text{sign}\left\{\log\left(\frac{f_+(\mathbf{X})}{f_-(\mathbf{X})}\right) + \log\left(\frac{\pi_+}{\pi_-}\right)\right\},$$

where $\pi_y = \Pr(Y = y)$ and $f_y(\mathbf{X})$ is the conditional density or probability mass function of X given $Y = y$. In high-dimensional classification, we also focus on discovering which predictors are responsible for the classification, besides achieving high classification accuracy. A canonical example is tumor classification with microarray data where we are interested in finding a few genes that contribute most to classification. A fundamental assumption there is that the underlying Bayes rule only depends on a few important variables $\{X_j, j \in D\}$, where D is a subset of $\{1, \ldots, p\}$ and its cardinality, d, is much smaller than the sample size, n, and the dimension, p. Sparse penalization techniques have been used to construct high-dimensional classifiers, including ℓ_1 penalized logistic regression (Section 4.4.4 in [13]) and the ℓ_1 penalized support vector machine [94]. These methods can perform variable selection and classification simultaneously.

In a series of papers, Fan and his co-authors advocated the use of variable screening in high-dimensional learning in a two-stage procedure [6, 1, 95, 93, 96]. First, a fast variable screening method is applied to reduce the dimension from p to d_n. Then a penalized model is fitted by using the reduced set of variables. In order for two-stage methods to be successful, the sure screening property must hold, i.e., the screening method employed in the first stage should keep all important variables [6]. With the sure screening property, the two-stage analysis could be computationally more efficient and yield better model estimation than one-stage penalization methods [6]. In the linear regression model [6] showed that, under suitable conditions, simple marginal correlation ranking has the sure screening property with overwhelming probability. For generalized linear models, [93] proposed maximum marginal likelihood screening. There are some recent papers on different sure screening techniques [38, 37, 97].

5.2.2 Marginal t-test screening and maximum marginal likelihood screening

[1] proposed marginal t-test screening based on the following independent normal model. For a diagonal matrix $\mathbf{\Sigma}$, assume that $\mathbf{X} \mid Y = y \sim N(\boldsymbol{\mu}_y, \mathbf{\Sigma})$. Under this model $D = \{j : \ \mu_{+j} \neq \mu_{-j}\}$. The t-test statistic for testing $\mu_{+j} = \mu_{-j}$ is given by $t_j = \hat{\sigma}_j^{-1}(\hat{\mu}_{+j} - \hat{\mu}_{-j})$, where $\hat{\sigma}_j^2$ is the pooled sample estimate of $\mathrm{var}(X_{yj})$ and $\hat{\mu}_{+j}, \hat{\mu}_{-j}$ are the sample mean estimates. Marginal t-test screening selects variables with $|t_j| > \nu_n$

where ν_n is a pre-defined threshold. [1] showed that it has the sure screening property with high probability.

[93] proposed maximum marginal likelihood screening (MMLE) under the logistic regression model

$$\log\left[q(x)/\{1-q(x)\}\right] = \beta_0^* + X^{\mathrm{T}}\beta^*,$$

where $q(x) = \Pr(Y = 1 \mid X = x)$. In this model $D = \{j : \beta_j^* \neq 0\}$. Recode Y by 0 and 1. For $j = 1, \ldots, p$, define

$$(\hat{\beta}_{0j}, \hat{\beta}_j^{\mathrm{M}}) = \arg\min_{(\beta_{0j},\beta_j)} n^{-1} \sum_{i=1}^{n} l(Y^i; \beta_{0j} + \beta_j X_j^i),$$

where

$$l(y; \beta_{0j} + \beta_j x_j) = \log\left(1 + e^{\beta_{0j}+\beta_j x_j}\right) - y(\beta_{0j} + \beta_j x_j), \quad y = 0, 1.$$

[93] rank variables according to the magnitudes of $\hat{\beta}_j^{\mathrm{M}}$ $(j = 1, \ldots, p)$, and established the sure screening property of maximum marginal likelihood screening under some other assumptions.

Although maximum marginal likelihood screening is more robust than marginal t-test screening, both depend on parametric modeling assumptions. Model mis-specification jeopardizes screening. In this chapter we introduce a new nonparametric screening method that is more robust and has wider applicability.

5.3 The Kolmogorov filter

5.3.1 Method

Let $F_{+j}(x)$ and $F_{-j}(x)$ denote the conditional cumulative probability functions of X_j given $Y = 1, -1$, respectively. Define

$$K_j = \sup_{-\infty<x<\infty} |F_{+j}(x) - F_{-j}(x)|.$$

The sample version of K_j is defined as

$$K_{nj} = \sup_{-\infty<x<\infty} |\hat{F}_{+j}(x) - \hat{F}_{-j}(x)|.$$

We rank all variables by the K_{nj} statistics. Because K_{nj} is the Kolmogorov–Smirnov test statistic for testing the equivalence of two distributions, we name the new screening method the Kolmogorov filter. By definition, this is invariant under any strictly monotone univariate transformations applied on individual variables. Such an invariance property is not shared by marginal t-test screening or by maximum marginal likelihood screening.

[96] were aware of the limitations of parametric model-based screening and proposed nonparametric independence screening for linear regression with generalized additive models. The idea there is to estimate $E(Y|X_j)$ non-parametrically using basis function expansion such as B-splines. Their idea can be generalized for variable screening in binary classification. Define

$$\frac{\Pr(Y = 1 \mid X_j = x)}{\Pr(Y = -1 \mid X_j = x)} = e^{m_j(x)}.$$

Maximum marginal likelihood screening amounts to assuming that $m_j(x)$ is a linear function of x. Using a non-parametric estimate of $m_j(x)$ in logistic regression can improve the robustness of maximum marginal likelihood screening.

It is interesting to compare the Kolmogorov filter and nonparametric maximum marginal likelihood screening. When using B-splines or other basis functions in nonparametric regression, a subtle but difficult problem is how to determine the number of basis functions to be used. For convenience, the user often uses the same number of basis functions for each component, as in [96], although the theoretical optimal choice and the probability of sure screening by nonparametric maximum marginal likelihood screening depend on the unknown smoothness of the individual $m_j(x)$; see Section 3 of [96]. In contrast, the Kolmogorov filter does not face these issues. Finally, as shown in Section 3, the Kolmogorov filter is almost as fast as t-test screening and is 10 times faster than nonparametric maximum marginal likelihood screening.

5.3.2 Sure screening property

We establish the sure screening property for the Kolmogorov filter in this section. The Dvoretzky–Kiefer–Wolfowitz inequality [98] plays a critical role in the theoretical analysis. See the proof in Appendix D.

We recommend using the Kolmogorov filter to select the subset

$$\hat{S}(d_n) = \{j : K_{nj} \text{ is amongst the first } d_n \text{ largest of all } K_{nj}s\}.$$

Obviously, $\Pr\{D \subseteq \hat{S}(d_n)\}$ is non-decreasing in d_n. Because it is often believed $|D| \ll n$, the default value for d_n is taken to be $\lceil n/\log(n) \rceil$, where $\lceil r \rceil = \min\{i : i \geq r \text{ and } i \text{ is an integer}\}$ for $r > 0$. A more conservative choice is $d_n = n$.

Theorem 5.3.1 *Define*

$$\zeta(\epsilon) = 2\{\exp(-\pi_+ n\epsilon^2/4) + \exp(-\pi_- n\epsilon^2/4) + \exp(-c_1 n\pi_+^2/4) + \exp(-c_2 n\pi_-^2/4)\}.$$

Then we have the following two conclusions:

(a) Let $\delta_D = \min_{j\in D}\{K_j\} - \max_{j\in D^C}\{K_j\}$. *Assume* $\delta_D > 0$ *and* $|D| < d_n$, *then*

$$\Pr\{D \subseteq \hat{S}(d_n)\} \geq 1 - p\zeta(\delta_D/2).$$

Thus, if $\delta_D \gg \{\log(p)/n\}^{1/2}$, *the sure screening property holds with probability going to 1.*

(b) If there exists $\tilde{S}$, *a subset of* $\{1, \ldots, p\}$, *such that* $d_n \geq |\tilde{S}|$, $D \subseteq \tilde{S}$ *and*

$$\delta_{\tilde{S}} = \min_{j\in\tilde{S}} K_j - \max_{j\in\tilde{S}^C} K_j > 0,$$

then we have

$$\Pr\{D \subseteq \hat{S}(d_n)\} \geq 1 - p\zeta(\delta_{\tilde{S}}/2).$$

Thus, if $\delta_{\tilde{S}} \gg \{\log(p)/n\}^{1/2}$, *the sure screening property holds with probability going to 1.*

In Theorem 5.3.1 part (b) is a generalization of part (a). When $\tilde{S} = D$ then part (b) reduces to part (a). To apply Theorem 5.3.1 the key condition is $\delta_D > 0$ or $\delta_{\tilde{S}} > 0$. Similar conditions have been used for establishing the sure screening property of marginal screening methods in the literature. See [93]. Theorem 5.3.1 can hold even when the informative variables and noise variables are correlated. To illustrate, consider the following semiparametric discriminant analysis model [83]

$$g(X) \mid (Y = y) \sim N(\mu_y, \Sigma), \tag{5.1}$$

where $g = (g_1, \ldots, g_p)$ is a set of p unspecified univariate monotone transformations and $\sigma_{jj} = 1$ $(j = 1, \ldots, p)$. The Bayes rule under this model is $\hat{Y} = \text{sign}(\beta_0 + g(X)^{\mathrm{T}}\beta)$, where $\beta = \Sigma^{-1}(\mu_+ - \mu_-)$. Hence, $D = \{j : \beta_j \neq 0\}$. The Kolmogorov–Smirnov test statistics are invariant under the g transformation. By normal theory and direct calculation, we have $K_j = 1 - 2\Phi(|\mu_{+j} - \mu_{-j}|/2)$ where Φ is the cumulative probability function of $N(0, 1)$.

Lemma 5.3.2 *Under model (5.1), we have the following conclusions:*

(a) Assume that Σ has a blockwise independence structure, i.e., $\sigma_{ij} = 0$ if $i \in D$ and $j \in D^C$, then $\delta_D > 0$ if and only if $\min_{j \in D} |\mu_{+j} - \mu_{-j}| > 0$.

(b) Assume that $D = \{1, \ldots, d\}$ and Σ has the autoregressive structure, i.e., $\sigma_{ij} = \rho^{|i-j|}$. If $\Omega_D(\min) = \min_{j \in D} |\mu_{+j} - \mu_{-j}| > 0$, then we let

$$k^* = \frac{\log\{\Omega_D(\min)/|\mu_{+d} - \mu_{-d}|\}}{\log(|\rho|)}$$

and define $\tilde{S}_1(k^*) = \{1, \ldots, d + \lceil k^* \rceil\}$. We have $\delta_{\tilde{S}_1(k^*)} > 0$.

(c) Assume that Σ has the compound symmetry structure, i.e., $\sigma_{ij} = \rho, i \neq j$. Define $\tilde{S}_2 = \{j : \mu_{+j} \neq \mu_{-j}\}$, then $\delta_{\tilde{S}_2} > 0$. Moreover, $D \subseteq \tilde{S}_2$ if and only if $1^{\mathrm{T}}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-) = 0$.

The proof of Lemma 5.3.2 is given in the appendix. Conclusion (a) is intuitively sound, while Conclusions (b) and (c) are somewhat counterintuitive. The key is to note that in a sure screening procedure it suffices to separate D from a large majority of the noise variables, as done in part (b) of Theorem 1. In the autoregressive correlation case, if $\min_{j \in D} |\mu_{+j} - \mu_{-j}| = |\mu_{+d} - \mu_{-d}|$, then $k^* = 0$ and hence there is no restriction on ρ. In general, to apply Conclusion (b) to Theorem 1, we need $d + \lceil k^* \rceil \leq d_n$, which means

$$|\rho| < \exp[\log\{\Omega_D(\min)/|\mu_{+d} - \mu_{-d}|\}/(d_n - d)].$$

Note that d_n is $\lceil n/\log(n) \rceil$ (or n), thus the upper bound on $|\rho|$ can approach 1 asymptotically. In the compound symmetry case there is no restriction on the correlation either.

5.4 A Simulation Study

In this section we use seven simulation models to demonstrate the performance of the Kolmogorov filter. In all cases, $n = 200, p = 2000$ and $D = \{1, \ldots, d\}$.

Model 1: $d = 8$. $\mathbf{X} \mid Y = y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_+ = (1.922 \times 1_8, 0_{p-8})$, and $\boldsymbol{\mu}_- = 0_p$. Conditioning on y, the informative and noise variables are independent and the correlation within the informative/noise group is 0·5. The Bayes error is 10%.

Model 2: $d = 5$. The true variables are independently generated from $X_j \mid Y = +1 \sim t_4$ and

$$X_j \mid Y = -1 \sim 0.5N(2.5, 1) + 0.5N(-2.5, 1)$$

, for $j = 1, \ldots, 5$. The noise variables independently follow $N(0, 1)$. The Bayes error is 3%.

