ESTIMATION OF SPARSE LOW-DIMENSIONAL LINEAR PROJECTIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Irina Gaynanova

May 2015

ESTIMATION OF SPARSE LOW-DIMENSIONAL LINEAR PROJECTIONS

Irina Gaynanova, Ph.D.

Cornell University 2015

Many multivariate analysis problems are unified under the framework of linear projections. These projections can be tailored towards the analysis of variance (principal components), classification (discriminant analysis) or network recovery (canonical correlation analysis). Traditional techniques form these projections by using all of the original variables, however in recent years there has been a lot of interest in performing variable selection. The main goal of this dissertation is to elucidate some of the fundamental issues that arise in high-dimensional multivariate analysis and provide computationally efficient and theoretically sound alternatives to existing heuristic techniques

BIOGRAPHICAL SKETCH

Irina Gaynanova was born and raised in Russia, where she earned a Diploma with honors (M.S. equivalent) in Applied Mathematics and Computer Science from the Lomonosov Moscow State University in 2009. She joined the MS/PhD program in Statistics at Cornell University in 2010 and received a M.S. in Statistics in 2013. She graduates with PhD in Statistics in 2015, under the direction of her advisors Prof. James Booth and Prof. Martin Wells. Irina's research interests include high-dimensional data analysis, multivariate analysis, statistical methods for analyzing biological data, computational statistics and machine learning. After graduation, Irina will join the Department of Statistics at Texas A&M University as an Assistant Professor.

To my husband Evan, and to my parents, Tanya and Valeriy.

ACKNOWLEDGEMENTS

I would like to thank my advisors, Jim Booth and Marty Wells, for their continuous support and encouragement during my years as a PhD student. Your doors were always open to me, and I have benefited greatly from your advice on research, writing, presenting, teaching and various other aspects of academic career. I am very thankful to you for stimulating my curiosity and for giving me freedom to explore a variety of research directions, it was invaluable for my development as an independent scholar.

I am grateful to my committee members, Jason Mezey and Marten Wegkamp, for sharing their valuable expertise and insights. Our conversations were fundamental to my desire to strive for a balance between application-driven research and theoretically sound methodology. I also like to thank Jacob Bien, Mladen Kolar, Johannes Lederer for fruitful research discussions and subsequent collaborations that resulted from them, and Françoise Vermeylen for generously sharing her expertise and advice in becoming a better statistical consultant and improving the communication with clients.

I am thankful to my fellow graduate students, in particular Didier Chetelat, Ben Risk, Jon Steingrimsson and Inder Tecuapetla, for being wonderful office mates, for inspiring me to be productive by setting your own example and for supporting me in times of failure.

Finally, I am grateful to my beloved husband Evan for his support and patience throughout these years. I would of not been able to do this without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Recent technological advances have generated high-dimensional data sets across a wide variety of application areas such as finance, atmospheric science, astronomy, biology and medicine. Not only do these data sets provide computational challenges, but they also motivate new statistical challenges as the traditional methods are no longer sufficient.

In this work we focus on classical multivariate analysis tools, which can be used for the analysis of variance (principal components), classification (discriminant analysis) or network recovery (canonical correlation). These methods perform poorly when applied to modern datasets due to the difficulty in estimation of large covariance matrix and over-selection of relevant features. In the literature, these problems have been addressed separately, however their joint consideration leads to significant improvements in terms of empirical performance, computational speed and subsequent scientific importance. While each multivariate technique is different, we consider them under the unified framework of sparse linear projections. These projections serve two main functions: the graphical visualization of the complex data tailored to the research question and the identification of key features that codify the underlying structure. Having these objectives, we focus on developing computationally efficient and theoretically sound statistical methods that are appropriate for the analysis of the data at hand. Our primary goal is to aid the discovery of scientifically meaningful low-dimensional structures in high-dimensional data.

1.2 Review of generalized eigenvalue problems

Let $Q \in \mathbb{R}^{p\times p}$ (for quadratic function) and $C \in \mathbb{R}^{p\times p}$ (for constraint) be two symmetric, semi positive-definite matrices. In addition, let C be strictly positive-definite. Consider the optimization problem:

$$v = \arg\max_{v\in\mathbb{R}^p} \left\{ v^\top Q v \right\} \quad \text{subject to} \quad v^\top C v \leq 1. \tag{1.1}$$

Problem (2.1) is called the generalized eigenvalue problem [61], since the maximum is achieved when v is the leading eigenvector of matrix $C^{-1}Q$. Generalized eigenvalue problems are the core of many multivariate analysis methods, we review the most common examples below.

1.2.1 Principal Component Analysis

Principal Component Analysis (PCA) seeks linear combinations of features that explain maximal variability in the data. Let $X \in \mathbb{R}^{n\times p}$ be the centered data matrix of n independent observations $X_i \in \mathbb{R}^p$. The lth principal component loading $v^{(l)}$ is defined as

$$v^{(l)} = \arg\max_{v\in\mathbb{R}^p} \left\{ \frac{1}{n} v^\top X^\top X v \right\} \quad \text{subject to} \quad v^\top v \leq 1, \quad v^\top v^{(k)} = 0 \quad \text{for all} \quad k < l.$$

This is the generalized eigenvalue problem (2.1) with $Q = \frac{1}{n}X^\top X$ and $C = I$. The most common applications of PCA are clustering and dimension reduction.

1.2.2 Discriminant analysis

Consider a problem of multi-group classification: given n independent observations $\{(X_i, Y_i), i = 1, \ldots, n\}$ from a joint distribution (X, Y) on $\mathbb{R}^p \times \{1, \ldots, G\}$, our goal is to learn a rule that will classify a new data point $X \in \mathbb{R}^p$ into one of the G groups.

Fisher's Linear Discriminant Analysis (FLDA) is a classical approach for obtaining a classification rule in the multi-group setting. It seeks the linear combinations of features that maximize Between Group Variability with respect to Within Group Variability [38, Chapter 11]. These linear combinations are called *canonical vectors* and they provide a low-dimensional representation of the data by reducing the original feature space dimension p to $G - 1$, where G is the total number of groups.

Denote by n_g the number of samples from the group g, $n_g = \#\{i \mid Y_i = g\}$, and the sample average in the group g as $\bar{X}_g = n_g^{-1} \sum_{i|Y_i=g} X_i$. Let W be a pooled sample covariance matrix,

$$W = (n - G)^{-1} \sum_{g=1}^{G} (n_g - 1) S_g, \tag{1.2}$$

where $S_g = (n_g - 1)^{-1} \sum_{i|Y_i=g} (X_i - \bar{X}_g)(X_i - \bar{X}_g)$. Furthermore, let B be the between-group sample covariance matrix,

$$B = \frac{1}{n} \sum_{g=1}^{G} n_g (\bar{X}_g - \bar{X})(\bar{X}_g - \bar{X})^\top, \tag{1.3}$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. FLDA estimates vectors $\{v_g\}_{g=1}^{G-1}$, which are linear com-

binations of p variables, through the following optimization program

$$\begin{aligned} v_g = \arg\max_{v \in \mathbb{R}^p} \left\{ v^\top B v \right\} \\ \text{s.t.} \quad v^\top W v = 1; \\ v^\top W v_{g'} = 0 \quad \text{for} \quad g' < g. \end{aligned} \tag{1.4}$$

This is a generalized eigenvalue problem (2.1) with $Q = B$ and $C = W$.

Given the matrix $V \in R^{p \times (G-1)}$ of vectors $\{v_g\}_{g=1}^{G-1}$, a new data point $X \in \mathbb{R}^p$ is classified into group $\hat{g}$ if

$$\hat{g} = \arg\min_{g \in \{1,\ldots,G\}} (X - \bar{X}_g)^\top V (V^\top W V)^{-1} V^\top (X - \bar{X}_g) - 2 \log \frac{n_g}{n}. \tag{1.5}$$

This rule is a sample version of the optimal classification rule derived under the assumption of multivariate Gaussian class-conditional distributions with a common covariance matrix [39, Chapter 3.9].

1.2.3 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a standard multivariate analysis tool that is used to find linear combinations of two sets of features with the maximum correlation.

Consider n iid observations $X_i \in \mathbb{R}^{p_1+p_2}$ with $X_i = (X_{1i} \quad X_{2i})$ and $X_{1i} \in \mathbb{R}^{p_1}$, $X_{2i} \in \mathbb{R}^{p_2}$. Let $\mathrm{Cov}(X_i) = \Sigma$ and partition it as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{11} = \mathrm{Cov}(X_1) \in \mathbb{R}^{p_1 \times p_1}$, $\Sigma_{22} = \mathrm{Cov}(X_2) \in \mathbb{R}^{p_2 \times p_2}$ and $\Sigma_{12} = \mathrm{Cov}(X_1, X_2) \in \mathbb{R}^{p_1 \times p_2}$.