Model 3: $d = 5$. We generated Y from a logistic regression model with log odds

$$\log\left(\frac{\Pr(Y = +1 \mid \mathbf{X} = x)}{\Pr(Y = -1 \mid \mathbf{X} = x)}\right) = -3 + 2x_1 + 2x_2 + 2x_3 + 3\sin(x_4) + 4x_5^2.$$

The variables Xs are independently $N(0, 1)$. The Bayes error is 9.7%.

Model 4: $d = 8$. $\mathbf{X} \mid Y = y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$, where $\sigma_{ij} = 0.8^{|i-j|}$ and $\boldsymbol{\beta} = (-0.41 \times 1_8, 0_{p-8})$ so that the Bayes error rate is 10%. The means $\boldsymbol{\mu}_- = 0_p$, $\boldsymbol{\mu}_+ = \boldsymbol{\Sigma}\boldsymbol{\beta}$.

Model 5: $d = 8$. We first generated $\mathrm{W} \mid Y = y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}_y, \boldsymbol{\Sigma}$ as in Model 4. Then we let $X_j = \exp(2W_j)$.

Model 6: $d = 4$. $\mathbf{X} \mid Y = y \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}_+ = 0{\cdot}63 \times (1, -1, -1, 1, 0_{p-4})$, $\mu_- = 0$, $\sigma_{ij} = 0{\cdot}8$, $i \neq j$. The Bayes error is 8%.

Model 7: $d = 4$. We first generated $W \mid Y = y \sim N(\mu_y, \Sigma)$, with μ_y, Σ as in Model 6. Then we let $X_j = \exp(2W_j)$.

To avoid choosing the thresholds, [93] reported the smallest model size required to contain all the signal features. We follow their convention and report the simulation results in Table 5.1. The two parametric screening methods work very well in Models 1, 4 and 6, so do the two nonparametric screening methods. Models 2, 3, 5 and 7 violate the parametric modeling assumptions from different perspectives. The t-test screening and maximum marginal likelihood screening fail miserably, but the Kolmogorov filter and nonparametric maximum likelihood screening are both able to reduce the dimension effectively. The Kolmogorov filter has the best screening performance in all seven

models. Models 4–7 are designed based on Lemma 1, and the performance of the Kolmogorov filter confirms Lemma 1 and Theorem 1. Our numerical results in Models 4–7 also confirm that the Kolmogorov filter is invariant under monotone transformations. However, nonparametric maximum likelihood screening has different results in Models 6 and 7, which suggests that it does not have the invariance property under monotone transformation. In Table 5.2 we also compare the computation costs of these screening methods. Obviously, t-test screening is the fastest. The Kolmogorov filter is only a little bit slower than t-test screening and is at least 10 times faster than nonparametric maximum likelihood screening.

Table 5.1: Minimum numbers of features needed to recover all the signal features. The numbers are medians of 400 replicates. In the parentheses are the standard errors, obtained by bootstrap.

	Bayes Error(%)	d	Kolmogorov	Nonparametric likelihood	MMLE	t
Model 1	10	8	8(0)	8(0)	8(0)	8(0)
Model 2	3	5	5(0)	5(0)	1675(17·8)	1673(19·0)
Model 3	9·7	5	27(2·0)	72(5·9)	1003·5(45·1)	990(46·1)
Model 4	10	8	8(0)	8(0)	8(0)	8(0)
Model 5	10	8	8(0)	8(0)	9(0)	47·5(8·0)
Model 6	8	4	4(0)	13(1·4)	4(0)	4(0)
Model 7	8	4	4(0)	18(2·2)	212·5(46·4)	210(33·7)

Table 5.2: Total computation time in seconds for 20 replicates.

	Kolmogorov	Nonparametric likelihood	MMLE	t
Model 1	18·9	194·4	132·9	15·8
Model 2	17·8	189·2	127·0	14·9
Model 3	17·7	189·0	125·6	14·8
Model 4	18·2	207·5	136·1	14·6
Model 5	18·3	247·8	140·4	14·1
Model 6	18·6	251·0	141·9	14·4
Model 7	18·4	247·8	139·2	13·9

5.5 The Spam Dataset

In this section, we use the Spam dataset to show how the Kolmogorov filter can influence the final classification. This dataset is available at the UCI Machine Learning Repository[99]. The basic task for the Spam dataset is to predict whether an email is spam. The response Y is coded as $+1$ if an email is spam ($Y = +1$) and as -1 if it is not. Around 40% of the observations in this dataset are spam. There are 57 predictors in total, including frequencies of words such as "you", total number of capital letters, etc.

We randomly split the dataset to form a training set of 300 observations and a testing set of 4301 observations. Note that, because of its nonparametric nature, the Kolmogorov filter can be followed by any classification techniques. While MMLE should be followed by the logistic regression, and the t-test screening should be followed by naive Bayes. Therefore, we apply the following five methods on this dataset: K-RF: Kolmogorov filter followed by random forest [2] (K-RF), Kolmogorov filter followed by logistic regression (K-GLM), Kolmogorov filter followed by naive Bayes [49] (K-NB), MMLE followed by logistic regression (MMLE-GLM) and t-test screening followed by naive Bayes (T-NB).

In Figure 5.1 we see that K-RF consistently give lower error rates than all the other methods regardless of $\hat{d}_n$. More surprisingly, even when the logistic regression is used for classification, K-GLM is still more accurate than MMLE-GLM. And so is the comparison between K-NB and T-NB. Therefore, the Kolmogorov filter does provide a more meaningful ranking than the t-test screening and MMLE.

To understand the superior performance of the Kolmogorov filter, we compare the ranking given by the Kolmogorov filter, t-test screening, MMLE with the ranking of variable importance in random forest. The variable importance in random forest is defined as follows. For each variable X_j, after fitting a model by random forest, statisticians permute X_j and refit the random forest on the permuted data. Then the increase in the prediction error is recorded as the importance of X_j. As in the Kolmogorov filter, this ranking does not rely on any assumptions on the distributions of $\mathbf{X}$.

We plotted the scatter plot of the rankings given by the four methods in Figure 5.2. It can be seen that the rankings given by the Kolmogorov filter closely resembles that by

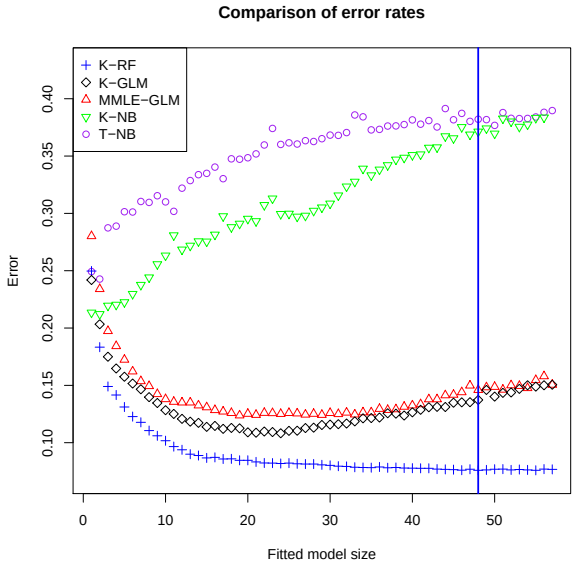

Figure 5.1: We compare several methods on the Spam dataset. The smallest error rate is achieved by K-RF with 48 predictors. The results are based on 100 replicates.

random forest, while the rankings given by the t-test screening and MMLE are generally very different. It further supports that the Kolmogorov filter is more reliable.

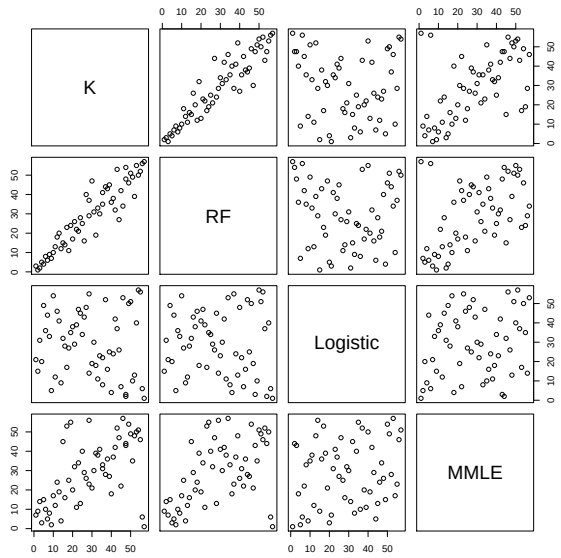

Figure 5.2: Comparison of rankings on the Spam dataset given by the Kolmogorov filter, t-test screening, MMLE and random forest. The Kolmogorov filter resembles random forest closely.

We further compare K-RF with other well-developed high-dimensional classifiers. First, we added independent standard normal noise variables to the Spam dataset so that the new dataset has 1000 dimensions. This yields a more difficult classification problem. Then we applied four methods: K-RF, random forest, ℓ_1 logistic regression (SL) and MMLE followed by ℓ_1 logistic regression (MMLE-SL). We repeatedly split the dataset into training and testing sets, and perform classification for 100 times. Box plots of the error rates are presented in Figure 5.3. It is obvious that K-RF is the most accurate, followed by random forest. This observation confirms the necessity of performing variable selection when the dimension is high. Also, the superior performance of K-RF over SL and MMLE-SL supports the use of robust screening techniques in practice.

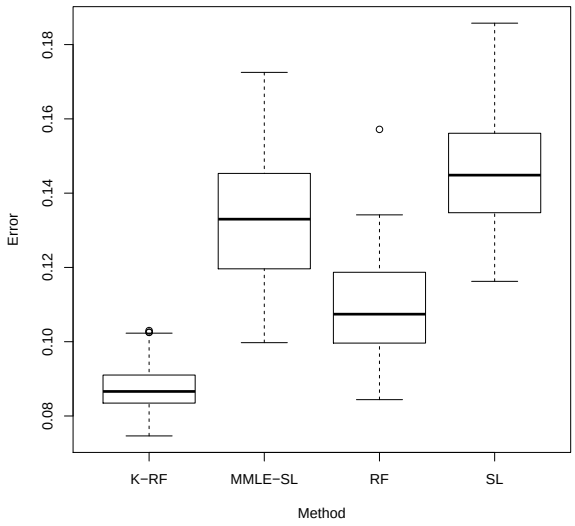

Figure 5.3: Error rates on the Spam dataset given by K-RF, random forest, SL and MMLE-SL.

5.6 Discussion

It would be interesting to use the Kolmogorov filter to develop an iterative screening and sparse model fitting procedure. Some ideas from existing works [1, 96, 95, 93, 97] can be borrowed for that purpose. Such development is left for future study.

Chapter 6

Concluding Remarks

Variable selection in high-dimensional classification is crucial for many applications in practice. By employing variable selection techniques, we can end up with accurate and interpretable classifiers with much less computation cost. However, very few such variable selection techniques exist in the literature, and these methods usually impose very strong assumptions and/or lack theoretical justifications.

This dissertation proposes three variable selection methods for high-dimensional classification, DSDA, SeSDA and the Kolmogorov filter. These methods provide solutions to variable selection for high-dimensional data under parametric, semiparametric and nonparametric models, respectively. Whereas all of the three methods share the advantages in computation efficiency, theoretical justifications and numerical performance.

In particular, DSDA generalizes the well-known LDA model to high dimensions. Unlike its predecessors, DSDA can consistently identify the important variables and estimate the Bayes rule without requiring the independence structure between predictors. A side effect of its consistency is that, through some connections we develop between it and SOS and SFDA, we can show the consistency of SOS and SFDA, which is otherwise unknown. On the other hand, because of its least squares formulation, DSDA can make use of the extremely fast implementations for penalized regression problems, and therefore is faster than other proposals of sparse LDA methods. DSDA also compares favorably to its competitors on simulated datasets and benchmark real datasets.

SeSDA improves the robustness of DSDA since it relaxes the Gaussian assumption

in the LDA model. Instead, it assumes the semiparametric model that the predictors can be nonparametrically transformed so that the LDA model is true. SeSDA is able to accurately estimate the unknown nonparametric transformations and then perform variable selection. Its performance is strongly supported by theoretical and numerical results. Moreover, in order to show the consistency of SeSDA, a new concentration inequality for Gaussian copulas is proved. This inequality can be used to obtain higher convergence rates for almost all methods based on Gaussian copulas.

The Kolmogorov filter is the first nonparametric screening method for classification problems. It utilizes the Kolmogorov-Smirnov statistic to rank the importance of each predictors and screens out the unimportant ones. Since the Kolmogorov-Smirnov statistic is used, the Kolmogorov filter imposes minimal assumptions on the distributions of each predictor to ensure the SURE screening property. It is almost as fast as the t-test screening and much faster than MMLE. It also significantly outperforms existing methods on numerical examples.

However, even with the fast development in this field during the past a few years, many open problems remain in variable selection for high-dimensional classification. Some of them are as follows.