Population CCA seeks linear combination $w = (w_1^\top, w_2^\top) \in \mathbb{R}^p$ such that

$$(w_1, w_2) = \arg \max_{w_1 \in \mathbb{R}^{p_1}, w_2 \in \mathbb{R}^{p_2}} w_1^\top \Sigma_{12} w_2 \quad \text{subject to} \quad w_1^\top \Sigma_{11} w_1 = 1, w_2^\top \Sigma_{22} w_2 = 1. \tag{1.6}$$

Let $K = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}$, set $N_1 = KK^\top = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^\top \Sigma_{11}^{-1/2}$ and $N_2 = K^\top K$. Then $w_1 = \Sigma_{11}^{-1/2} u_1$ and $w_2 = \Sigma_{22}^{-1/2} v_1$, where u_1 is the first eigenvector of N_1 and v_1 is the first eigenvector of N_2. Hence,

$$w_1 = \arg \max_{v \in \mathbb{R}^{p_1}} v^\top \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^\top v \quad \text{subject to} \quad v^\top \Sigma_{11} v = 1, \tag{1.7}$$

$$w_2 = \arg \max_{v \in \mathbb{R}^{p_2}} v^\top \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12} v \quad \text{subject to} \quad v^\top \Sigma_{22} v = 1. \tag{1.8}$$

Therefore, (1.7) is a generalized eigenvalue problem (2.1) with $Q = \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^\top$ and $C = \Sigma_{11}$, and (1.8) is a generalized eigenvalue problem (2.1) with $Q = \Sigma_{12}^\top \Sigma_{11}^{-1} \Sigma_{12}$ and $C = \Sigma_{22}$.

Sample CCA solves (1.6) by substituting sample covariance matrix S instead of Σ. Usually, the datasets X_1 and X_2 are centered and scaled, so that $S_{11} = \frac{1}{n} X_1^\top X_1$, $S_{22} = \frac{1}{n} X_2^\top X_2$ and $S_{12} = \frac{1}{n} X_1^\top X_2$.

1.3 Thesis overview

The remainder of this thesis is organized as follows. Chapter 2 focuses on the variable selection properties of generalized eigenvalue problem (2.1) in high-dimensional settings. A common trend in the literature is to perform variable selection through the addition of an ℓ_1 penalty to the objective function in (2.1). We show that this approach doesn't work well for the generalized eigenvalue problems, as it fails to provide interpretable solutions, which are crucial for scientific discovery. We demonstrate that it is possible to obtain interpretable so-

lutions by considering an ℓ_1 constraint instead of an ℓ_1 penalty. We contrast the two techniques and discover that they have different variable selection performance in nonconvex settings. We prove, both empirically and theoretically, that an ℓ_1 penalty can fail to generate solutions with pre-defined level of sparsity, while an ℓ_1 constraint has no such drawback.

Chapter 3 describes a novel methodology for discriminant analysis in high-dimensional settings. It has been observed that classical Fisher's linear discriminant analysis performs poorly when the number of samples is small compared to the number of variables [7, 52]. The existing high-dimensional approaches provided inadequate solutions due to unrealistic assumption of independence between the features and lack of theoretical support. Driven by the desire to achieve both computational efficiency and theoretical guarantees, we have investigated a possibility of combining variable selection and covariance estimation through the convex optimization framework. The proposed method in Chapter 3 achieves the state of the art performance for multi-group discriminant analysis by directly estimating the canonical vectors using empirical risk minimization. The approach is based on the geometric observation that canonical vectors can be expressed in a closed form (through the covariance matrix Σ and the mean vectors μ_g) up to orthogonal rotation that affects neither the classification rule nor the sparsity pattern. We develop a computationally efficient block-coordinate descent algorithm and provide sound statistical guarantees on the variable selection and classification consistency, in contrast to other existing multi-group approaches.

Chapter 4 focuses on the optimal variable selection rate for multi-group classification using discriminant analysis. Several multi-group methods have been

proposed in the literature, however their variable selection performance is either unknown or suboptimal to the two-group case. We provide sharp conditions for the consistent recovery of relevant variable using the methodology of Chapter 3. The resulting rates of convergence attain the optimal scaling of the sample size n, number of variables p and the sparsity level s. These rates are significantly faster than the best known results in the multi-group case. Moreover, they coincide with the optimal minimax rates for the two-group case. The theoretical results are validated with numerical analysis.

Chapter 5 summarizes the main findings of this thesis and outlines directions for future research.

1.4 Notation

For a vector $v \in \mathbb{R}^p$ we define $\|v\|_2 = \sqrt{\sum_{i=1}^p v_i^2}$, $\|v\|_1 = \sum_{i=1}^p |v_i|$ and $\|v\|_\infty = \max_i |v_i|$. We use e_j to define a unit norm vector with jth element being equal to one. For a matrix $M \in \mathbb{R}^{n \times p}$ we define by m_i the ith row of M and by M_j the jth column of M. We also define $\|M\|_\infty = \max_{i=1,..,n} \|m_i\|_1$, $\|M\|_{\infty,2} = \max_{i=1,..,n} \|m_i\|_2$, $\|M\|_1 = \sum_{i=1}^n \sum_{j=1}^p |m_{ij}|$, $\|M\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^p m_{ij}^2}$, $\|M\|_2 = \Sigma_1(M)$ and $\|M\|_* = \sum_{i=1}^{\min(n,p)} \Sigma_i(M)$, where $\Sigma_i(M)$ is the ith singular value of M. Given an index set A, we define M_{AA} to be the submatrix of M with rows and columns indexed by A. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = \mathcal{O}(b_n)$ to define $a_n < Cb_n$ for some positive constant C. We write $a_n = o(b_n)$ to define $a_n b_n^{-1} \to 0$. We define $\mathbb{O}^p$ to be the space of $p \times p$ matrices $R \in \mathbb{O}^p$ such that $RR^\top = R^\top R = I$.

CHAPTER 2

PENALIZED VERSUS CONSTRAINED GENERALIZED EIGENVALUE PROBLEMS

2.1 Introduction

There has been a lot of recent interest in extending the traditional multivariate analysis techniques to high-dimensional settings. A common strategy is to enforce a low-dimensional structure by performing variable selection. With the success of the LASSO [56], it is very common to achieve this goal by adding the ℓ_1 penalty to the objective function [2, 53, 59, 70, 76]. One such method is the penalized discriminant analysis approach of [69], which inspired this work. It has been observed in simulations [37] and data applications [69] that the method performs poorly in terms of variable selection, consistently selecting a much larger number of features than the competitors, sometimes more than 90% of the original features. No sound explanation has been given for this phenomenon.

In this chapter we demonstrate that the core reason for the poor variable selection performance in a variety of settings is the nonconvexity of the underlying optimization problem. The ℓ_1 penalty is motivated by the dual problem in LASSO, where using an ℓ_1 constraint geometrically means projecting the solution vector onto the polytope that forces certain components to be exactly zero. The constraint tuning parameter controls the level of sparsity; the larger the constraint, the smaller the number of selected features. The convexity of LASSO ensures the solutions to the ℓ_1-penalized and the ℓ_1-constrained problems coincide [6, Proposition 5.2.1], and so the behavior of ℓ_1 penalty and the ℓ_1 constraint is identical.

Penalized discriminant analysis solves a nonconvex optimization problem. Hence, there is no guarantee that the ℓ_1 penalty behaves the same way as the ℓ_1 constraint. In particular, we show that very sparse solutions are not attainable using an ℓ_1 penalty. For example, a solution with 10 or fewer variables may not be achievable for any value of the tuning parameter. We derive a theoretical lower bound on the number of selected variables that supports our empirical results. While this research is motivated by penalized discriminant analysis, the results apply to any generalized eigenvalue problem with an ℓ_1 penalty.

To our knowledge, this is the first work in model selection that recognizes and quantifies the difference between the ℓ_1 penalty and the ℓ_1 constraint in nonconvex settings. Moreover, we show that the poor variable selection performance of ℓ_1-penalized nonconvex criterions can be remedied by considering their ℓ_1-constrained versions. The constrained version of discriminant analysis, unlike penalized version in [69], can select an arbitrarily small number of variables.

2.2 Penalized generalized eigenvalue problem

Consider the generalized eigenvalue problem

$$v = \arg\max_{v \in \mathbb{R}^p} \left\{ v^\top Q v \right\} \quad \text{subject to} \quad v^\top C v \leq 1, \tag{2.1}$$

where $Q \in \mathbb{R}^{p \times p}$ (for quadratic function) and $C \in \mathbb{R}^{p \times p}$ (for constraint) are two symmetric, semi positive-definite matrices. In addition, let C be strictly positive-definite. A common approach to enforce sparsity in the solution vector v is to restrict the ℓ_1 norm of v by modifying (2.1), either by penalizing the objective function or by adding the ℓ_1 norm constraint. We define the ℓ_1-penalized

problem (2.1) as

$$v_\lambda = \arg\max_{v\in\mathbb{R}^p} \left\{ v^\top Q v - \lambda \|v\|_1 \right\} \text{ subject to } v^\top C v \leq 1. \tag{2.2}$$

Here $\lambda \geq 0$ is the tuning parameter and the ℓ_1 norm is part of the objective function.

For clarity of exposition, we only consider the case $C = I$. Problem (2.2) simplifies to

$$v_\lambda = \arg\max_{v\in\mathbb{R}^p} \left\{ v^\top Q v - \lambda \|v\|_1 \right\} \text{ subject to } v^\top v \leq 1. \tag{2.3}$$

Following [69], (2.3) can be recast as a biconvex optimization problem

$$\text{maximize}_{u,v} \left\{ 2u^\top Q^{1/2} v - \lambda \sum_{j=1}^{p} |v_j| - u^\top u \right\} \text{subject to } v^\top v \leq 1, \tag{2.4}$$

since maximizing with respect to u gives $u = Q^{1/2}v$. The problem (5.6) is convex with respect to u when v is fixed and is convex with respect to v when u is fixed. This property allows the use of Alternate Convex Search (ACS) to find the solution [26, Section 4.2.1]. ACS ensures that all accumulation points are partial optima and have the same function value [26, Theorem 4.9].