6.1 Variable Selection in Multi-Class Problems

It is worth noting that DSDA, SeSDA and the Kolmogorov filter all deal with binary classification problems. This is actually the case for most of the current variable selection techniques. However, although binary classification is indeed the most important special case in classification, researchers often have to classify observations into several classes. Therefore, extensions of these methods to multi-class problems are highly desirable.

In the low-dimensional scenario, many researchers have found that binary classifiers can be generalized to multi-class classifiers by the one-versus-all or all-versus-all method. Suppose there are K classes in total, the one-versus-all method creates K different binary classifiers so that the j'th classifier estimates the probability of an observation belonging to j'th class, denoted as $\hat{p}_j$. Then the observation is classified to Class i, with

$$i = \arg\max_j \hat{p}_j. \tag{6.1}$$

The all-versus-all method, on the other hand, creates $\frac{K(K-1)}{2}$ binary classifiers so that the classifier δ_{ij} predicts whether an observation belongs to Class i or Class j. Then the predictions are combined to form a final prediction.

However, high-dimensional classifiers cannot be extended so easily, because we also have to consider variable selection. An important issue before any further investigation is the definition of sparsity in multi-class problems. Take the all-versus-all method for an example. One could assume that each δ_{ij} is sparse, and yet each of them can depend on different predictors. Another reasonable assumption is that all δ_{ij}'s depend on the same set of predictors. It is easy to see that the former is weaker than the later, while the later could result in sparser variable selection results. Therefore, it is by far not clear how to generalize the binary classification techniques to multi-class problems and more studies of these two assumptions and corresponding formulations for variable selection can be valuable.

Meanwhile, it could be useful to develop more straightforward variable selection techniques for multi-class problems. For example, SOS and Witten's method for sparse LDA are directly applicable to multi-class data, although these two methods are not yet justified when more than two classes are present. Similar extensions with justifications will be interesting.

6.2 Structured Variable Selection

Practitioners often find that some special structure may exist between predictors. For example, predictors can be grouped [17, 20] or have a linear ordering [64]. In such cases, structured variable selection is expected to be superior to variable selection without structure. With the least-squares formulation in DSDA and SeSDA, the generalization should be easy in terms of implementation, since one could simply replace the Lasso or SCAD penalty with any suitable penalty functions, including the grouped Lasso and the fused Lasso. However, the performance of such generalizations is worth studying.

On the other hand, it is unclear how the Kolmogorov filter should be applied if there is a structure within the predictors. [1] and [38] briefly discussed grouped selection, but more intensive study is by far unavailable.

6.3 Transformations in Sufficient Dimension Reduction

Another potential project is to study the transformations in sufficient dimension reduction. This is actually not related to variable selection in high-dimensional classification, but it shares the idea of SeSDA.

Sufficient dimension reduction techniques have proven to be useful in many applications. For a random pair $(Y, \mathbf{X})$, where $Y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$, such techniques aim to find $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ so that $Y \perp \mathbf{X} \mid \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}$. Some successful sufficient dimension reduction techniques include the sliced inverse regression [100], sliced average variance estimation [101], and inverse regression estimator [102], among others. These techniques rely on a fundamental assumption that the predictors should be elliptically contoured, or, more specifically, normal. When this assumption is violated, all the sufficient dimension reduction techniques mentioned above cannot guarantee to find good estimates of $\boldsymbol{\beta}$.

As in SeSDA, we can relax this assumption for sufficient dimension reduction techniques. We assume that there exists a set of monotone transformations $\mathbf{T} = (T_1, \ldots, T_p)$ such that $\mathbf{T}(\mathbf{X})$ is normal. Then, in practice, we can estimate $\mathbf{T}$ similar to SeSDA and apply the sufficient dimension techniques to $(Y, \hat{\mathbf{T}}(\mathbf{X}))$. The properties of the resulting techniques remain to be investigated, though.

References

[1] J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *Ann. Statist.*, 36:2605–2637, 2008.

[2] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[3] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58:267–288, 1996.

[4] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.*, 96:1348–1360, 2001.

[5] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, 35:2313–2351, 2007.

[6] J. Fan and J. Lv. Sure independence screening for ultra-high dimensional feature space. *J. R. Statist. Soc. B*, 20:101–148, 2008.

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

[8] Q. Mai and H. Zou. The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100:229–234, 2013.

[9] Q. Mai and H. Zou. On the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics*, Accepted, 2012.

[10] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.

[11] Y. Freund and R. Schapire. A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[12] V. Vapnik. *Statistical Learning Theory.* Wiley, New York, 1998.

[13] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *Elements of statistical learning: data mining, inference, and prediction.* Springer Verlag, second edition, 2009.

[14] H. Akaike. Information theory and an extension of the maximum likelihood principle. *In Second International Symposium on Information Theory (Edited by B. N. Petrov and F. Csaki).*, pages 267–281, 1973.

[15] G. Schwartz. Estimating the dimension of a mode. *Ann. Statist.*, 6:461–464, 1978.

[16] C. L. Mallows. Some comments on c_p. *Technometrics*, 15:661–675, 1973.

[17] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67(2):301–320, 2005.

[18] H. Zou. The adaptive Lasso and its oracle properties. *J. Am. Statist. Assoc.*, 101:1418–1429, 2006.

[19] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38:894–942, 2010.

[20] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68:49–67, 2006.

[21] L. Breiman. Better subset selection using the nonnegative garrote. *Technometrics*, 37, 1995.

[22] I.E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–148, 1993.

[23] W. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7, 1998.

[24] M. R. Osborne, B. Presnell, and B. A. Turlach. On the Lasso and its dual. *Journal of Computational and Graphical Statistics*, 9, 2000.

[25] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32:407–499, 2004.

[26] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2008.

[27] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28, 2000.

[28] E. Greenstein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10, 2004.

[29] J. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50, 2004.

[30] D. Donoho. For most large underdetermined systems of equations, the minimal ℓ_1-norm solution is the sparsest solution. *Communications on pure and applied mathematics*, 59, 2006.

[31] N. Meinshausen. Lasso with relaxation. *Computational statistics and data analysis*, 52, 2007.

[32] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

[33] M.J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. Info. Theory*, 55:2183–2202, 2009.

[34] H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.*, 36:1509–1533, 2008.

[35] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Spam: Sparse additive models. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

[36] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B*, 70:849–911, 2008.

[37] G. Li, H. Peng, J. Zhang, and L.-X. Zhu. Robust rank correlation based screening. *Ann. Statist.*, 40:1846–1877, 2012.

[38] R. Li, W. Zhong, and L.-P. Zhu. Feature screening via distance correlation learning. *J. Am. Statist. Assoc.*, 107:1129–1139, 2012.

[39] M. L. Rizzo and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35, 2007.

[40] D. Michie, D.J. Spiegelhalter, and C.C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, first edition, 1994.

[41] D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21:1–14, 2006.

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

[43] P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc. B*, 67:427–444, 2005.

[44] P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *Ann. Statist.*, 36:199–227, 2008.

[45] T. Cai, C. Zhang, and H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38:2118–2144, 2010.

[46] A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electron. J. Statist.*, 2:494–515, 2008.

[47] J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *Ann. Statist.*, 36:2605–2637, 2008.

[48] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Nat. Acad. Sci.*, 99:6567–6572, 2002.

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

[50] Y. Guo, T. Hastie, and Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100, 2006.

[51] S. Wang and J. Zhu. Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, 23:972–979, 2007.

[52] S. Dudoit and M. Van der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer Series in Statistics. New York: Springer, 2008.

[53] B. Efron. *Large-Scale Inference: empirical Bayes methods for estimation, testing and prediction*. Cambridge University Press, 2010.

[54] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300, 1995.

[55] J. Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64:479–498, 2002.

[56] J. Storey, J. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates; a unified approach. *J. R. Statist. Soc. B*, 66:187–206, 2004.

[57] C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Ann. Statist.*, 32:1035–1061, 2004.

[58] B. Efron. Local false discovery rate. Technical report, Stanford University, 2005.

[59] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *The Ann. Statist*, 32:962–994, 2004.

[60] W. Sun and T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Statist. Assoc.*, 102:901–912, 2007.

[61] B. Efron. Empirical Bayes estimates for large-scale prediction problems. *J. Am. Statist. Assoc.*, 104:1015–1028, 2009.

[62] M. Wu, L. Zhang, Z. Wang, D. Christiani, and X. Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25:1145–1151, 2008.

[63] D. Witten and R. Tibshirani. Penalized classification using fisher's linear discriminant. *J. R. Statist. Soc. B*, 73:753–772, 2011.

[64] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Keith. Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc. B*, 67:91–108, 2005.

[65] J. Lv and Y. Fan. A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, 37:3498–3528, 2009.

[66] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.

[67] C.H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36:1567–1594, 2008.

[68] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Mack, and J Levine. 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*, 96:6745–6750, 1999.

[69] D. Singh, P. G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, Loda M., P.W. Kantoff, T.R. Golub, and W.R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[70] M. Yuan, R. Joseph, and Y. Lin. An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49:430–439, 2007.

[71] E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *Ann. Appl. Statist.*, 2:245–263, 2008.

[72] L. Clemmensen, T. Hastie, and B. Ersbøll. Sparse discriminant analysis. *Technometrics*, 53:406–413, 2011.

[73] N. T. Trendafilov and I. T. Jolliffe. DALASS: Variable selection in discriminant analysis via the lasso. *Computational Statistics and Data Analysis*, 51:3718–3736, 2007.

[74] J. Shao, Y. Wang, X. Deng, and S. Wang. Sparse linear discriminant analysis with high dimensional data. *Ann. Statist.*, 2011.

[75] T. Cai and W. Liu. A direct estimation approach to sparse linear discriminant analysis. *J. Am. Statist. Assoc.*, 106:1566–1577, 2011.

[76] Jianqing Fan, Y. Feng, and X. Tong. A ROAD to classification in high dimensional space. *J. R. Statist. Soc. B*, 74:745–771, 2012.

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

[78] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[79] P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5:232–253, 2011.

[80] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization theory and applications*, 109:47–494, 2001.

[81] T. Hastie, R. Tibshirani, and A. Buja. Flexible discriminant analysis by optimal scoring. *J. Am. Statist. Assoc.*, 89:1255–1270, 1994.

[82] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixtures. *J. R. Statist. Soc. B*, 58:155–176, 1996.

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

[84] Q. Mai and H. Zou. The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100:229–234, 2013.

[85] T. Hastie and R. Tibshirani. *Generalized Additive Models*. New York: Chapman and Hall, 1990.

[86] G. James, P. Radchenko, and J. Lv. Dasso: connections between the dantzig selector and lasso. *J. R. Statist. Soc. B*, 71, 2009.

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

[88] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. High dimensional semiparametric gaussian copula graphical models. *Ann. Statist.*, 40:2293–2326, 2012.

[89] L. Xue and H. Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40, 2012.

[90] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34:1436–1462, 2006.

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

[92] G. A. Heap, G. Trynka, R. C. Jansen, and M. et al. Bruinenberg. Complex nature of snp genotype effects on gene expression in primary human leucocytes. *BMC Med Genomics*, 7, 2009.

[93] J. Fan and R. Song. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.*, 38(6):3567–3604, 2010.

[94] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *The Annual Conference on Neural Information Processing Systems 16*. MA: MIT Press, 2004.

[95] J. Fan, R. Samworth, and Y. Wu. Ultra-dimensional variable selection via independenct learning: beyond the linear model. *J. Mach. Learn. Res.*, 10:2013–2038, 2009.

[96] J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Am. Statist. Assoc.*, 106:544–557, 2011.

[97] L.-P. Zhu, L. Li, R. Li, and L.-X. Zhu. Model-free feature screening for ultrahigh dimensional data. *J. Am. Statist. Assoc.*, 106:1464–1475, 2011.

[98] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956.

[99] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[100] K. C. Li. Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Assoc.*, 86:316–342, 1991.

[101] R. D. Cook and S. Weisberg. Discussion of sliced inverse regression for dimension reduction, by k.-c. li. *J. Amer. Statist. Assoc.*, 86:328–332, 1991.

[102] R. D. Cook and L. Ni. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Am. Statist. Assoc.*, 100(470):410–428, 2005.

[103] R.J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, first edition, 1980.

[104] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Prob.*, 18:1269–1283, 1990.

Appendix A

Technical Details in Chapter 2

A.1 Proofs

Proof [Proof of Proposition 1] 1. Let $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ and $\boldsymbol{\beta}^{Bayes} = \mathbf{\Omega}(\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)$. Write $\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega}_{\tilde{A},\tilde{A}} & \mathbf{\Omega}_{\tilde{A},\tilde{A}^c} \\ \mathbf{\Omega}_{\tilde{A}^c,\tilde{A}} & \mathbf{\Omega}_{\tilde{A}^c,\tilde{A}^c} \end{pmatrix}$. Note that $A \subseteq \tilde{A}$ is equivalent to $\boldsymbol{\beta}^{Bayes}_{\tilde{A}^c} = 0$. On the other hand, we have

$$\boldsymbol{\beta}^{Bayes}_{\tilde{A}^c} = \mathbf{\Omega}_{\tilde{A}^c,\tilde{A}}(\boldsymbol{\mu}_{-,\tilde{A}} - \boldsymbol{\mu}_{+,\tilde{A}})$$

and

$$\mathbf{\Omega}_{\tilde{A}^c,\tilde{A}} = -(\mathbf{\Sigma}_{\tilde{A}^c,\tilde{A}^c} - \mathbf{\Sigma}_{\tilde{A}^c,\tilde{A}}\mathbf{\Sigma}^{-1}_{\tilde{A},\tilde{A}}\mathbf{\Sigma}_{\tilde{A},\tilde{A}^c})^{-1}\mathbf{\Sigma}_{\tilde{A}^c,\tilde{A}}\mathbf{\Sigma}^{-1}_{\tilde{A},\tilde{A}}.$$

Therefore, part 1 is proven.