Starting with an initial value $v^{(0)}$ the algorithm proceeds by iterating the following two steps:

Step 1 $u^{(k)} = \arg\max_u \left\{ 2u^\top Q^{1/2} v^{(k)} - u^\top u \right\} = Q^{1/2} v^{(k)}$

Step 2 $v^{(k+1)} = \arg\max_v \left\{ 2(u^{(k)})^\top Q^{1/2} v - \lambda \sum_{j=1}^{p} |v_j| \right\}$ subject to $v^\top v \leq 1$.

Following [69, Proposition 2], it is useful to reformulate Step 2 as

$$q^{(k+1)} = \arg\max_q \left\{ 2(u^{(k)})^\top Q^{1/2} q - \lambda \sum_{j=1}^{p} |q_j| - q^\top q \right\} \tag{2.5}$$

Algorithm 1 Optimization algorithm for ℓ_1-penalized problem with $C = I$.

Given: $\lambda > 0$, Q, $k = 1$

$v^{(0)} \leftarrow$ dominant eigenvector of Q

$v^{(0)} \leftarrow v^{(0)}/\sqrt{(v^{(0)})^\top v^{(0)}}$

repeat

  for $l \in \{1, ..., p\}$ **do**

    $v_l^{(k)} \leftarrow \text{sign}\left((Qv^{(k-1)})_l\right)\left(|(Qv^{(k-1)})_l| - \lambda/2\right)_+$

  end for

  if $\{v^{(k)} \neq 0\}$ **then**

    $v^{(k)} \leftarrow v^{(k)}/\sqrt{(v^{(k)})^\top v^{(k)}}$

  end if

  $k \leftarrow k + 1$

until $k = k_{\max}$ or $v^{(k)}$ satisfies stopping criterion.

where, if $q^{(k+1)} = 0$, then $v^{(k+1)} = 0$, else $v^{(k+1)} = q^{(k+1)}/\sqrt{(q^{(k+1)})^\top q^{(k+1)}}$. Since problem (2.5) is convex with respect to q, the solution $q^{(k+1)}$ satisfies KKT conditions [11]

$$2Q^{1/2}u^{(k)} - 2q^{(k+1)} - \lambda\Gamma = 0, \tag{2.6}$$

where Γ is a p-vector and each Γ_j is a subgradient of $|q_j^{(k+1)}|$, i.e. $\Gamma_j = 1$ if $q_j^{(k+1)} > 0$, $\Gamma_j = -1$ if $q_j^{(k+1)} < 0$ and Γ_j is between -1 and 1 if $q_j^{(k+1)} = 0$. From (3.12)

$$q_j^{(k+1)} = \text{sign}((Q^{1/2}u^{(k)})_j)\left(|(Q^{1/2}u^{(k)})_j| - \frac{\lambda}{2}\right)_+. \tag{2.7}$$

Algorithm 1 results from combining Steps 1 and 2 with the update (2.7).

We use Algorithm 1 to find the local solution to problem (2.3). While the convergence to the global solution v_λ is not guaranteed due to nonconvexity, the observed empirical behavior of local solution is consistent with the theoretical results of Section 2.4.

2.3 Empirical evidence for restriction on solution sparsity

First, we consider the following synthetic scenarios:

1. $p \in \{500, 2000\}$, $\mathrm{rank}(Q) = 1$ with eigenvalue $\gamma = 1$ and the dominant eigenvector l with components l_i coming from the uniform distribution on $[0, 1]$, standardized as $l^\top l = 1$.
2. $p \in \{500, 2000\}$, $\mathrm{rank}(Q) = 50$, where Q is the sample covariance matrix of 50 observations x_i with $x_{ij} \sim N(0, 1)$ for $j = 1, ..., p$.

Figure 2.1 illustrates that the number of selected features decreases when the value of λ increases. What is surprising, however, is the sudden drop to zero which is observed in all cases and is most severe when p is 2000 and $\mathrm{rank}(Q) = 1$. Based on Figure 2.1, it is impossible to select fewer than 1000 features in this scenario. It appears there exists a λ_0 such that for all $\lambda < \lambda_0$ the solution has at least M non-zero components and for all $\lambda \geq \lambda_0$ the solution is exactly zero.

Next, we consider colon cancer dataset [4] and 14 cancer dataset [47]. We have chosen these two datasets as they are publicly available and have been extensively studied in the literature [69, 76]. Colon cancer dataset is available from `http://genomics-pubs.princeton.edu/oncology/affydata/index.html`. It contains the expression of 2000 genes from 40 tumor tissues and 22 normal tissues. 14 cancer dataset is available from `http://statweb.stanford.edu/~tibs/ElemStatLearn/`. It contains 16063 gene expression measurements collected on 198 samples from 14 cancer classes. For the analysis, we select 144 samples that are designated as the training set. Following the recommendation of [29, p. 654], we standardize the data to have

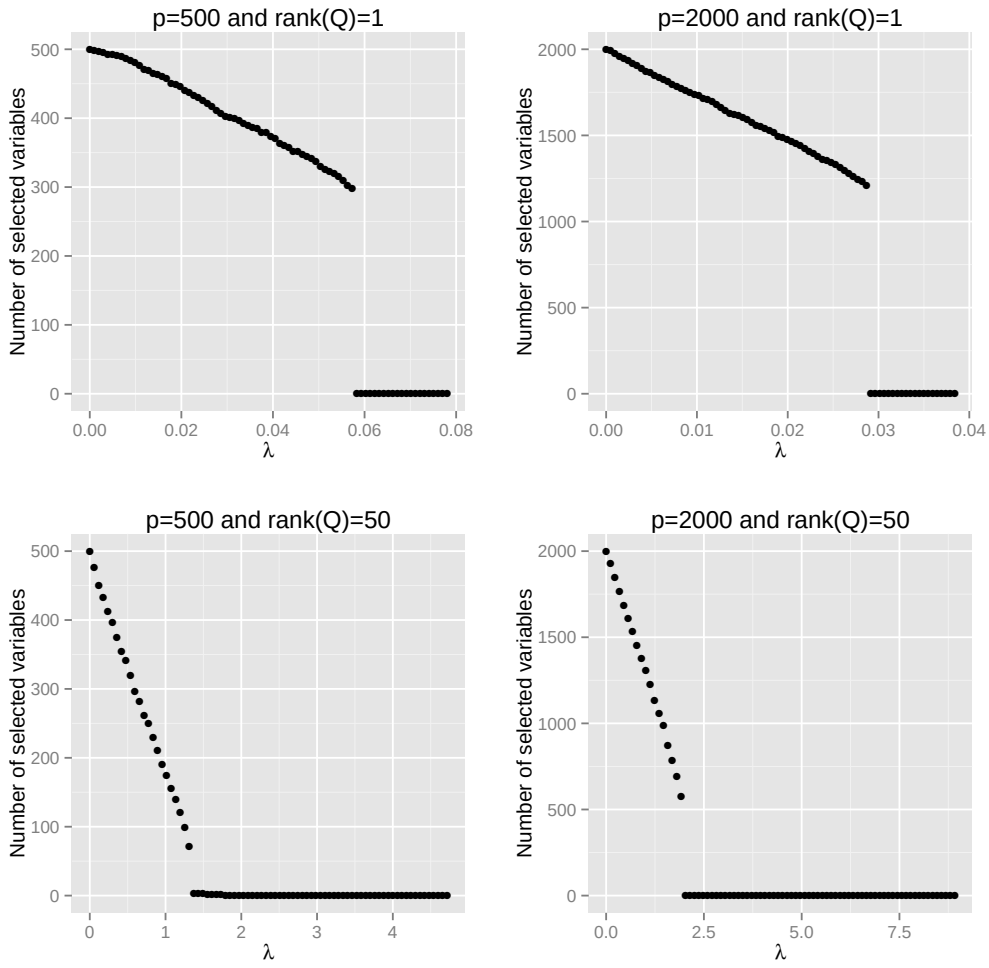


Figure 2.1: Number of non-zero features obtained empirically versus the tuning parameter λ.

mean zero and standard deviation one for each patient. To reduce computational costs, we only consider 2684 features that have standard deviation above 0.45.

We apply penalized linear discriminant analysis (LDA) and penalized principal components analysis (PCA) to both datasets, more details are provided in Section 2.6.1. Thus, there are four scenarios:

1. Penalized LDA on colon cancer dataset, $p = 2000$, $\mathrm{rank}(Q) = 1$.

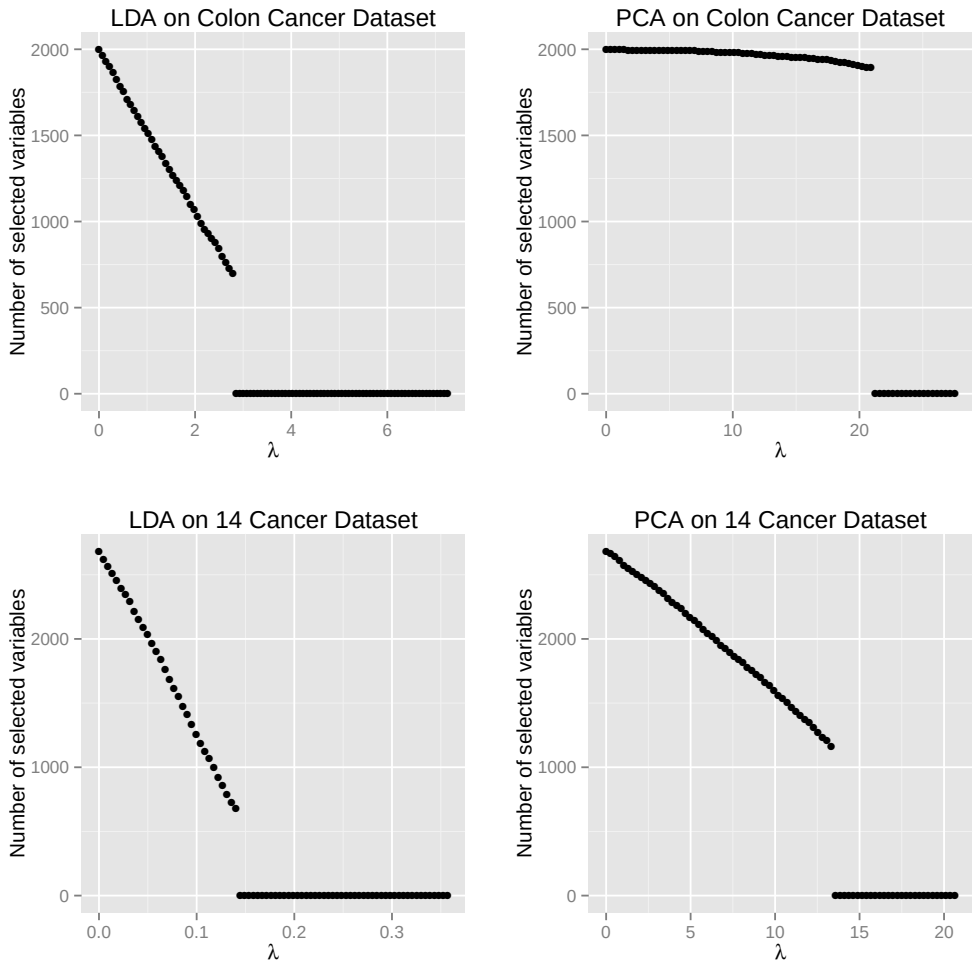


Figure 2.2: Number of nonzero features versus the tuning parameter λ.

2. Penalized PCA on colon cancer dataset, $p = 2000$, $\mathrm{rank}(Q) = 61$.

3. Penalized LDA on 14 cancer dataset, $p = 2684$, $\mathrm{rank}(Q) = 13$.

4. Penalized PCA on 14 cancer dataset, $p = 2684$, $\mathrm{rank}(Q) = 141$.

Figure 2.2 shows the number of nonzero features versus the tuning parameter for colon cancer and 14 cancer datasets. As in the case with synthetic data, there is a sudden drop to zero as the tuning parameter increases. This behavior appears to be especially problematic with penalized PCA, making it impossible to select less than 1000 features for either dataset.

These results demonstrate that the ℓ_1-penalized method is not effective as a variable selection tool. The large number of nonzero features makes it impossible to interpret the results or further validate the features in a lab setting. In Section 2.4 we demonstrate that this behavior is not an artifact of the chosen optimization algorithm, but rather is intrinsic to penalized generalized eigenvalue problem (2.2).

2.4 Lower bound on the number of non-zero components

We start by deriving an upper bound on the objective function in (2.2) for v such that $\|v\|_0 \leq k$.

Proposition 1. *Let q_i be the ith row of Q and let $q_i(j)$ be the subvector of q_i of length j with the maximal ℓ_2 norm. Then*

$$\max_{\|v\|_0 \leq k, v^\top C v \leq 1} \left\{ v^\top Q v - \lambda \|v\|_1 \right\} \leq \frac{\|\tilde{q}(k)\|_2}{\sigma_{\min}(C)},$$

where $\tilde{q}_i = \max\left(\|q_i(k)\|_2 - \lambda \sqrt{\sigma_{\min}(C)}, 0 \right)$.

Proof. When $\|v\|_0 \leq k$ and $v^\top C v \leq 1$,

$$\begin{aligned} v^\top Q v - \lambda \|v\|_1 &= \sum_{i=1}^{p} |v_i| \left(s_i q_i^\top v - \lambda \right) \leq \sum_{i=1}^{p} |v_i| (\|q_i(k)\|_2 \|v\|_2 - \lambda) \\ &\leq \sum_{i=1}^{p} |v_i| \left(\frac{\|q_i(k)\|_2}{\sqrt{\sigma_{\min}(C)}} - \lambda \right) \leq \frac{1}{\sqrt{\sigma_{\min}(C)}} \sum_{i=1}^{p} |v_i| \tilde{q}_i \\ &\leq \frac{1}{\sqrt{\sigma_{\min}(C)}} \|\tilde{q}_{(k)}\|_2 \|v\|_2 \leq \frac{1}{\sigma_{\min}(C)} \|\tilde{q}_{(k)}\|_2. \end{aligned}$$

□

The upper bound grows with the value of k. In particular, if $\|\tilde{q}(l)\|_2 = 0$, then $\|\tilde{q}(m)\|_2 = 0$ for all $m < l$. As such, we can derive the lower bound on the number of nonzero components in v_λ.

Corollary 1. *Let q_i be the ith row of Q and let $q_i(j)$ be the subvector of q_i of length j with the maximal ℓ_2 norm. Let $m_\lambda = j_{\min} \in \{1, ..., p\}$ such that $\max_i \|q_i(j)\|_2 > \lambda\sqrt{\sigma_{\min}(C)}$. Then*

$$v_\lambda = 0 \quad \textit{or} \quad \|v_\lambda\|_0 \geq m_\lambda.$$

We can further use this result to derive a value of $\lambda_{\max}$.

Corollary 2. *Let q_i be the ith row of Q and let $\lambda_{\max} = \max_i \|C^{-1/2} q_i\|_2$. Then for all $\lambda \geq \lambda_{\max}$, $v_\lambda = 0$.*

Since the bound of Proposition 1 applies to any v in the feasible region with $\|v\|_0 \leq k$, the results of Proposition 1 and Corollaries 1 and 2 also apply to the solution of Algorithm 1.

Figure 2.3 demonstrates the bound from Corollary 1 as a function of λ using the examples of Section 3.4. As the tuning parameter λ increases, so does the value of m_λ (dashed line on Figure 2.3). On the other hand, as the tuning parameter λ increases, so does the weight of the penalty in the objective function of (2.2) leading to the smaller number of nonzero features (dots on Figure 2.3). A perfect prediction of the minimal number of nonzero features requires the dashed line to take the same value as dotted line at the value of $\lambda = \lambda_0$ where the drop happens. For example, when $p = 500$ and $\mathrm{rank}(Q) = 1$, $\lambda_0 \approx 0.06$, $m_{\lambda_0} \approx 150$, whereas the drop happens at ≈ 300 features. This discrepancy between the actual minimal number of nonzero features and the value of m_{λ_0} is not surprising, since the value of m_λ is based on an upper bound of the objective

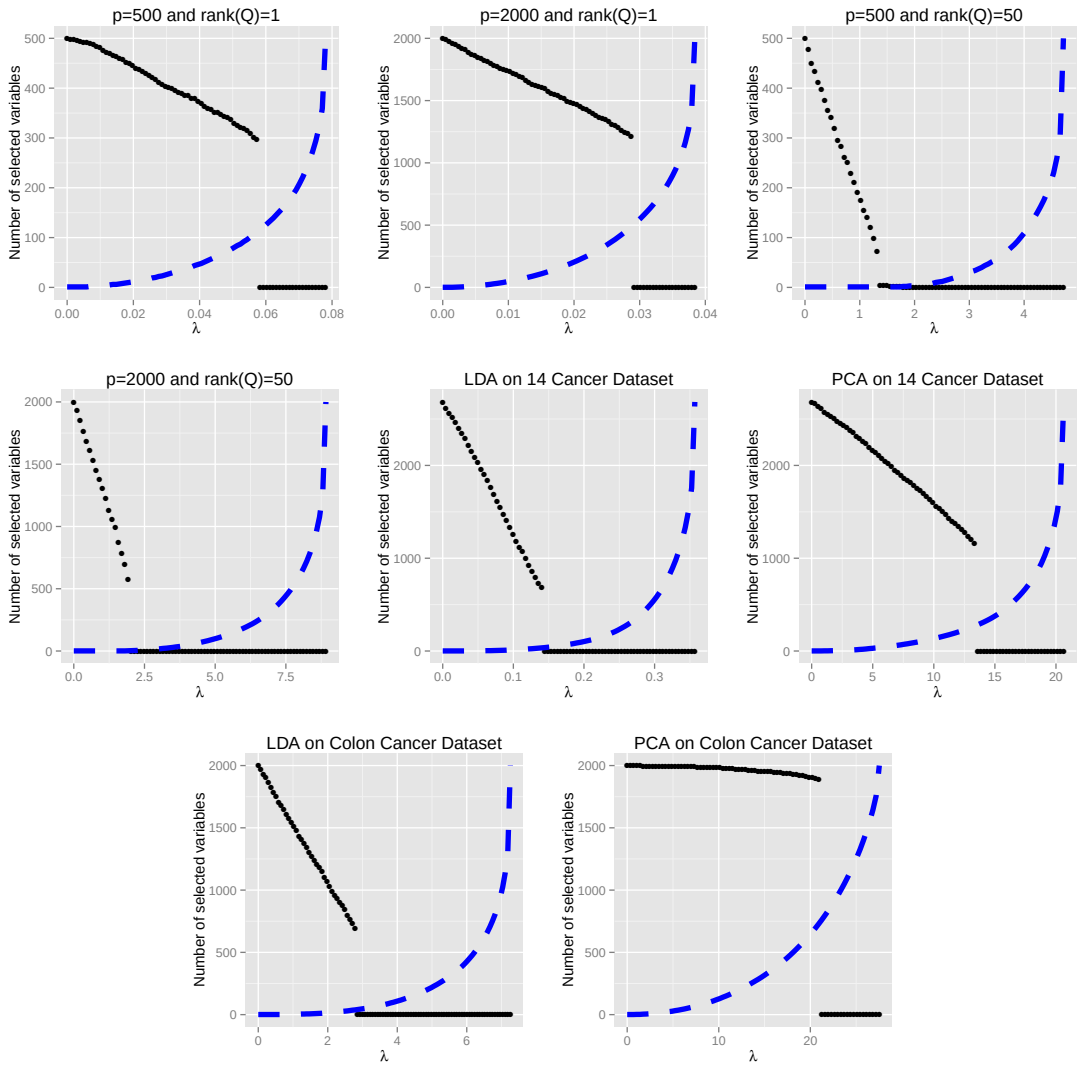


Figure 2.3: Number of non-zero features obtained empirically versus the tuning parameter λ, the dashed line shows the value of m_λ from Corollary 1.