2. By definition, $\tilde{A} \subseteq A \iff A^c \subseteq \tilde{A}^c \iff \boldsymbol{\mu}_{-,A^c} = \boldsymbol{\mu}_{+,A^c}$. Now using $\boldsymbol{\mu}_- - \boldsymbol{\mu}_+ = \mathbf{\Sigma}\boldsymbol{\beta}^{Bayes}$ we have

$$\boldsymbol{\mu}_{-,A} - \boldsymbol{\mu}_{+,A} = \mathbf{\Sigma}_{A,A}\boldsymbol{\beta}^{Bayes}_A$$

and

$$\boldsymbol{\mu}_{-,A^c} - \boldsymbol{\mu}_{+,A^c} = \mathbf{\Sigma}_{A^c,A}\boldsymbol{\beta}^{Bayes}_A.$$

Hence, it yields that $\boldsymbol{\mu}_{-,A^c} - \boldsymbol{\mu}_{+,A^c} = \mathbf{\Sigma}^{-1}_{A,A}(\boldsymbol{\mu}_{-,A} - \boldsymbol{\mu}_{+,A})$. Then part 2 is proven.

Proof [Proof of Proposition 2] We recode the response variable as $y^* = 1, -1$. Note that $\tilde{\beta}_0^{opt.} = \arg\min_{\tilde{\beta}_0} E(y^*_{new} \neq \text{sign}(\mathrm{x}^T_{new}\tilde{\boldsymbol{\beta}} + \tilde{\beta}_0)|\text{training data})$. Since $y^*_{new}, \mathrm{x}_{new}$ are

independent from the training data, $(Y^*_{new}, \mathrm{z}_{new} = \mathrm{x}^T_{new}\tilde{\boldsymbol{\beta}})$ obeys a one-dimensional LDA model, that is,

$$\mathrm{z}_{new}|y^*_{new} = -1 \sim N(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\mu}_-, \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}), \quad \Pr(y^*_{new} = -1) = \pi_-$$

and

$$\mathrm{z}_{new}|y^*_{new} = +1 \sim N(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\mu}_+, \tilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\Sigma}\tilde{\boldsymbol{\beta}}), \quad \Pr(y^*_{new} = +1) = \pi_+.$$

Then by some straightforward calculation we obtain (2.8).

Proof [Proof of Proposition 3] It suffices to prove that, there exists a constant $c > 0$ such that $\tilde{\boldsymbol{\beta}}(Bayes)_A = c\boldsymbol{\beta}(Bayes)_A$. Note that $C_{AA} = \boldsymbol{\Sigma}_{AA} + \pi_+\pi_-(\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})(\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})^T$ and $C_{AA}\tilde{\boldsymbol{\beta}}(Bayes)_A = \boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A}$. Let

$$c = n\left(n - 2 + \pi_+\pi_-(\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})^T\boldsymbol{\Sigma}^{-1}_{AA}(\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\right)^{-1} > 0$$

then we have $\tilde{\boldsymbol{\beta}}(Bayes)_A = c\boldsymbol{\beta}(Bayes)_A$.

We now prove theorems 1, 2 and 3. The following two lemmas provide some useful concentration inequalities that are repeatedly used in the proof.

Lemma A.1.1 *There exists some constants ϵ_0 and c_1, c_2 such that for any $\epsilon \le \epsilon_0$ we have*

$$\Pr(|C^{(n)}_{ij} - C_{ij}| \ge \epsilon) \le 2\exp(-n\epsilon^2 c_1), \tag{A.1}$$

for each (i, j) pair; and

$$\Pr(|(\hat{\mu}_{-j} - \hat{\mu}_{+j}) - (\mu_{-j} - \mu_{+j})| \ge \epsilon) \le 2\exp(-n\epsilon^2 c_2). \tag{A.2}$$

for each j. Moreover, we have

$$\Pr(\|C^{(n)}_{AA} - C_{AA}\|_\infty \ge \epsilon) \le 2s^2\exp(-\frac{n}{s^2}\epsilon^2 c_1), \tag{A.3}$$

$$\Pr(\|C^{(n)}_{A^cA} - C_{A^cA}\|_\infty \ge \epsilon) \le 2(p-s)s\exp(-\frac{n}{s^2}\epsilon^2 c_1), \tag{A.4}$$

$$\Pr(\|(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+) - (\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)\|_\infty \ge \epsilon) \le 2p\exp(-n\epsilon^2 c_2), \tag{A.5}$$

$$\Pr(\|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_\infty \ge \epsilon) \le 2s\exp(-n\epsilon^2 c_2). \tag{A.6}$$

Lemma A.1.2 *There exists some constants ϵ_0, c_1 such that for any $\epsilon \leq \min(\epsilon_0, \frac{1}{\varphi})$, we have*

$$\Pr(\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1} - C_{A^cA}(C_{AA})^{-1}\|_\infty \geq \frac{(\kappa+1)\epsilon\varphi}{1-\varphi\epsilon}) \leq 2ps\exp(-\frac{n}{s^2}\epsilon^2 c_1). \tag{A.7}$$

Proof [Proof of Lemma 1]

Note that inequalities in (A.3)–(A.6) can be obtained from (A.1)-(A.2) by simple union bounds. So we only prove (A.1) and (A.2). First, it is easy to see that $\Pr(|\hat{\mu}_{+j} - \mu_{+j}| \geq \epsilon | Y) \leq 2\exp(-n_+\frac{\epsilon^2}{2\sigma_j^2})$. Also, $n_+ \sim Bernoulli(n, \pi_+)$. Hence,

$$\Pr(|n_+ - \pi_+ n| \geq n\epsilon) \leq 2\exp(-nc_2'\epsilon^2)$$

for some $c_2' > 0$. Therefore,

$$\Pr(|\hat{\mu}_+ - \mu_+| \geq \epsilon) \leq 2\exp(-n\frac{\pi_+}{2}\frac{\epsilon^2}{2\sigma_j^2}) + 2\exp(-nc_2'(\frac{\pi}{2})^2) \leq 2\exp(-nc_2^{(1)}\epsilon^2)$$

for some small enough $c_2^{(1)}$ and $\epsilon > 0$. Similarly, we have $\Pr(|\hat{\mu}_- - \mu_-| \geq \epsilon) \leq 2\exp(-nc_2^{(2)}\epsilon^2)$. Thus (A.2) holds.

To prove (A.1), note that $C^{(n)}_{ij} = \frac{1}{n}\sum_{k=1}^n x_{ki}x_{kj} - \bar{x}_i\bar{x}_j$. Since $\bar{x}_v = \hat{\pi}_+\hat{\mu}_{1v} + \hat{\pi}_-\hat{\mu}_{2v}$, for $v = i, j$, by the previous arguments, we know that there exists $c_1' > 0$ such that

$$\Pr(|\bar{x}_i\bar{x}_j - Ex_iEx_j| \geq \epsilon) \leq 2\exp(-nc_0''\epsilon^2). \tag{A.8}$$

$\frac{1}{n}\sum_{k=1}^n x_{ki}x_{kj} - E(x_ix_j) = \sum_{l=1}^2 \frac{n_l}{n}\left(\frac{1}{n_l}\sum_{g_k=l} x_{ki}x_{kj} - E(x_ix_j|g=l)\right) + \sum_{l=1}^2 E(x_ix_j|g=l)(\frac{n_l}{n} - \pi_l)$ and $E(x_ix_j|g=l) = \Sigma_{ij} + \mu_{li}\mu_{lj}$ for $l = 1, 2$. Then it suffices to show that there exists some constant $c_1^{(l)}$ such that

$$\Pr\left(|\frac{1}{n_l}\sum_{g_k=l} x_{ki}x_{kj} - E(x_ix_j|g=l)| \geq \epsilon | Y\right) \leq 2\exp(-n_l c_1^{(l)}\epsilon^2). \tag{A.9}$$

We further have that

$$\begin{aligned} n^{-1}\sum_{k=1}^n x_{ki}x_{kj} - E(x_ix_j) &= \sum_{l=1}^2 n_l/n\left\{n_l^{-1}\sum_{g_k=l} x_{ki}x_{kj} - E(x_ix_j \mid g=l)\right\} \\ &\quad + \sum_{l=1}^2 E(x_ix_j \mid g=l)(n_l/n - \pi_l) \end{aligned}$$

and

$$E(x_i x_j \mid g = l) = \Sigma_{ij} + \mu_{li}\mu_{lj}$$

for $l = +1, -1$. Note that

$$n_l^{-1} \sum_{g_k=l} x_{ki}x_{kj} = n_l^{-1} \sum_{g_k=l} (x_{ki} - \mu_{li})(x_{kj} - \mu_{lj}) + \mu_{li}(\mu_{lj} - \hat{\mu}_{lj}) + \mu_{lj}(\mu_{li} - \hat{\mu}_{li}) + \mu_{li}\mu_{lj}.$$

[44] showed that, for $\epsilon < \epsilon_0$,

$$\Pr(|n_l^{-1} \sum_{g_k=l} (x_{ki} - \mu_{li})(x_{kj} - \mu_{lj}) - \Sigma_{ij}| > \epsilon \mid Y) \leq 2\exp(-c_3 n\epsilon^2). \tag{A.10}$$

Combining the concentration results for $\hat{\mu}_{l\nu}$, n_l and (A.10), we have (A.1).

Proof [Proof of Lemma 2] Let $\eta_1 = \|C_{AA} - C_{AA}^{(n)}\|_\infty$, $\eta_2 = \|C_{A^cA} - C_{A^cA}^{(n)}\|_\infty$ and $\eta_3 = \|(C_{AA}^{(n)})^{-1} - (C_{AA})^{-1}\|_\infty$. First we have

$$\begin{aligned}
& \|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}(C_{AA})^{-1}\|_\infty \qquad (A.11)\\
\leq\ & \|C_{A^cA}^{(n)} - C_{A^cA}\|_\infty \cdot \|(C_{AA}^{(n)})^{-1} - (C_{AA})^{-1}\|_\infty + \|C_{A^cA}^{(n)} - C_{A^cA}\|_\infty \cdot \|(C_{AA})^{-1}\|_\infty \\
& + \|C_{A^cA}(C_{AA})^{-1}\|_\infty \cdot \|C_{AA} - C_{AA}^{(n)}\|_\infty \cdot \|(C_{AA})^{-1}\|_\infty \\
& + \|C_{A^cA}(C_{AA})^{-1}\|_\infty \cdot \|C_{AA} - C_{AA}^{(n)}\|_\infty \cdot \|(C_{AA}^{(n)})^{-1} - (C_{AA})^{-1}\|_\infty \\
\leq\ & (\kappa\eta_1 + \eta_2)(\varphi + \eta_3)
\end{aligned}$$

Moreover,

$$\begin{aligned}
\eta_3 &\leq \|(C_{AA}^{(n)})^{-1}\|_\infty \cdot \|(C_{AA}^{(n)} - C_{AA})\|_\infty \cdot \|(C_{AA})^{-1}\|_\infty \qquad (A.12)\\
&= (\varphi + \eta_3)\varphi\eta_1.
\end{aligned}$$

So as long as $\varphi\eta_1 < 1$ we have $\eta_3 \leq \frac{\varphi^2\eta_1}{1-\varphi\eta_1}$ and hence

$$\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}(C_{AA})^{-1}\|_\infty \leq \frac{(\kappa\eta_1 + \eta_2)\varphi}{1 - \varphi\eta_1}. \tag{A.13}$$

Then we consider the event of $\max(\eta_1, \eta_2) \leq \epsilon$ and use Lemma 1 to obtain Lemma 2.

With $y = \frac{n}{n_-}$ or $-\frac{n}{n_+}$ and the centered predictor matrix $\tilde{X}$, we can rewrite the Lasso-DSDA estimator as

$$\hat{\boldsymbol{\beta}}(lasso) = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\boldsymbol{\beta}^{\mathrm{T}}(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}})\boldsymbol{\beta} - 2(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{A.14}$$

Similar to the Lasso-DSDA, the SCAD-DSDA estimator can be written as

$$\arg\min \frac{1}{n}\boldsymbol{\beta}^{\mathrm{T}}(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}})\boldsymbol{\beta} - 2(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\boldsymbol{\beta} + \sum_{j=1}^{p} P_{\lambda,a}(|\beta_j|), \tag{A.15}$$

where $P_{\lambda,a}(\cdot)$ is the SCAD penalty function.