function in (2.2). While this bound appears to be somewhat crude for the case when $\mathrm{rank}(Q) = 50$, it predicts at least 150 features for $p = 500$, $\mathrm{rank}(Q) = 1$ case and at least 500 features for $p = 2000$, $\mathrm{rank}(Q) = 1$ case. Similarly, it predicts at least 34 features for penalized LDA on 14 cancer dataset, at least 267 features for penalized PCA on 14 cancer dataset, at least 38 features for penalized LDA on colon cancer dataset and at least 728 features for penalized PCA on colon cancer dataset.

2.5 ℓ_1 penalty versus ℓ_1 constraint

In Sections 2.3 and 2.4 we demonstrated both empirically and theoretically that ℓ_1-penalized generalized eigenvalue problem can fail to obtain very sparse solutions. Consider ℓ_1-constrained generalized eigenvalue problem

$$v_\tau = \arg\max_{v \in \mathbb{R}^p} \left\{ v^\top Q v \right\} \quad \text{subject to} \quad v^\top C v \leq 1, \|v\|_1 \leq \tau. \tag{2.8}$$

Here $\tau \geq 0$ is a tuning parameter which constrains the ℓ_1 norm.

It is natural to ask whether the solutions to problems (2.2) and (2.8) are the same, and whether the same restriction on solution sparsity applies to (2.8). A partial answer to this question is given in Proposition 3.

Proposition 2. *For every $\lambda \geq 0$ there exists $\tau \geq 0$ such that $v_\lambda = v_\tau$.*

Proof of Proposition 3. Fix any $\lambda \geq 0$ and let v_λ be the solution to (2.2). It follows that for any v such that $v^T v \leq 1$,

$$v_\lambda^T F v_\lambda - \lambda \|v_\lambda\|_1 \geq v^T F v - \lambda \|v\|_1. \tag{2.9}$$

Consider (2.8) with $t = \|v_\lambda\|_1$. From (2.9) for each v such that $v^T v \leq 1$ and $\|v\|_1 \leq t$

$$v_\lambda^T F v_\lambda \geq v^T F v + \lambda(\|v_\lambda\|_1 - \|v\|_1) = v^T F v + \lambda(t - \|v\|_1) \geq v^T F v.$$

This means v_λ is the solution to (2.8), hence $v_t = v_\lambda$. □

The reverse is true for convex problems such as LASSO [6, Proposition 5.2.1], however the generalized eigenvalue problem is nonconvex. Following [6, Chapter 5] and [11, Chapter 5.3], we use a geometry-based approach to visualize the

relationship between the solutions to the ℓ_1-constrained and the ℓ_1-penalized optimization problems in the following example.

Example: Let $p = 2$, $C = I$ and $\mathrm{rank}(Q) = 1$. Let $\gamma = 1$ be the positive eigenvalue of Q and l be the corresponding eigenvector, so that $Q = \gamma l l^\top$. We consider two scenarios:

1. $x = (0.2, 0.8)^\top$, $l = x/\|x\|_2$;
2. $x = (0.5, 0.6)^\top$, $l = x/\|x\|_2$.

The corresponding ℓ_1-constrained optimization problem (2.8) becomes

$$v_\tau = \arg\min_{v \in \mathbb{R}^p} -(v^\top l)^2 \text{ subject to } v^\top v \leq 1, \|v\|_1 \leq \tau. \tag{2.10}$$

This problem defines the set S of constrained pairs

$$S = \left\{ (h, f) \,\middle|\, h = \|v\|_1, f = -(v^\top l)^2 \text{ for all } v \in \mathbb{R}^p, v^\top v \leq 1 \right\}. \tag{2.11}$$

Using the set S, (2.10) can be viewed as a minimal common point problem: finding a point (h', f') with a minimal f-coordinate among the points common to set S and halfspace $h \leq \tau$,

$$\left\{ (h', f') \in S \,\middle|\, f' = \min_{(h,f) \in S, h \leq \tau} f \right\}. \tag{2.12}$$

By definition of v_τ, $f' = -(v_\tau^\top l)^2$ and $h' = \|v_\tau\|_1$. We construct the corresponding sets S for both scenarios in Figure 2.4 and identify the minimal common point using $\tau = 1.1$.

Consider the corresponding ℓ_1-penalized optimization problem (2.2):

$$v_\lambda = \arg\min_{v \in \mathbb{R}^p} -(v^\top l)^2 + \lambda \|v\|_1 \text{ subject to } v^\top v \leq 1. \tag{2.13}$$

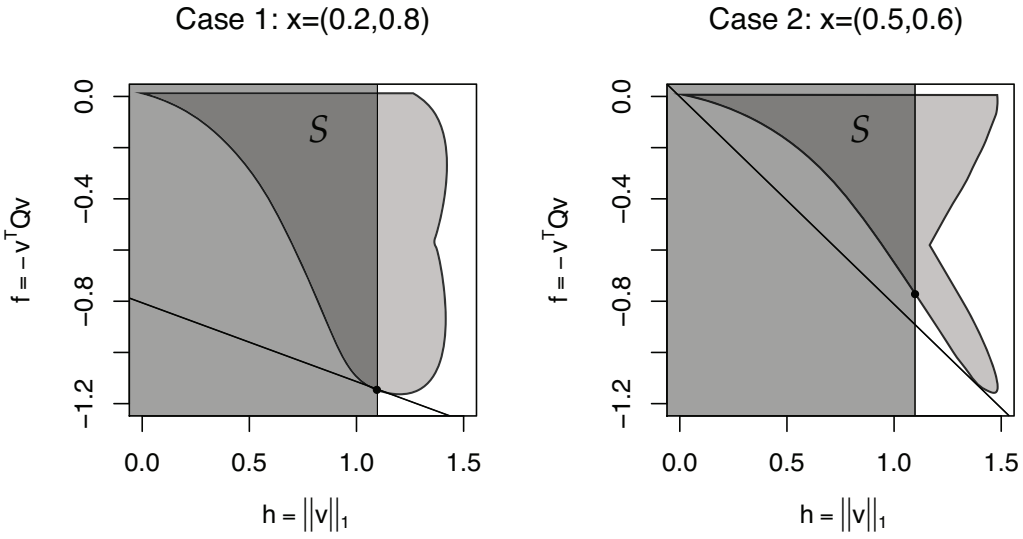


Figure 2.4: Visualization of the set S, the minimum common point of S and $h \leq 1.1$ and the supporting hyperplane for the set S. The eigenvector of matrix Q is equal to $l = x/\|x\|_2$.

Using the set S in (2.11), we can view (2.13) as finding the point $(h'', f'') \in S$ such that

$$(h'', f'') = \arg \min_{(h,f)\in S} \{f + \lambda h\}.$$

By definition of v_λ, $h'' = \|v_\lambda\|_1$ and $f'' = -(v_\lambda^\top l)^2$.

The solutions to (2.10) and (2.13) are the same if $(h', f') = (h'', f'')$. This occurs when $f = -\lambda h$ is the supporting hyperplane to the set S at the point (h', f'). Figure 2.4 shows whether such a hyperplane can be constructed in both scenarios. In the first scenario the hyperplane can be constructed for each $\tau \geq 1$, and in particular for $\tau = 1.1$. In the second scenario, the hyperplane cannot be constructed for $\tau = 1.1$ as it has to lie below the point $(0, 0)$ and the minimal point of S corresponding to $h = 1.4$. Moreover, this is true not only for $\tau = 1.1$ but for all values of τ between 1 and 1.4. Hence, for these τ there exists no λ such that $v_\lambda = v_\tau$.

Consider the shape of the set S in the second scenario. For all $\tau < 1.4$, $(h', f') = (0, 0)$ is the only point at which it is possible to construct the supporting hyperplane to the set S. This implies $h'' = \|v_\lambda\|_1 = 0$, hence $v_\lambda = 0$ is the corresponding solution to the dual problem (2.13) for all $\tau < 1.4$. In contrast, $v_\tau = 0$ only for $\tau = 0$. Therefore there exists no $\lambda \geq 0$ such that $\|v_\lambda\|_1 = \tau$ for $\tau \in (0, 1.4)$, leading to a constraint on the sparsity level of the solution v_λ.

In the language of optimization theory, the Lagrangian dual problem defines the supporting hyperplane to S in (2.11), and hence the optimal (primal) solution is greater than the dual solution (weak duality). If the supporting hyperplane intersects S at a single point, as in scenario one above, the optimization problem is said to have the zero duality gap (strong duality) property. If the objective function is convex, as in the LASSO, strong duality is guaranteed by Slaters constraint [11, Chapter 5]. Unlike the LASSO, (2.1) is not a convex problem and therefore this guarantee no longer applies.