Proof [Proof of Theorem 4.5.4] Part (1). By definition we can write

$$\hat{\boldsymbol{\beta}}_A = (\frac{1}{n}\tilde{X}_A^T\tilde{X}_A)^{-1}((\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - \frac{\lambda}{2}t_A) \tag{A.16}$$

where t_A represents the so-called subgradient which is defined as

$$t_j = \begin{cases} \text{sign}(\hat{\beta}_j), & \text{if } \hat{\beta}_j \neq 0, \\ t_j \in (-1,1), & \text{if } \hat{\beta}_j = 0. \end{cases}$$

From (A.16) we can write

$$\begin{aligned} \hat{\boldsymbol{\beta}}_A &= (C_{AA})^{-1}(\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A}) + (C_{AA}^{(n)})^{-1}((\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})) \\ &\quad -((C_{AA}^{(n)})^{-1} - (C_{AA})^{-1})(\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A}) - \frac{\lambda}{2}(C_{AA}^{(n)})^{-1}t_A. \end{aligned} \tag{A.17}$$

In order to show $\hat{\boldsymbol{\beta}}(lasso) = (\hat{\boldsymbol{\beta}}_A, 0)$ it suffices to verify

$$\|\frac{1}{n}\tilde{X}_{A^c}^T\tilde{X}_A\hat{\boldsymbol{\beta}}_A - (\hat{\boldsymbol{\mu}}_{+A^C} - \hat{\boldsymbol{\mu}}_{-A^C})\|_\infty \leq \frac{\lambda}{2}. \tag{A.18}$$

The left hand side of (A.18) is equal to

$$\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1}\frac{\lambda}{2}t_A - (\hat{\boldsymbol{\mu}}_{+A^C} - \hat{\boldsymbol{\mu}}_{-A^C})\|_\infty. \tag{A.19}$$

Using $C_{A^cA}C_{AA}^{-1}(\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A}) = (\mu_{+A^C} - \mu_{-A^C})$, (A.19) has an upper bound:

$$\begin{aligned} U_1 &= \|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}C_{AA}^{-1}\|_\infty\Delta + \|(\hat{\boldsymbol{\mu}}_{+A^C} - \hat{\boldsymbol{\mu}}_{-A^C}) - (\mu_{+A^C} - \mu_{-A^C})\|_\infty \\ &\quad +(\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}C_{AA}^{-1}\|_\infty + \kappa)\|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_\infty \\ &\quad +(\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}C_{AA}^{-1}\|_\infty + \kappa)\frac{\lambda}{2} \end{aligned} \tag{A.20}$$

Pick ϵ such that $\epsilon < \epsilon_0$ and $\epsilon < \frac{\frac{\lambda}{4\varphi}(1-\kappa)}{\frac{\lambda}{2}+(1+\kappa)\Delta}$. Check $\epsilon < \frac{1-\kappa}{2}\frac{1}{\varphi}$. If

$$\|C_{A^cA}^{(n)}(C_{AA}^{(n)})^{-1} - C_{A^cA}C_{AA}^{-1}\|_\infty \leq \frac{(\kappa+1)\epsilon\varphi}{1-\varphi\epsilon} \tag{A.21}$$

and

$$\|(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+) - (\boldsymbol{\mu}_- - \boldsymbol{\mu}_+)\|_\infty \leq \frac{\lambda}{4}\frac{1-\kappa-2\epsilon\varphi}{1+\kappa} \tag{A.22}$$

then $U_1 \leq \frac{\lambda}{2}$. Therefore, by Lemma A.1.1 and Lemma A.1.2, we have

$$\begin{aligned} & \Pr(\|\frac{1}{n}\tilde{\mathbf{X}}_{A^c}^{\mathrm{T}}\tilde{\mathbf{X}}_A\hat{\boldsymbol{\beta}}_A - (\mu_{+A^C} - \mu_{-A^C})\|_\infty \leq \frac{\lambda}{2}) \qquad (A.23)\\ \equiv\ & 1-\delta_1 \\ \geq\ & 1 - 2ps\exp(-\frac{n}{s^2}\epsilon^2 c_1) - 2p\exp(-nc_2(\frac{\lambda}{4}\frac{1-\kappa-2\epsilon\varphi}{1+\kappa})^2). \end{aligned}$$

part (2). Let $\zeta = \frac{|\boldsymbol{\beta}^*|_{min}}{\Delta\varphi}$. Write $\eta_1 = \|C_{AA} - C_{AA}^{(n)}\|_\infty$ and $\eta_3 = \|(C_{AA}^{(n)})^{-1} - (C_{AA})^{-1}\|_\infty$. Then for any $j \in A$,

$$|\hat{\beta}_j| \ \geq\ \zeta\Delta\varphi - (\eta_3+\varphi)(\frac{\lambda}{2} + \|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - \boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_\infty) - \eta_3\Delta. \tag{A.24}$$

When $\eta_1\varphi < 1$ we have shown that $\eta_3 < \frac{\varphi^2\eta_1}{1-\eta_1\varphi}$, thus

$$|\hat{\beta}_j| \ \geq\ \zeta\Delta\varphi - \frac{1}{1-\eta_1\varphi}(\frac{\lambda\varphi}{2} + \|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_\infty\varphi + \varphi^2\eta_1\Delta) \equiv L_1 \tag{A.25}$$

Note that $\zeta \leq 1$, because $\|\boldsymbol{\beta}^*\|_\infty \leq \Delta\varphi$. Hence $\lambda \leq \frac{1}{2}|\boldsymbol{\beta}^*|_{min}/\varphi \leq \frac{2}{3+\zeta}|\boldsymbol{\beta}^*|_{min}/\varphi$. Pick ϵ such that $\epsilon < \min(\epsilon_0, \frac{1}{\varphi}\frac{\zeta}{3+\zeta}, \frac{\Delta}{2}\frac{\zeta}{3+\zeta})$ Under the events $\eta_1 \leq \epsilon$ and $\|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_\infty \leq \epsilon$ we have $L_1 > 0$. Therefore,

$$\Pr(L_1 > 0) \ \geq\ 1 - 2s^2\exp(-\frac{nc_1}{s^2}\epsilon^2) - 2s\exp(-nc_2\epsilon^2). \tag{A.26}$$

Part (3). By (A.17) and $\eta_1\varphi < 1$, we have

$$\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}^*\|_\infty \leq \frac{1}{1-\eta_1\varphi}(\frac{\lambda}{2}\varphi + \|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_\infty\varphi + \varphi^2\eta_1\Delta). \tag{A.27}$$

Pick ϵ such that $\epsilon < \min(\epsilon_0, \frac{\lambda}{2\varphi\Delta}, \lambda)$. Under the events $\eta_1 < \epsilon$ and $\|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_\infty \leq \epsilon$ we have $\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}^*\|_\infty \leq 4\varphi\lambda$. Thus,

$$\Pr(\|\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}^*\|_\infty \leq 4\varphi\lambda) \ \geq\ 1 - 2s^2\exp(-\frac{nc_1}{s^2}\epsilon^2) - 2s\exp(-nc_2\epsilon^2). \tag{A.28}$$

This completes the proof.

Proof [Proof of Theorem 2.5.3] Part (1). Fix any positive ϵ satisfying $\frac{\epsilon}{\epsilon+2\Delta\varphi} \leq \min(\epsilon_0\varphi, \epsilon_0\frac{1}{\Delta})$. Under the events $\eta_1\varphi < \frac{1}{1+\frac{2\Delta\varphi}{\epsilon}}$ and $\|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_\infty \leq$

$\frac{\epsilon}{\epsilon+2\Delta\varphi}\Delta$, we have

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}(oracle)-\boldsymbol{\beta}^*\|_\infty &\leq (\eta_3+\varphi)\|(\hat{\boldsymbol{\mu}}_{+A}-\hat{\boldsymbol{\mu}}_{-A})-(\boldsymbol{\mu}_{+A}-\boldsymbol{\mu}_{-A})\|_\infty+\eta_3\|\boldsymbol{\mu}_{+A}-\boldsymbol{\mu}_{-A}\|_\infty \\ &\leq \frac{1}{1-\eta_1\varphi}(\|(\hat{\boldsymbol{\mu}}_{+A}-\hat{\boldsymbol{\mu}}_{-A})-(\boldsymbol{\mu}_{+A}-\boldsymbol{\mu}_{-A})\|_\infty\varphi+\varphi^2\eta_1\Delta) \\ &< \epsilon. \end{aligned} \tag{A.29}$$

Then (2.21) is obtained by Lemma A.1.1.

Part (2). Let $g(\boldsymbol{\beta})=\frac{1}{n}\boldsymbol{\beta}^{\mathrm{T}}(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}})\boldsymbol{\beta}-2(\hat{\boldsymbol{\mu}}_- -\hat{\boldsymbol{\mu}}_+)^T\boldsymbol{\beta}+\sum_{j=1}^p P_\lambda(|\beta_j|)$. If the following two conditions hold

$$|\hat{\boldsymbol{\beta}}(oracle)_A|_{min}>a\lambda \tag{A.30}$$

$$\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}(\hat{\boldsymbol{\mu}}_{+A}-\hat{\boldsymbol{\mu}}_{-A})-(\hat{\boldsymbol{\mu}}_{+A^C}-\hat{\boldsymbol{\mu}}_{-A^C})\|_\infty<\frac{\lambda}{2} \tag{A.31}$$

then $\hat{\boldsymbol{\beta}}(oracle)$ is a local minimizer of $g(\boldsymbol{\beta})$. To see this, consider any $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(oracle)+b$ with a sufficiently small b satisfying $\|b\|_2<\min(\lambda,\frac{1}{2}(|\hat{\boldsymbol{\beta}}(oracle)_A|_{min}-a\lambda))$. Let $z=C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}(\hat{\boldsymbol{\mu}}_{+A}-\hat{\boldsymbol{\mu}}_{-A})-(\hat{\boldsymbol{\mu}}_{+A^C}-\hat{\boldsymbol{\mu}}_{-A^C})$. Then it is easy to check that

$$\begin{aligned} g(\boldsymbol{\beta})-g(\hat{\boldsymbol{\beta}}(oracle)) &= b^{\mathrm{T}}\mathbf{C}^{(n)}b+(\sum_{j\in A^c}\lambda|b_j|)+2z^{\mathrm{T}}b_{A^c} \\ &\geq \sum_{j\in A^c}(\lambda-2\|z\|_\infty)|b_j| \;\geq\; 0. \end{aligned} \tag{A.32}$$

Clearly, "=" is taken if and only if $b=0$. Thus, within a sufficiently small ball centered at $\hat{\boldsymbol{\beta}}(oracle)$, $\hat{\boldsymbol{\beta}}(oracle)$ is the unique (strict) minimizer of the objective function.

First, we derive a bound for the probability of (A.30). Pick some ϵ in part (1) and let $\epsilon<|\boldsymbol{\beta}^*|_{min}-a\lambda$. Then $\Pr(|\hat{\boldsymbol{\beta}}(oracle)_A|_{min}>a\lambda)>\Pr(|\hat{\boldsymbol{\beta}}(oracle)_A|_{min}>|\boldsymbol{\beta}^*|_{min}-\epsilon)$ and (2.21) implies

$$\Pr(|\hat{\boldsymbol{\beta}}(oracle)_A|_{min}>|\boldsymbol{\beta}^*|_{min}-\epsilon)\geq 1-2s^2\exp(-\frac{nc_1}{4s^2}\frac{\epsilon^2}{\varphi^2(\epsilon+2\Delta\varphi)^2})-2s\exp(-\frac{nc_2}{4}\frac{\epsilon^2\Delta^2}{(\epsilon+2\Delta\varphi)^2}). \tag{A.33}$$

To derive a bound for the probability of (A.31), we use similar arguments as in the proof of Theorem 4.5.4. Consider three events $\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}-C_{A^cA}C^{-1}_{AA}\|_\infty\leq\epsilon<\frac{\lambda}{6\Delta}$, $\|(\hat{\boldsymbol{\mu}}_{+A^C}-\hat{\boldsymbol{\mu}}_{-A^C})-(\mu_{+A^C}-\mu_{-A^C})\|_\infty\leq\epsilon<\frac{\lambda}{6}$ and $\|(\hat{\boldsymbol{\mu}}_{+A}-\hat{\boldsymbol{\mu}}_{-A})-(\boldsymbol{\mu}_{+A}-\boldsymbol{\mu}_{-A})\|_\infty\leq$

$\epsilon < \frac{\lambda}{6\kappa + \frac{\lambda}{\Delta}}$. Then we have

$$\begin{aligned} & \|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\hat{\boldsymbol{\mu}}_{+A^C} - \hat{\boldsymbol{\mu}}_{-A^C})\|_\infty \\ \leq \; & \|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1} - C_{A^cA}C^{-1}_{AA}\|_\infty \Delta + \|(\hat{\boldsymbol{\mu}}_{+A^C} - \hat{\boldsymbol{\mu}}_{-A^C}) - (\mu_{+A^C} - \mu_{-A^C})\|_\infty \\ & + (\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1} - C_{A^cA}C^{-1}_{AA}\|_\infty + \kappa)\|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_\infty \\ < \; & \frac{\lambda}{2}. \end{aligned} \tag{A.34}$$

By Lemma A.1.1 and Lemma A.1.2, we also have

$$\begin{aligned} & \Pr(\|C^{(n)}_{A^cA}(C^{(n)}_{AA})^{-1}(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\hat{\boldsymbol{\mu}}_{+A^C} - \hat{\boldsymbol{\mu}}_{-A^C})\|_\infty < \frac{\lambda}{2}) \\ \geq \; & 1 - 2p\exp(-nc_2\epsilon^2) - 2ps\exp(-\frac{nc_1}{s^2}\frac{1}{\varphi^2}(\frac{\epsilon}{\epsilon + \kappa + 1})^2). \end{aligned} \tag{A.35}$$

We obtain the expression for δ_3 in (2.22) by combining (A.33) and (A.35). This completes the proof.