Our example demonstrates the existence of a duality gap between problems (2.2) and (2.8); there exist values of $\tau > 0$ such that the solution v_τ cannot be obtained by solving (2.2). Moreover, these unattainable values of τ correspond to sparse solutions, v_τ with very few non-zero components. Therefore, there is a restriction on the sparsity of solutions obtained by solving the ℓ_1-penalized problem (2.2), but there is no restriction on the sparsity of the solutions obtained by solving the corresponding ℓ_1-constrained problem (2.8).

2.6 Variable selection with ℓ_1 constraint

2.6.1 Fisher's Linear Discriminant Analysis

Let (X_i, Y_i), $i = 1, .., n$, be independent pairs with $X_i \in \mathbb{R}^p$ and $Y_i \in \{1, .., G\}$, where G is the number of classes. Let W and B be the within-group sample covariance matrix and the between-group sample covariance matrix respectively. Further, assume that X is scaled so that $\text{diag}(W) = I$. [69] find the first penalized discriminant vector as

$$v_\lambda = \arg\max_{v \in \mathbb{R}^p} \left\{ v^\top B v - \lambda \sum_{j=1}^{p} |v_j| \right\} \text{ subject to } v^\top v \leq 1. \tag{2.14}$$

We use (2.14) for the analysis of the colon cancer dataset [4] and 14 cancer dataset [47]. Figure 2.2 shows the number of non-zero features selected by Algorithm 1 versus the tuning parameter λ. Empirically it is impossible to select less than 600 features for colon cancer dataset and less than 500 features for 14 cancer dataset.

Now consider the constrained version of (2.14):

$$v_\tau = \arg\max_{v \in \mathbb{R}^p} \left\{ v^\top B v \right\} \text{ subject to } v^\top v \leq 1, \quad \|v\|_1 \leq \tau. \tag{2.15}$$

The local solution to Problem (2.15) can be found using Algorithm 1 with the following modification: for each iteration k choose $\lambda^{(k)}$ such that $\|v^{(k)}_{\lambda^{(k)}}\|_1 = \tau$. Usually, such a $\lambda^{(k)}$ is found by performing a binary search on the grid $[0, \lambda_{\max}]$. Figure 2.5 shows the number of non-zero features in v_τ versus the tuning parameter τ. As τ increases, so does the number of features. Moreover, it is possible to select an arbitrarily small number of variables.

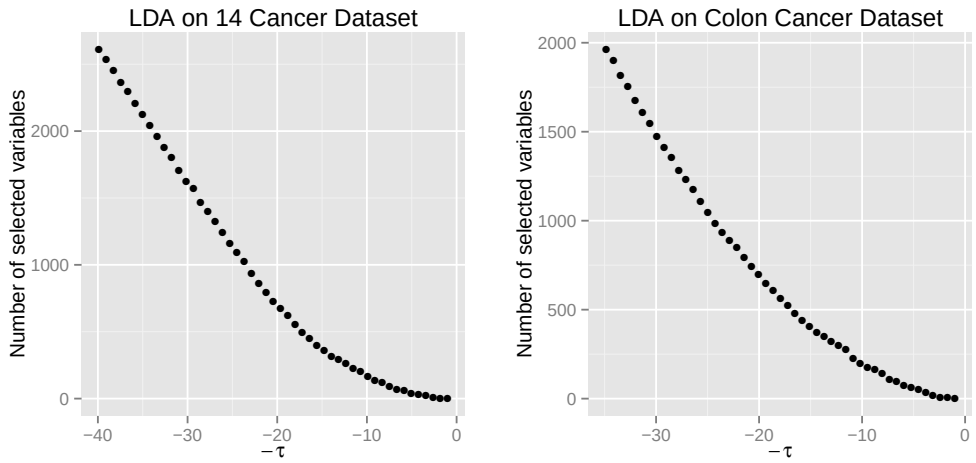


Figure 2.5: Number of nonzero features versus the tuning parameter for the ℓ_1-constrained LDA.

2.6.2 Principal Component Analysis

Let X_i, $i = 1, ..., n$, be independent samples with $X_i \in \mathbb{R}^p$. Let S be the sample covariance matrix and assume that X is scaled so that $S = \frac{1}{n}X^\top X$. The first penalized principal component v is defined as

$$v_\lambda = \arg\max_{v \in \mathbb{R}^p} \left\{ \frac{1}{n} v^\top X^\top X v - \lambda \sum_{j=1}^{p} |v_j| \right\} \quad \text{subject to} \quad v^\top v \leq 1. \tag{2.16}$$

We use (2.16) for the analysis of the colon cancer dataset [4] and 14 cancer dataset [47]. Figure 2.2 shows the number of non-zero features selected by Algorithm 1 versus the tuning parameter λ. Empirically it is impossible to select less than 1700 features for colon cancer dataset and less than 1000 features for 14 cancer dataset.

Now consider the constrained version of (2.16):

$$v_\tau = \arg\max_{v \in \mathbb{R}^p} \left\{ \frac{1}{n} v^\top X^\top X v \right\} \text{ subject to } v^\top v \leq 1, \quad \|v\|_1 \leq \tau. \tag{2.17}$$

The local solution to Problem (2.17) can be found using Algorithm 1 in the same

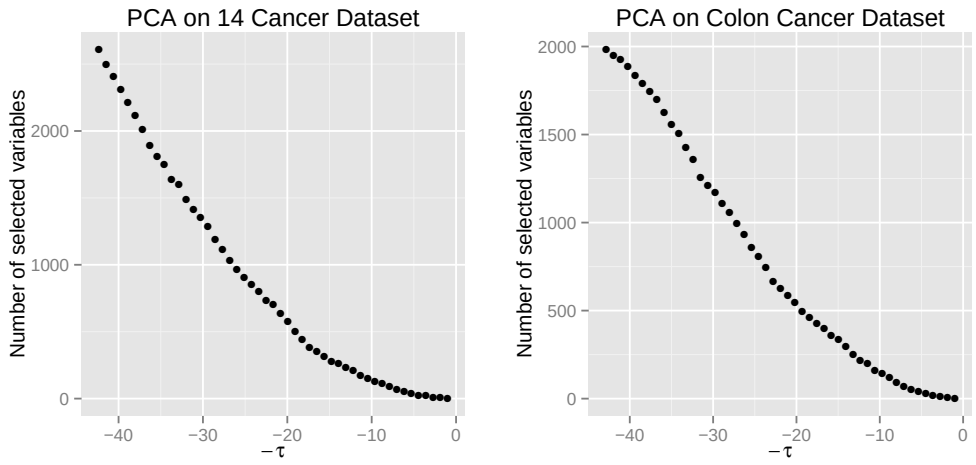


Figure 2.6: Number of nonzero features versus the tuning parameter for the ℓ_1-constrained PCA.

way as the local solution to (2.15). Figure 2.6 shows the number of non-zero features in v_τ versus the tuning parameter τ. As τ increases, so does the number of features. Moreover, it is possible to select an arbitrarily small number of variables.

2.7 Discussion

We conjecture that the restriction on the solution sparsity is an intrinsic property of any ℓ_1-penalized criterion with a nonconvex objective function due to the likely non-zero duality gap. Other examples of such criterions include [3] and [9]. The restriction on the solution sparsity directly affects the variable selection properties of corresponding estimators. In future research we are planning to generalize our results to these nonconvex criterions.

CHAPTER 3

SIMULTANEOUS SPARSE ESTIMATION OF CANONICAL VECTORS WHEN $P >> N$

3.1 Introduction

The classical use of Linear Discriminant Analysis (LDA) when $p \gg N$ fails to provide useful results because of the singularity of covariance matrix and over-selection of relevant features [21, 7]. As a result, the extension of LDA to high-dimensional settings has recently received a lot of attention. A number of these proposals result in non-sparse classifiers. [24], [33] and [72] regularize the within-class covariance matrix in order to obtain a positive definite estimate. Other approaches that lead to sparse discriminant vectors have also been considered. [57] propose the shrunken centroids methodology by adapting the naive Bayes classifier and soft-thresholding the mean vectors. [28] combine the shrunken centroids approach with a ridge-type penalty on the within-class covariance matrix. [69] apply an ℓ_1 penalty to the Fisher's discriminant problem in order to obtain sparse discriminant vectors. [16] use an optimal scoring approach which essentially reduces the sparse discriminant vector construction to a penalized regression problem.

In the two-group setting, [12] and [37] propose direct estimation of the canonical vector thus avoiding separate estimation of the covariance matrix. Simulations and real data applications show that the direct estimation approach results in reduced misclassification rates in comparison to alternative methods. The corresponding optimization problems can be solved efficiently for large data sets and have desirable theoretical properties. Unfortunately, the extension

of two-group methods to the multi-group case is nebulous [29, p. 658]. Popular approaches include "one-versus-all" and "one-versus-one" methods, where the final classification assignment is usually based on the "majority vote". As such, computation of more vectors is required (G and $G(G-1)/2$ versus $G-1$).

[69] and [16] propose estimating canonical vectors in a sequential fashion in the multi-group setting: starting with the first canonical vector v_1, with subsequent v_i found subject to orthogonality constraints. This approach is undesirable from a computational viewpoint, as well as from an estimation perspective. Each subsequent canonical vector v_i relies on all the previous estimates v_k for $k < i$, hence propagation of the estimation error is possible. In addition, the corresponding optimization problems are nonconvex, hence the convergence of the optimization algorithms to the global solution is not assured. This computational burden poses additional theoretical challenges in the analysis and, as a result, these methods do not have any theoretical guarantees.

The objective of this chapter is to bridge the computational and theoretical gap in the literature between the two-group and multi-group methods. Inspired by the superior performance of the direct estimation methods by [12] and [37] in the two-group case, our goal is to introduce and develop a novel methodology that has the same guaranteed performance in the multi-group setting. Our proposal is based on the observation that canonical vectors can be expressed in a closed form up to an orthogonal transformation. Moreover, this transformation affects neither the classification rule nor the sparsity pattern. The definitive contribution of this work is the development of novel methodology that provides:

1. simultaneous estimation of all $G-1$ canonical vectors without prior estimation of the covariance structure;

2. simultaneous variable selection from all canonical vectors;
3. theoretical guarantees on variable selection and classification consistency in multi-group settings.

To our knowledge, this is the first method for multi-group sparse discriminant analysis that achieves all of these goals.

In addition, while the motivation for our approach is quite different from [37], we show that the two methods are equivalent in the two-group setting. We use this connection to extend the theoretical results of [37] to the multi-group setting and show that the proposed method can consistently identify the true support of canonical vectors.

The proposed optimization problem for our methodology is convex and therefore can be solved efficiently for large data sets. Our algorithm doesn't require additional regularization of the sample covariance matrix or generation of an initial starting point, as the form of the optimization problem suggests a natural choice for the grid of tuning parameters. This is not the case for the methods of [12], [69] or [16], where the appropriate grid must be carefully chosen by the user. While these advantages are mostly computational, they simplify the implementation of the method leading to more consistent results across users. More implementation details are provided in Section 3.4.3.

The rest of this chapter is organized as follows. Section 3.2 discusses the canonical vectors estimation problem and provides new insights into the form of the canonical vectors in the standard $N \gg p$ setting. We use these insights to propose a direct estimation procedure and describe computational aspects of the algorithm. Section 3.3 develops bounds on the estimation error and proves

that the proposed method identifies the true support of canonical vectors with a high probability. Section 3.4 provides simulation results while Section 3.5 describes applications to real datasets. We conclude with a discussion of future research in Section 3.6.

3.2 Methodology

3.2.1 Estimation problem

We assume that $X_i \in \mathbb{R}^p$, $i = 1, ..., N$, are independent and come from G groups with different means and the same covariance matrix, i.e. $\mathbb{E}(X_i|Y_i = g) = \mu_g$ and $\mathrm{Cov}(X_i|Y_i = g) = \Sigma_W$, where $Y_i \in \{1, ..., G\}$. The between-group population covariance matrix Σ_B is defined as

$$\Sigma_B = \sum_{g=1}^{G} \pi_g (\mu_g - \mu)(\mu_g - \mu)^\top,$$

where $\pi_g = P(Y_i = g)$ are group-specific probabilities and $\mu = \sum_{i=1}^{G} \pi_g \mu_g$ is the overall population mean. The *population canonical vectors* Ψ are defined as eigenvectors corresponding to non-zero eigenvalues of $\Sigma_W^{-1}\Sigma_B$. Although the eigenvectors are unique only up to normalization [25], we take advantage of the uniqueness of the eigenspace in defining a scale-invariant classification rule. For a new observation value of X, $x \in \mathbb{R}^p$, the population classification rule $h_\Psi(x)$ is defined as

$$h_\Psi(x) = \arg\min_{1 \le g \le G} (x - \mu_g)^\top \Psi (\Psi^\top \Sigma_W \Psi)^{-1} \Psi^\top (x - \mu_g) - 2 \log \pi_g. \qquad (3.1)$$

The classification rule is based on the closest Mahalanobis distance in the projected space defined by Ψ, after adjustment for potential discrepancy in the prior

group probabilities π_g. Through the addition of $2\log\pi_g$ term, the resulting classification rule mimics the optimal classification rule under the assumption that the data comes from the multivariate normal group-conditional distribution [39, Chapter 3.9.3]. Our goal is to identify the eigenspace spanned by Ψ based on the sample observations $X_i \in \mathbb{R}^p$ and sample labels $Y_i \in \{1, ..., G\}$.

Consider the within-group sample covariance matrix $W = \frac{1}{N-G}\sum_{g=1}^{G}(n_g - 1)S_g$ and the between-group sample covariance matrix $B = \frac{1}{N}\sum_{g=1}^{G} n_g(\bar{X}_g - \bar{X})(\bar{X}_g - \bar{X})^\top$, where n_g is the number of observations in group g, S_g is the sample covariance matrix for group g, $\bar{X}_g$ is the sample mean for group g and $\bar{X}$ is the overall sample mean. Recall that W is nonsingular when $N \gg p$ and therefore we can define the *sample canonical vectors* V as $G-1$ eigenvectors corresponding to non-zero eigenvalues of $W^{-1}B$ [38, Chapter 11.5]. Similarly to (3.1), the sample classification rule $\hat{h}_V(x)$ is defined as

$$\hat{h}_V(x) = \arg\min_{1\le g\le G}(x - \bar{X}_g)^\top V(V^\top WV)^{-1}V^\top(x - \bar{X}_g) - 2\log\frac{n_g}{N}. \tag{3.2}$$

Further we show that canonical vectors can be expressed in a closed form up to an orthogonal transformation. For this purpose, we establish the connection between the eigenspaces of matrices $\Sigma_W^{-1}\Sigma_B$ and $(\Sigma_W + \Sigma_B)^{-1}\Sigma_B$, as well as derive a closed form low-rank decomposition of Σ_B and B.

Proposition 1. *Let Υ_ρ be the matrix of eigenvectors corresponding to non-zero eigenvalues of $(\Sigma_W + \rho\Sigma_B)^{-1}\rho\Sigma_B$ for some positive ρ. Then Υ_ρ is also the matrix of eigenvectors corresponding to non-zero eigenvalues of $\Sigma_W^{-1}\Sigma_B$.*

Proof. From the definition of Υ_ρ, $(\Sigma_W + P\Sigma_B)^{-1}\rho\Sigma_B\Upsilon_\rho = \Upsilon_\rho\Lambda$. It follows that

$$\rho\Sigma_B\Upsilon_\rho = \Sigma_W\Upsilon_\rho\Lambda + \rho\Sigma_B\Upsilon_\rho\Lambda;$$

$$\rho\Sigma_B\Upsilon_\rho(I-\Lambda) = \Sigma_W\Upsilon_\rho\Lambda;$$

$$\Sigma_W^{-1}\Sigma_B\Upsilon_\rho = \Upsilon_\rho\frac{1}{\rho}\Lambda(I-\Lambda)^{-1}.$$

From the last equation it follows that Υ_ρ is the matrix of eigenvectors of $\Sigma_W^{-1}\Sigma_B$.

□

Proposition 2. *The following decompositions hold:* $\Sigma_B = \Delta\Delta^\top$ *and* $B = DD^\top$*, where for* $r = 1, ..., G-1$ *the* r*th column of* Δ *has the form*

$$\Delta_r = \frac{\sqrt{\pi_{r+1}}\left(\sum_{i=1}^r \pi_i(\mu_i - \mu_{r+1})\right)}{\sqrt{\sum_{i=1}^r \pi_i \sum_{i=1}^{r+1}\pi_i}} \tag{3.3}$$

and the r*th column of* D *has the form*

$$D_r = \frac{\sqrt{n_{r+1}}\left(\sum_{i=1}^r n_i(\bar{X}_i - \bar{X}_{r+1})\right)}{\sqrt{N}\sqrt{\sum_{i=1}^r n_i \sum_{i=1}^{r+1} n_i}}. \tag{3.4}$$

Proof. The proof is only given for matrix B, the proof for matrix Σ_B is similar.

1. Consider the equal group case: $n_1 = ... = n_G = n$ and $N = Gn$. It follows that $\bar{X} = \sum_{i=1}^G \bar{X}_G/G$ and therefore $B = \frac{1}{N}\sum_{g=1}^G n(\bar{X}_g - \bar{X})(\bar{X}_g - \bar{X})^\top = \frac{1}{N}X^\top\{\frac{1}{\sqrt{n}}1_g\}C\{\frac{1}{\sqrt{n}}1_g\}^\top X$, where C is the centering matrix and $\{\frac{1}{\sqrt{n}}1_g\}$ is a $N \times G$ matrix formed by G columns $\frac{1}{\sqrt{n}}1_g$ such that $(1_g)_j = 1$ if jth observation belongs to the gth group and $(1_g)_j = 0$ otherwise. Note that $C = H^\top H$ where H is the Helmert matrix of size G with its first row removed [50]. Therefore $B = DD^\top$, where $D = \frac{1}{\sqrt{N}}X^\top\{\frac{1}{\sqrt{n}}1_g\}H^\top$.