Proof [Proof of Theorem 3] Theorem 3 directly follows Theorems 1 and 2.

Appendix B

Technical Details in Chapter 3

B.1 Proofs

Sparse optimal scoring works with centered predictors. Both the ℓ_1-Fisher's discriminant analysis and the direct sparse discriminant analysis use an intercept term in their formulation, which allows us to assume that X is centered without loss of generality.

Proof [Proof of Theorem 3.8.1] First, note the following facts

$$\sum_{k=+1,-1} \hat{\pi}_k \hat{\boldsymbol{\mu}}_k \hat{\boldsymbol{\mu}}_k^{\mathrm{T}} = \hat{\pi}_1 \hat{\pi}_2 (\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}},$$
$$y^{\mathrm{T}}\mathbf{X} = 2n(\hat{\boldsymbol{\mu}}_- - \boldsymbol{\mu}_+)^{\mathrm{T}},\ \mathbf{X}^{\mathrm{T}}\mathbf{X} = (n-2)\hat{\boldsymbol{\Sigma}} + n\hat{\boldsymbol{\Sigma}}_b,$$

where $\hat{\boldsymbol{\Sigma}}_b = \hat{\pi}_+ \hat{\pi}_- (\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}$. Hence, we can write $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}} = \arg\min L_3(\boldsymbol{\beta}, \lambda)$, where

$$L_3(\boldsymbol{\beta}, \lambda) = -2n(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\boldsymbol{\beta} + (n-2)\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + n\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1. \tag{B.1}$$

For notation convenience, write $c_1 = c_1(\lambda)$ and $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda)/c_1(\lambda)$. Then $(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}} = 1$. Denote $L_1(\boldsymbol{\beta}, \lambda) = \boldsymbol{\beta}^{\mathrm{T}}[(n-2)/n\hat{\boldsymbol{\Sigma}}]\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1$. Let $\tilde{\lambda} = \lambda/(n|c_1|)$. Now it suffices to check that, for any $\tilde{\boldsymbol{\beta}}'$ such that $(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}}' = 1$, we have

$$L_1(\tilde{\boldsymbol{\beta}}', \tilde{\lambda}) \geq L_1(\tilde{\boldsymbol{\beta}}, \tilde{\lambda}). \tag{B.2}$$

This is indeed true, because

$$\begin{aligned} L_3(c_1\tilde{\boldsymbol{\beta}}', \lambda) &= -2nc_1(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\tilde{\boldsymbol{\beta}}' + (n-2)c_1^2\tilde{\boldsymbol{\beta}}'^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}\tilde{\boldsymbol{\beta}}' + nc_1^2\tilde{\boldsymbol{\beta}}'^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_b\tilde{\boldsymbol{\beta}}' + |c_1|\lambda\|\tilde{\boldsymbol{\beta}}'\|_1, \\ &= -2nc_1 + nc_1^2 + nc_1^2[\tilde{\boldsymbol{\beta}}'^{\mathrm{T}}[(n-2)/n\hat{\boldsymbol{\Sigma}}]\tilde{\boldsymbol{\beta}}' + \tilde{\lambda}\|\tilde{\boldsymbol{\beta}}'\|_1], \\ &= -2nc_1 + nc_1^2 + nc_1^2 L_1(\tilde{\boldsymbol{\beta}}', \tilde{\lambda}), \end{aligned}$$

which yields

$$L_1(\tilde{\boldsymbol{\beta}}', \tilde{\lambda}) = \frac{1}{nc_1^2}[L_3(c_1\tilde{\boldsymbol{\beta}}', \lambda) + 2nc_1 - nc_1^2]. \tag{B.3}$$

Similarly,

$$L_1(\tilde{\boldsymbol{\beta}}, \tilde{\lambda}) = \frac{1}{nc_1^2}[L_3(c_1\tilde{\boldsymbol{\beta}}, \lambda) + 2nc_1 - nc_1^2]. \tag{B.4}$$

Because $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda) = c_1\tilde{\boldsymbol{\beta}}$ minimizes $L_3(\boldsymbol{\beta}, \lambda)$, we have $L_3(c_1\tilde{\boldsymbol{\beta}}, \lambda) \leq L_3(c_1\tilde{\boldsymbol{\beta}}', \lambda)$. Combine this fact with (B.3)–(B.4) and we have (B.2).

Proof [Proof of Theorem 3.8.2] For convenience, write $\hat{\boldsymbol{\beta}}^{\mathrm{SOS}} = \hat{\boldsymbol{\beta}}^{\mathrm{SOS}}(\lambda)$, $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}} = \hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}(\lambda)$. It is easy to check that, if $(\hat{\theta}, \hat{\boldsymbol{\beta}}^{\mathrm{SOS}})$ is a solution to SOS, then $(-\hat{\theta}, -\hat{\boldsymbol{\beta}}^{\mathrm{SOS}})$ is also a solution. Therefore, we restrict our attention to $\{\boldsymbol{\beta} : (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)^{\mathrm{T}}\boldsymbol{\beta} > 0\}$. It is given in [72] that $\hat{\theta}(\boldsymbol{\beta}) = c_2\tilde{\theta}$, where

$$\tilde{\theta} = (\mathbf{I} - 11^{\mathrm{T}}\mathbf{D}_\pi)\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{X}\boldsymbol{\beta}, \mathbf{D}_\pi = \frac{1}{n}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{Y}^{\mathrm{dm}}, c_2 = \frac{1}{\sqrt{\frac{1}{n}\tilde{\theta}^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{Y}^{\mathrm{dm}}\tilde{\theta}}}.$$

Note that

$$\mathbf{I} - 11^{\mathrm{T}}\mathbf{D}_\pi = \begin{pmatrix} \hat{\pi}_+ & -\hat{\pi}_+ \\ -\hat{\pi}_- & \hat{\pi}_- \end{pmatrix}, \mathbf{D}_\pi^{-1}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{X} = n(\hat{\boldsymbol{\mu}}_+^{\mathrm{T}}, \hat{\boldsymbol{\mu}}_-^{\mathrm{T}}).$$

Therefore, $\tilde{\theta}^{\mathrm{T}}Y^{\mathrm{dm}^{\mathrm{T}}}\mathbf{X}\boldsymbol{\beta} = n^2\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}$ and $\tilde{\theta}^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{Y}^{\mathrm{dm}}\tilde{\theta} = n^3\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}$. It follows that

$$\hat{\theta}^{\mathrm{T}}\mathbf{Y}^{\mathrm{dm}^{\mathrm{T}}}\mathbf{X}\boldsymbol{\beta} = n\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_b\boldsymbol{\beta}} = n\sqrt{\hat{\pi}_+\hat{\pi}_-}(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\boldsymbol{\beta}.$$

So $\hat{\boldsymbol{\beta}}^{\mathrm{SOS}} = \arg\min_{\boldsymbol{\beta}} L_2(\boldsymbol{\beta}, \lambda)$, where

$$L_2(\boldsymbol{\beta}, \lambda) = -2n(\hat{\pi}_+\hat{\pi}_-)^{1/2}(\hat{\boldsymbol{\mu}}_- - \hat{\boldsymbol{\mu}}_+)^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1. \tag{B.5}$$

Now, for any $\boldsymbol{\beta}$, define $\boldsymbol{\beta}' = \boldsymbol{\beta}/\sqrt{\hat{\pi}_+\hat{\pi}_-}$. Compare (B.5) with (B.1) and it is easy to see that

$$L_2(\boldsymbol{\beta}, \lambda) = (\hat{\pi}_+\hat{\pi}_-)L_3(\boldsymbol{\beta}', \frac{\lambda}{\sqrt{\hat{\pi}_+\hat{\pi}_-}}). \tag{B.6}$$

By (B.6) and the definition of $\hat{\boldsymbol{\beta}}^{\mathrm{DSDA}}$, we have the desired conclusion.

Appendix C

Technical Details in Chapter 4

C.1 Proofs

The following properties of the normal distribution are repeatedly used in our proof.

Proposition C.1.1 *Let $\phi(t)$ and $\Phi(t)$ be the pdf and CDF of $N(0,1)$.*

1. *For $t \geq 1$,*
$$\frac{\phi(t)}{2t} \leq 1 - \Phi(t) \leq \frac{\phi(t)}{t};$$

2. *For $t \geq 0.99$.*
$$\Phi^{-1}(t) \leq \sqrt{2\log(\frac{1}{1-t})};$$

Proposition 1 is an elementary and classic result in probability and hence its proof is omitted for the sake of space.

To prove Theorem 4.5.1, we first study the accuracy of $\hat{h}_j$. The behavior of $\hat{h}_j = \Phi^{-1} \circ \hat{F}_j$ drastically varies on the real line. Define

$$A_n = [-\sqrt{\gamma_1 \log n}, \sqrt{\gamma_1 \log n}], \tag{C.1}$$

where $0 < \gamma_1 < 1$ is a fixed number and n is the sample size. The following lemma shows that $\hat{h}_j(x)$ is an accurate estimator of $h_j(x)$ for $h_j(x) \in A_n$.

Lemma C.1.2 *For sufficiently large n and $0 < \gamma_1 < 1$, we have*

$$\Pr(\sup_{h_j(x)\in A_n} |\hat{h}_j(x) - h_j(x)| \geq \epsilon) \leq 2\exp(-n^{1-\gamma_1}\frac{\epsilon^2}{32\pi^2\gamma_1\log n}) + 2\exp(-\frac{n^{1-\gamma_1}}{16\pi\gamma_1\log n}).$$

Proof [Proof of Lemma C.1.2] By mean value theorem,

$$\hat{h}_j(x) - h_j(x) = (\Phi^{-1})'(\xi)(\hat{F}_j(x) - F_j(x)),$$

for some $\xi \in [\min(\hat{F}_j(x), F_j(x)), \max(\hat{F}_j(x), F_j(x))]$.

First, we bound $|(\Phi^{-1})'(\xi)|$. This is achieved by bounding $F_j(x)$ and $\hat{F}_j(x)$. By definition, for any $h_j(x) \in A_n$,

$$\frac{n^{-\frac{\gamma_1}{2}}}{2\sqrt{2\pi\gamma_1\log n}} \leq \Phi(-\sqrt{\gamma_1\log n}) \leq F_j(x) \leq \Phi(\sqrt{\gamma_1\log n}) \leq 1 - \frac{n^{-\frac{\gamma_1}{2}}}{2\sqrt{2\pi\gamma_1\log n}}.$$

On the other hand, for x such that $h_j(x) \in A_n$

$$\begin{aligned}
& \Pr(\frac{n^{-\frac{\gamma_1}{2}}}{4\sqrt{2\pi\gamma_1\log n}} \leq \hat{F}_j(x) \leq 1 - \frac{n^{-\frac{\gamma_1}{2}}}{4\sqrt{2\pi\gamma_1\log n}}) \\
\geq & \Pr(\sup_{h_j(x)\in A_n} |\tilde{F}_j(x) - F_j(x)| \leq \frac{n^{-\frac{\gamma_1}{2}}}{4\sqrt{2\pi\gamma_1\log n}}) \\
\geq & 1 - 2\exp(-\frac{n^{1-\gamma_1}}{16\pi\gamma_1\log n}),
\end{aligned}$$

where the last inequality follows from Dvoretzky-Kiefer-Wolfowitz (DKW) inequality.

Consequently, with a probability no less than $1 - 2\exp(-\frac{n^{1-\gamma_1}}{16\pi\gamma_1\log n})$,

$$\frac{n^{-\frac{\gamma_1}{2}}}{4\sqrt{2\pi\gamma_1\log n}} \leq \xi \leq 1 - \frac{n^{-\frac{\gamma_1}{2}}}{4\sqrt{2\pi\gamma_1\log n}},$$

and, combining this fact with Proposition C.1.1, we have

$$\begin{aligned}
|(\Phi^{-1})'(\xi)| &= \frac{1}{\phi(\Phi^{-1}(\xi))} = \sqrt{2\pi}\exp(\frac{\Phi^{-1}(\xi)^2}{2}) \\
&\leq \sqrt{2\pi}\exp(\log(4n^{\frac{\gamma_1}{2}}\sqrt{2\pi\gamma_1\log n})) \\
&= 8\pi n^{\frac{\gamma_1}{2}}\sqrt{\gamma_1\log n} \equiv M_n.
\end{aligned}$$

Then

$$\begin{aligned}
& \Pr(\sup_{h_j(x)\in A_n} |\hat{h}_j(x) - h_j(x)| > \epsilon) \\
\leq & \Pr(M_n \sup_{h_j(x)\in A_n} |\hat{F}_j(x) - F_j(x)| > \epsilon) + 2\exp(-\frac{n^{1-\gamma_1}}{16\pi\gamma_1\log n}).
\end{aligned}$$

For the first term on the right hand side,

$$\begin{aligned} & \Pr(M_n \sup_{h_j(x)\in A_n} |\hat{F}_j(x) - F_j(x)| > \epsilon) \\ \leq\ & \Pr(M_n \sup_{h_j(x)\in A_n} |\hat{F}_j(x) - \tilde{F}_j(x)| > \frac{\epsilon}{2}) + \Pr(M_n \sup_{h_j(x)\in A_n} |F_j(x) - \tilde{F}_j(x)| > \frac{\epsilon}{2}). \end{aligned}$$

Because $\sup_{h_j(x)\in A_n} |\hat{F}_j(x) - \tilde{F}_j(x)| \leq \delta_n = \frac{1}{n^2}$, $\delta_n M_n \to 0$ and so the first term is 0 for sufficiently large n. Apply the DKW inequality to the second term and the conclusion follows.

The above lemma guarantees that $\hat{h}_j(X_j)$ is very close to $h_j(X_j)$ on A_n. Now we consider observations in A_n^c. Because such observations are relatively few, Their influence is limited in estimating μ_{yj} and Σ_{jk}. Partition A_n^c to three regions:

$$\begin{aligned} B_n &= [-\gamma_2 \log n, -\sqrt{\gamma_1 \log n}) \cup (\sqrt{\gamma_1 \log n}, \gamma_2 \log n]; \\ C_n &= [-n^{\gamma_3}, -\gamma_2 \log n) \cup (\gamma_2 \log n, n^{\gamma_3}]; \\ D_n &= (-\infty, -n^{\gamma_3}) \cup (n^{\gamma_3}, \infty). \end{aligned}$$

Define $\#B_n = \#\{i : h_j(X_j^i) \in B_n\}$ and $\#C_n$, $\#D_n$ analogously. We have the following lemma.

Lemma C.1.3 *For sufficiently large n and positive constants α_1, α_2 such that $\alpha_1 > 1 - \frac{\gamma_1}{2}$, we have*

$$\sup_{h_j(x)\in B_n} |\hat{h}_j(x) - h_j(x)| \leq 2\sqrt{\log n} + \gamma_2 \log n; \tag{C.2}$$

$$\sup_{h_j(x)\in C_n} |\hat{h}_j(x) - h_j(x)| \leq 2\sqrt{\log n} + n^{\gamma_3}; \tag{C.3}$$

$$\Pr(\#B_n > n^{\alpha_1}) \leq \exp(-\frac{n^{2\alpha_1 - 1}}{4}); \tag{C.4}$$

$$\Pr(\#C_n > n^{\alpha_2}) \leq \exp(-\frac{n^{2\alpha_2 - 1}}{4}); \tag{C.5}$$

$$\Pr(\#D_n > 1) \leq \frac{2n^{1-\gamma_3}}{\sqrt{2\pi}} \exp(-\frac{n^{2\gamma_3}}{2}). \tag{C.6}$$

Proof [Proof of Lemma C.1.3] Equations (C.2)–(C.3) are direct consequences of the definitions of $\hat{h}$ and B_n, C_n. Indeed, because $\hat{F} < 1 - \delta_n$, by Proposition C.1.1, for

$x \in B_n \cup C_n$

$$|\hat{h}_j(x)| \leq \Phi^{-1}(1-\delta_n) \leq \sqrt{2\log\frac{1}{\delta_n}} = 2\sqrt{\log n}.$$

Combining this bound with the definitions of B_n, C_n, we have the desired conclusions.

For (C.4), note that, for sufficiently large n,

$$\Pr(h_j(X_j) \in B_n) \leq 2\Pr(h_j(X_j) > \sqrt{\gamma_1 \log n}) \leq \frac{\sqrt{2}n^{-\frac{\gamma_1}{2}}}{\sqrt{\pi\gamma_1 \log n}} \leq n^{-\frac{\gamma_1}{2}}.$$

Therefore, by Hoeffding's inequality

$$\begin{aligned}
& \Pr(\#B_n > n^{\alpha_1}) \\
\leq\ & \Pr(\sum_{i=1}^{n}[(I(h_j(X_j^i) \in B_n)) - \Pr(h_j(X_j^i) \in B_n)] > n^{\alpha_1} - n^{1-\frac{\gamma_1}{2}}) \\
\leq\ & \exp(-\frac{n^{2\alpha_1-1}(1-n^{1-\frac{\gamma_1}{2}-\alpha_1})^2}{2}) \\
\leq\ & \exp(-\frac{n^{2\alpha_1-1}}{4}),
\end{aligned}$$

for sufficiently large n.

For (C.5), note that

$$\Pr(h_j(X_j^i) \in C_n) \leq \frac{2n^{-\frac{\gamma_2^2 \log n}{2}}}{\gamma_2 \log n}.$$

So (C.5) can be proven similarly.

For (C.6),

$$\Pr(\#D_n > 1) \leq 2n\Pr(h_j(X_j^i) > n^{\gamma_3}) \leq \frac{2n^{1-\gamma_3}}{\sqrt{2\pi}}\exp(-\frac{n^{2\gamma_3}}{2}).$$

Proof [Proof of Theorem 4.5.1] We first prove (4.10).

$$\begin{aligned}
\Pr(|\hat{\mu}_j - \mu_j| > \epsilon) &\leq \Pr(\frac{1}{n}\sum_{i=1}^{n}|\hat{h}_j(X_j^i) - h_j(X_j^i)| > \frac{\epsilon}{2}) + \Pr(|\frac{1}{n}\sum_{i=1}^{n}h_j(X_j^i) - \mu_j| > \frac{\epsilon}{2}) \\
&\equiv L_1 + L_2.
\end{aligned}$$

By the Chernoff bound, $L_2 \leq 2\exp(-Cn\epsilon^2)$.

$$
\begin{aligned}
L_1 \;\leq\; & \Pr(\sup_{h_j(x)\in A_n} |\hat{h}_j(x) - h_j(x)| > \frac{\epsilon}{8}) + \Pr(\frac{\#B_n}{n} \sup_{h_j(x)\in B_n} |\hat{h}_j(x) - h_j(x)| > \frac{\epsilon}{8}) \\
+ \;& \Pr(\frac{\#C_n}{n} \sup_{h_j(x)\in C_n} |\hat{h}_j(x) - h_j(x)| > \frac{\epsilon}{8}) + \Pr(\frac{\#D_n}{n} \sup_{h_j(x)\in D_n} |\hat{h}_j(x) - h_j(x)| > \frac{\epsilon}{8}).
\end{aligned}
$$

By Lemma C.1.3, it can be checked that, under Condition (C1), if $\#B_n \leq n^{\alpha_1}$ and $\#D_n = 0$ then

$$
\begin{aligned}
\Pr(\frac{\#B_n}{n} \sup_{h_j(x)\in B_n} |\hat{h}_j(x) - h_j(x)| > \frac{\epsilon}{8}) = 0, \\
\Pr(\frac{\#D_n}{n} \sup_{h_j(x)\in D_n} |\hat{h}_j(x) - h_j(x)| > \frac{\epsilon}{8}) = 0,
\end{aligned}
$$

for sufficiently large n. If $\gamma_3 + \alpha_2 < 1$, similarly we have

$$
\Pr(\frac{\#C_n}{n} \sup_{h_j(x)\in C_n} |\hat{h}_j(x) - h_j(x)| > \frac{\epsilon}{8}) = 0.
$$

It follows that, if $\alpha_1 < 1$ and $\gamma_3 + \alpha_2 < 1$, then we have

$$
L_1 \leq 4\exp(-Cn^{1-\gamma_1}\frac{\epsilon^2}{\gamma_1}) + \exp(-Cn^{2\alpha_1-1}) + \exp(-Cn^{2\alpha_2-1}) + \frac{2n^{1-\gamma_3}}{\sqrt{2\pi}}\exp(-\frac{n^{2\gamma_3}}{2}),
$$

Take $\gamma_1 = \theta, \alpha_1 = 1 - \frac{\theta}{4}, \alpha_2 = \frac{3}{4} - \frac{\theta}{2}, \gamma_3 = \frac{1}{4} - \frac{\theta}{2}$ and the conclusion follows.

Now we prove (4.11). By the proof in [87], it suffices to bound

$$
\Pr(|\frac{1}{n}\sum_{i=1}^{n} h_j(X_j^i)(\hat{h}_k(X_k^i) - h_k(X_k^i))| > \epsilon).
$$

We can decompose the summation into four terms.

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} h_j(X_j^i)(\hat{h}_k(X_k^i) - h_k(X_k^i)) \\
= \;& \frac{1}{n}\Big(\sum_{h_j(X_j^i)\in D_n \text{ or } h_k(X_k^i)\in D_n} + \sum_{h_j(X_j^i)\notin D_n, h_k(X_k^i)\in C_n} \\
& + \sum_{h_j(X_j^i)\in A_n\cup B_n, h_k(X_k^i)\in B_n} + \sum_{h_j(X_j^i)\in A_n, h_k(X_k^i)\in A_n} \Big)(h_j(X_j^i)(\hat{h}_k(X_k^i) - h_k(X_k^i))) \\
\equiv \;& S_1 + S_2 + S_3 + S_4.
\end{aligned}
$$

Write $\#D_{nj} = \#\{i : h_j(X_j^i) \in D_n\}$. Then

$$\begin{aligned}\Pr(|S_1| > \epsilon) &\leq \Pr(\#D_{nj} > 1) + \Pr(\#D_{nk} > 1) \\ &\leq \frac{4n^{1-\gamma_3}}{\sqrt{2\pi}} \exp(-\frac{n^{2\gamma_3}}{2}).\end{aligned}$$

Note that, for a pair of α_2, γ_3, such that $\alpha_2 + 2\gamma_3 - 1 < 0$, we have $n^{\alpha_2+2\gamma_3-1} \to 0$. Therefore, for sufficiently large n,

$$\begin{aligned}\Pr(|S_2| > \epsilon) &\leq \Pr(\frac{1}{n} \sum_{h_k(X_k^i) \in C_n} |\hat{h}_k(X_k^i) - h_k(X_k^i)| > \frac{\epsilon}{n^{\gamma_3}}) \\ &\leq \Pr(\#C_n > n^{\alpha_2}) + \Pr(n^{\alpha_2-1}(2\sqrt{\log n} + n^{\gamma_3}) > \frac{\epsilon}{n^{\gamma_3}}) \\ &\leq \exp(-\frac{n^{2\alpha_2-1}}{4}) + 0,\end{aligned}$$

Similarly, for $0 < \alpha_1 < 1$,

$$\begin{aligned}\Pr(|S_3| > \epsilon) &\leq \Pr(\#B_n > n^{\alpha_1}) + \Pr(n^{\alpha_1-1}(\gamma_2 \log n)(2\sqrt{\log n} + \gamma_2 \log n) > \epsilon) \\ &\leq \exp(-\frac{n^{2\alpha_1-1}}{4}) + 0,\end{aligned}$$

where $0 < \alpha_1 < 1$. Finally,

$$\begin{aligned}\Pr(|S_4| > \epsilon) &\leq \Pr(\sup_{h_k(X_k^i) \in A_n} |\hat{h}_k(X_k^i) - h_k(X_k^i)| > \frac{\epsilon}{\sqrt{\gamma_1 \log n}}) \\ &\leq 4 \exp(-C \frac{n^{1-\gamma_1}\epsilon^2}{\gamma_1^2 \log^2 n}).\end{aligned}$$

Pick $\gamma_1 = \rho, \gamma_3 = \frac{1}{6} - \rho, \alpha_2 = \frac{2}{3} - \frac{\rho}{2}, \alpha_1 = 1 - \frac{\rho}{2}$ and the conclusion follows.

Proof [Proof of Corollary 4.5.3] Note that n_+ is a summation of n i.i.d random variables with distribution Bernoulli$(1, \pi_+)$. Therefore, by Chernoff bound, there exists $c > 0$ such that $\Pr(n_+ > \frac{\pi_+}{2}n) > 1 - 2\exp(-cn)$. Hence, by Theorem 4.5.1,

$$\Pr(|\hat{\mu}_{+j} - \mu_{+j}| \geq \frac{\epsilon}{2}) < \zeta_1^*(\frac{\sqrt{\pi_+}\epsilon}{2}) + 2\exp(-cn).$$

Similarly,

$$\Pr(|\hat{\mu}_{-j} - \mu_{-j}| \geq \frac{\epsilon}{2}) < \zeta_1^*(\frac{\sqrt{\pi_-}\epsilon}{2}) + 2\exp(-cn).$$

Hence, we have (4.12). Equation (4.13) can be proven similarly.

Proof [Proof of Theorem 4.5.4 and Theorem 4.5.5] By [7], the consistency is implied by accurate estimators of $\hat{\mu}_y$, $\hat{C}_{ij}$. Therefore, Theorem 4.5.4 can be proven by following the proof in their paper and applying Corollary 4.5.3.

Theorem 4.5.5 is direct consequence of Theorem 4.5.4. Hence, the proof is omitted here for the sake of space.

Lemma C.1.4 *For any* $\epsilon < \min\{\epsilon_0, \frac{\lambda}{2\phi\Delta_1}, \lambda\}$ *and large enough* n *such that* $\epsilon \gg sn^{-1/4}$, *we have*

1.

$$\Pr(\|\hat{\beta}_A - \beta_A\|_1 \geq \epsilon) \leq 2s^2\zeta_2(\frac{\epsilon}{s}) + 2s\zeta_1(\frac{\epsilon}{s}). \tag{C.7}$$

2. If we further assume that $0 < c < \pi_+, \pi_- < C$, *then*

$$\begin{aligned}\Pr(|\hat{\beta}_0 - \beta_0| \geq C\epsilon) \leq\ & 2\exp(-Cn) + Cs\zeta_1(\frac{\epsilon}{s(\phi\Delta_1 + \Delta_2)}) \\ & +2p\zeta_1(\frac{\lambda(1-\kappa+2\epsilon\phi)}{4(1+\kappa)}) + 2s^2\zeta_2(\frac{\epsilon}{s\Delta_2}) + 2ps\zeta_2(\frac{\epsilon}{s})\end{aligned} \tag{C.8}$$

Proof We first prove (C.7). Similar to the proof of Conclusion 3, Theorem 1 in [7], we have

$$\|\hat{\beta}_A - \beta_A\|_1 \leq \frac{1}{1-\eta_1\phi}(\frac{\lambda}{2} + \phi\|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_1 + \phi^2\eta_1\Delta_1) \tag{C.9}$$

where $\eta_1 = \|C_{AA} - C_{AA}^{(n)}\|_\infty$. Under the events $\eta_1 < \epsilon$ and $\|(\hat{\boldsymbol{\mu}}_{+A} - \hat{\boldsymbol{\mu}}_{-A}) - (\boldsymbol{\mu}_{+A} - \boldsymbol{\mu}_{-A})\|_1 < \epsilon$ we have $\|\hat{\beta}_A - \beta_A\|_1 \leq \epsilon$. Hence, (C.7) follows.

For (C.8), assume that $\hat{\beta}_{A^C} = 0$. Then we have

$$\begin{aligned}|\hat{\beta}_0 - \beta_0| &= |(\log\frac{n_+}{n_-} - \log\frac{\pi_2}{\pi_1}) - \frac{(\hat{\mu}_{1A} + \hat{\mu}_{1A})^{\mathrm{T}}\hat{\beta}_A}{2}| \\ &\leq |\log\hat{\pi}_+ - \log\pi_-| + |\log\hat{\pi}_1 - \log\pi_1| \\ &\quad +\frac{1}{2}|((\hat{\mu}_{1A} + \hat{\mu}_{2A}) - (\mu_{1A} + \mu_{2A}))^{\mathrm{T}}(\hat{\beta}_A - \beta_A)| \\ &\quad +|(\mu_{1A} + \mu_{2A})^{\mathrm{T}}(\hat{\beta}_A - \beta_A)| + \frac{1}{2}|(\mu_{+A} + \mu_{-A})^{\mathrm{T}}(\hat{\beta}_A - \beta_A)|\end{aligned}$$

Under the events $|\hat{\pi}_j - \pi_j| \leq \min\{\frac{c}{2}, \frac{2\epsilon}{c}\}$, $\|\hat{\mu}_{jA} - \mu_{jA}\|_1 \leq \frac{\epsilon}{\phi\Delta_1}$ and $\|\hat{\beta}_A - \beta_A\|_1 \leq \frac{\epsilon}{\Delta_2}$, we have $|\hat{\beta}_0 - \beta_0| \leq C\epsilon$.

Proof [Proof of Theorem 4.5.7] Note that

$$\begin{aligned} R_n &\leq 1 - \Pr(Y = \text{sign}(h(X)^{\mathrm{T}}\beta + \beta_0), \text{sign}(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) = \text{sign}(h(X)^{\mathrm{T}}\beta + \beta_0)) \\ &\leq R + \Pr(\text{sign}(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) \neq \text{sign}(h(X)^{\mathrm{T}}\beta + \beta_0)) \end{aligned}$$

Therefore,

$$\begin{aligned} R_n - R &\leq \Pr(\text{sign}(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) \neq \text{sign}(h(X)^{\mathrm{T}}\beta + \beta_0)) && \text{(C.10)} \\ &\leq \Pr(|h(X)^{\mathrm{T}}\beta + \beta_0| \leq \epsilon) && \text{(C.11)} \\ &\quad + \Pr(|(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) - (h(X)^{\mathrm{T}}\beta + \beta_0)| \geq \frac{\epsilon}{2}) \end{aligned}$$

Now

$$\Pr(|h(X)^{\mathrm{T}}\beta + \beta_0| \leq \epsilon) \leq \frac{C\epsilon}{\sqrt{2\pi}} \qquad \text{(C.12)}$$

For the second term, assume that $\hat{\beta}_{A^C} = 0$, $|\hat{\beta}_0 - \beta_0| \leq C\epsilon$, $\|\hat{\beta}_A - \beta_A\|_1 \leq \frac{\epsilon}{\sqrt{\log n}}$ and $\sup_{t \in A_n} |\hat{h}_j(t) - h_j(t)| \leq C\frac{\epsilon}{\phi\Delta_1}$ for all j, where A_n is defined as in (C.1). Then

$$\begin{aligned} & |(\hat{h}(X_A)^{\mathrm{T}}\hat{\beta}_A + \hat{\beta}_0) - (h(X_A)^{\mathrm{T}}\beta_A + \beta_0)| && \text{(C.13)} \\ \leq\ & |\hat{\beta}_0 - \beta_0| + \|\hat{h}(X_A)\|_\infty \|\hat{\beta}_A - \beta_A\|_1 + \|\hat{h}(X_A) - h(X_A)\|_\infty \|\beta_A\|_1 && \text{(C.14)} \\ \leq\ & |\hat{\beta}_0 - \beta_0| + 2\sqrt{\log n}\|\hat{\beta}_A - \beta_A\|_1 + \phi\Delta_1\|\hat{h}(X_A) - h(X_A)\|_\infty, && \text{(C.15)} \end{aligned}$$

which is smaller than ϵ as long as $h_j(X_j) \in A_n$ for all j. Therefore, take $\gamma_1 = 1/2$ in A_n, we have

$$\Pr(|(\hat{h}(X)^{\mathrm{T}}\hat{\beta} + \hat{\beta}_0) - (h(X)^{\mathrm{T}}\beta + \beta_0)| \geq \frac{\epsilon}{2}) \leq \Pr(\cup_{j \in A} h_j(X_j) \in A_n) \leq \frac{Csn^{-1/4}}{\sqrt{\log n}}, \qquad \text{(C.16)}$$

which will be smaller than ϵ for sufficiently large n.

Therefore, by Lemma C.1.2, (C.7), (C.8), we have the desired conclusion.

Appendix D

Technical Details in Chapter 5

D.1 Proofs

To prove Theorem 5.3.1, we first establish a concentration inequality for K_{nj} which is built upon the Dvoretzky–Kiefer–Wolfowitz inequality [98, 103, 104].

Lemma D.1.1 *For any* $\epsilon > 0$, $\Pr(|K_{nj} - K_j| > \epsilon) \leq \zeta(\epsilon)$.

Proof [of Lemma D.1.1] We use the Dvoretzky–Kiefer–Wolfowitz inequality in the proof. For the distribution function F of a random variable, $-\infty < X < \infty$, and its sample estimate $\hat{F}$, we have, for any $\epsilon > 0$ $\Pr\{\sup|\hat{F}(x) - F(x)| > \epsilon\} \leq 2\exp(-2n\epsilon^2)$. Hence, for any $\epsilon > 0$,

$$\begin{aligned}
\Pr(|K_{nj} - K_j| > \epsilon \mid Y) &\leq \Pr\{\sup|\hat{F}_{+j}(x) - F_{+j}(x)| + \sup|\hat{F}_{-j}(x) - F_{-j}(x)| > \epsilon \mid Y\} \\
&\leq \sum_{y=+1,-1} \Pr\{\sup|\hat{F}_{yj}(x) - F_{yj}(x)| > \epsilon/2 \mid Y\} \\
&\leq 2\exp(-n_+\epsilon^2/2) + 2\exp(-n_-\epsilon^2/2).
\end{aligned}$$

As n_+ and n_- are both sums of n independent and identically distributed Bernoulli random variables, by the Chernoff bound, there exist constants $c_1, c_2 > 0$ such that $\Pr(n_+ < \pi_+ n/2) < \exp(-c_1 n\pi_+^2/4)$ and $\Pr(n_- < \pi_- n/2) < \exp(-c_1 n\pi_-^2/4)$. Consequently, we have

$$\Pr(|K_{nj} - K_j| > \epsilon) = E\{\Pr(|K_{nj} - K_j| > \epsilon \mid Y)\} \leq \zeta(\epsilon).$$

Proof [of Theorem 5.3.1] We only prove part (b). Then part (a) can be proved by setting $\tilde{S} = D$. Let $\tilde{s} = |\tilde{S}|$. Note that $\hat{S}(\tilde{s}) = \tilde{S}$ under the event that $\max_{j=1,\ldots,p} |K_{nj} - K_j| < \delta_{\tilde{S}}/2$, because we must have $\min_{j\in\tilde{S}} K_{nj} > \max_{j\in\tilde{S}^c} K_{nj}$. Moreover, for any $d_n \geq \tilde{s}$, $\hat{S}(\tilde{s}) \subseteq \hat{S}(d_n)$, which implies $D \subseteq \tilde{S} \subseteq \hat{S}(d_n)$. On the other hand, we have

$$\Pr(\max_{j=1,\ldots,p} |K_{nj} - K_j| > \delta_{\tilde{S}}/2) \geq 1 - \sum_{j=1}^{p} \Pr(|K_{nj} - K_j| > \delta_{\tilde{S}}/2) \geq 1 - p\zeta(\delta_{\tilde{S}}/2),$$

where the last inequality follows Lemma D.1.1. This completes the proof.

Proof [of Lemma 5.3.2] By $K_j = 1 - 2\Phi(|\mu_{+j} - \mu_{-j}|/2)$, so it suffices to show that $\mu_{+j} = \mu_{-j}$ for $j \in D^c$ in order to prove Conclusion (a). For that, we write $\mu_+ - \mu_- = \Sigma\beta$. By the blockwise independence assumption on Σ and $\beta_j = 0$ for all $j \in D^c$, we must have $\mu_{+j} = \mu_{-j}$ for $j \in D^c$.

For Conclusion (b), note that, for $j > d$, $|\mu_{+j} - \mu_{-j}| = |\rho^{j-d} \sum_{i=1}^{d} \rho^{|i-d|}\beta_i| = |\rho|^{|j-d|}|\mu_{+d} - \mu_{-d}|$. It follows that $|\mu_{+j} - \mu_{-j}| < |\mu_{+(d+k^*)} - \mu_{-(d+k^*)}|$ for $j > d + k^*$ and $\min_{j\in\tilde{S}_1(k^*)} |\mu_{+j} - \mu_{-j}| = |\mu_{+(d+k^*)} - \mu_{-(d+k^*)}|$. Thus $\delta_{\tilde{S}_1(k^*)} = \min_{j\in\tilde{S}_1(k^*)} |\mu_{+j} - \mu_{-j}| - \max_{j\notin\tilde{S}_1(k^*)} |\mu_{+j} - \mu_{-j}| > 0$.

To prove Conclusion (c), we write $\Sigma = (1-\rho)I + \rho J$ where J is a $p \times p$ matrix of 1. Then $\Sigma^{-1} = (1-\rho)^{-1}I - \rho[\{1 + (p-1)\rho\}(1-\rho)]^{-1}J$. Write $c = 1^{\mathrm{T}}(\mu_+ - \mu_-) = \sum_{j\in\tilde{S}_2}(\mu_{+j} - \mu_{-j})$. For any $j \in \tilde{S}_2^C$, we have $\beta_j = -\rho[\{1 + (p-1)\rho\}(1-\rho)]^{-1}c$. Thus, $D \subseteq \tilde{S}_2 \Leftrightarrow 1^{\mathrm{T}}(\mu_+ - \mu_-) = 0$.