2. Consider the general case where each group has size n_g. Similar to the equal group case, $B = \frac{1}{N}X^\top\{\frac{1}{\sqrt{n_i}}1_i\}\tilde{C}\{\frac{1}{\sqrt{n_i}}1_i\}^\top X$, where $\tilde{C} = I_G - \frac{1}{\sqrt{N}}KK^\top$ and $K = (\sqrt{n_1}...\sqrt{n_G})^\top$. Next we show that similar to C, $\tilde{C}$ can be decomposed as

$\tilde{C} = \tilde{H}^\top \tilde{H}$ and $\tilde{H}$ is a $G-1 \times G$ adjusted Helmert matrix. Since $\tilde{H}$ satisfies $I_G - \frac{1}{\sqrt{N}} K K^\top = \tilde{H}^\top \tilde{H}$, $(K, \tilde{H}^\top)$ is an orthogonal matrix. The $G-1$ orthogonal contrasts for unbalanced data have the following form [50, p 51]:

$$\Delta_r = \sqrt{n_{r+1}} \left(\sum_{h=1}^{r} n_h (\bar{X}_h - \bar{X}_{r+1}) \right).$$

Denote by h_r the rows of $\tilde{H}$. Then it follows that for some constant C_r, $h_r \left\{ \frac{1}{\sqrt{n_i}} 1_i \right\}^\top X = C_r \Delta_r$. This means that $h_{rj} = C_r \sqrt{n_{r+1} n_j}$ for $j = 1, .., r$; $h_{r(r+1)} = -C_r \sum_{i=1}^{r} n_i$ and $h_{rj} = 0$ for $j > (r+1)$. To find C_r, we use the fact that $h_r h_r^\top = 1$. Let $s_r = \sum_{i=1}^{r} n_i$. Then C_r satisfies $C_r^2 \left(\sum_{j=1}^{r} n_{r+1} n_j + s_r^2 \right) = 1$, or equivalently $C_r^2 s_{r+1} s_r = 1$. From the last equation $C_r = \frac{1}{\sqrt{s_{r+1} s_r}}$. Combining the results it follows that $B = D D^\top$, where $D = \frac{1}{\sqrt{N}} X^\top \left\{ \frac{1}{\sqrt{n_g}} 1_i \right\} \tilde{H}^\top$ and $D_r = \frac{1}{\sqrt{N}} C_r \Delta_r = \frac{\sqrt{n_{r+1}} \left(\sum_{h=1}^{r} n_h (\bar{X}_h - \bar{X}_{r+1}) \right)}{\sqrt{N s_{r+1} s_r}}$. □

The low-rank decomposition of matrices Σ_B and B is not unique. Our choice of Δ and D in Proposition 2 is motivated by the fact that these matrices can be expressed in a closed form (unlike the eigenvectors of Σ_B and B) and have intuitive interpretation in terms of the differences between the group means. Specifically, the columns of Δ and D define orthogonal contrasts between the means of G groups. In the case $G = 2$, $\Delta = \sqrt{\pi_1 \pi_2} (\mu_2 - \mu_1)$ and $D = \frac{\sqrt{n_1 n_2}}{N} (\bar{X}_2 - \bar{X}_1)$.

We use Propositions 1 and 2 to derive the explicit form of the matrix of eigenvectors of $\Sigma_W^{-1} \Sigma_B$.

Proposition 3. *Define Δ as in* (3.3)*. There exists matrices $P_1, P_2 \in \mathbb{O}^{G-1}$ such that $\Sigma_W^{-1} \Delta P_1$ and $(\Sigma_W + \Sigma_B)^{-1} \Delta P_2$ are matrices of eigenvectors of $\Sigma_W^{-1} \Sigma_B$ corresponding to non-zero eigenvalues.*

Proof. Denote $\Psi = \Sigma_W^{-1}\Delta P$, where P is an orthogonal matrix such that $\Delta^\top \Sigma_W^{-1}\Delta = P\Lambda P^\top$. It follows that $\Sigma_W^{-1}\Sigma_B\Psi = \Sigma_W^{-1}\Delta\Delta^\top\Sigma_W^{-1}\Delta P = \Sigma_W^{-1}\Delta P\Lambda = \Psi\Lambda$. Hence, Ψ is the matrix of eigenvectors of $\Sigma_W^{-1}\Sigma_B$. The proof for $(\Sigma_W + \Sigma_B)^{-1}\Delta P$ is analogous using the results of Proposition 1. □

Both $\Sigma_W^{-1}\Delta P_1$ and $(\Sigma_W + \Sigma_B)^{-1}\Delta P_2$ satisfy the definition of the matrix of eigenvectors of $\Sigma_W^{-1}\Sigma_B$ since the eigenvectors are only unique up to normalization. We discuss the advantages of using $(\Sigma_W + \Sigma_B)^{-1}\Delta P_2$ over $\Sigma_W^{-1}\Delta P_1$ in Section 3.2.2.

Finally, we show that the orthogonal transformation has no effect on the classification rule.

Proposition 4. *For any matrix $R \in \mathbb{O}^{G-1}$, the classification rule based on V is the same as the classification rule based on VR: $h_V(x) = h_{VR}(x)$ and $\hat{h}_V(x) = \hat{h}_{VR}(x)$ for all $x \in \mathbb{R}^p$.*

Proof. The proof is only given for the sample classification rule $\hat{h}_V(x)$. Define $Z = XV$. Using V, a new observation $x \in \mathbb{R}^p$ is classified to group $\hat{h}_V(x)$, where

$$\hat{h}_V(x) = \arg\min_{1\le j\le G}(V^\top x - \bar{Z}_j)^\top (V^\top W V)^{-1}(V^\top x - \bar{Z}_j) - 2\log\frac{n_j}{N}.$$

Consider a new classification rule $\hat{h}_{V'}(x)$ based on $V' = VR$ with $R \in \mathbb{O}^{G-1}$. Then $Z' = XV' = XVR = ZR$ and

$$\begin{aligned}
\hat{h}_{V'}(x) &= \arg\min_{1\le j\le G}(V'^\top x - \bar{Z}'_j)^\top (V'^\top W V')^{-1}(V'^\top x - \bar{Z}'_j) - 2\log\frac{n_j}{N} \\
&= \arg\min_{1\le j\le G}(R^\top V^\top x - R^\top \bar{Z}_j)^\top (R^\top V^\top W V R)^{-1}(R^\top V^\top x - R^\top \bar{Z}_j) - 2\log\frac{n_j}{N} \\
&= \arg\min_{1\le j\le G}(V^\top x - \bar{Z}_j)^\top R R^{-1}(V^\top W V)^{-1}(R^\top)^{-1}R^\top(V^\top x - \bar{Z}_j) - 2\log\frac{n_j}{N} \\
&= \hat{h}_V(x).
\end{aligned}$$

□

3.2.2 Proposed estimation criterion

From Section 3.2.1 it follows that for classification and variable selection it is sufficient to estimate the matrix of eigenvectors of $\Sigma_W^{-1}\Sigma_B$ up to orthogonal transformation. Proposition 3 gives two possible population objectives: $\tilde{\Psi} = \Sigma_W^{-1}\Delta$ and $\Psi' = (\Sigma_W + \Sigma_B)^{-1}\Delta$.

First, we consider $\tilde{\Psi} = \Sigma_W^{-1}\Delta$ with the goal of choosing a suitable loss function to capture the deviations of the estimator from the target.

By definition $\tilde{\Psi}$ satisfies

$$\tilde{\Psi} = \arg\min_{V \in \mathbb{R}^{p\times(G-1)}} \frac{1}{2}\|\Sigma_W^{1/2}V - \Sigma_W^{-1/2}\Delta\|_F^2 = \arg\min_{V \in \mathbb{R}^{p\times(G-1)}} \frac{1}{2}\operatorname{Tr}\left(V^\top \Sigma_W V - 2\Delta^\top V\right). \tag{3.5}$$

In the two-group case, V is a vector and the objective function in (3.5) reduces to

$$\begin{aligned}
&\frac{1}{2}(\Sigma_W^{1/2}V - \Sigma_W^{-1/2}\Delta)^\top(\Sigma_W^{1/2}V - \Sigma_W^{-1/2}\Delta) \\
&\quad = \frac{1}{2}(V - \Sigma_W^{-1}\Delta)^\top \Sigma_W (V - \Sigma_W^{-1}\Delta) \\
&\quad = \frac{1}{2}(V - \tilde{\Psi})^\top \Sigma_W (V - \tilde{\Psi}).
\end{aligned}$$

This objective function is the same as the quadratic loss function considered by [49], who observed that it is invariant with respect to linear transformation of the data. Hence, we can define an estimator $\tilde{V}$ by substituting Σ_W and Δ with W and D:

$$\tilde{V} = \arg\min_{V \in \mathbb{R}^{p\times(G-1)}} \frac{1}{2}\operatorname{Tr}\left(V^\top W V - 2D^\top V\right). \tag{3.6}$$

Unfortunately, the objective function in (3.6) is unbounded when W is singular due to the existence of non-zero $\bar{V}$ with $\operatorname{Tr}(\bar{V}^\top W \bar{V}) = 0$ and $\operatorname{Tr}(D^\top \bar{V}) > 0$. A simple solution is to use $\tilde{W} = W + \rho I$ instead of W, which is a common

regularization in the LDA context [24, 28, 12], and leads to

$$\tilde{V}(\rho) = \arg \min_{V \in \mathbb{R}^{p \times G-1}} \frac{1}{2} \operatorname{Tr}(V^\top W V) + \frac{\rho}{2} \|V - D\|_F^2.$$

The second component of the objective function encourages $\hat{V}(\rho)$ to be close to D, especially when ρ is large. In contrast, $\hat{V}(\rho)$ should be close to $W^{-1}D$ according to Proposition 3. This discrepancy suggests that strong regularization of W may have a negative affect on classification performance.

Consider now the second population objective from Proposition 3, $\Psi' = (\Sigma_W + \Sigma_B)^{-1}\Delta$. Following the same arguments as with $\tilde{\Psi} = \Sigma_W^{-1}\Delta$ leads to

$$\tilde{V} = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \operatorname{Tr}\left(V^\top W V\right) + \frac{1}{2} \|D^\top V - I\|_F^2. \tag{3.7}$$

The objective function in (3.7) is bounded from below even when W is singular, hence no additional regularization of W is needed. For this reason, we choose $(\Sigma_W + \Sigma_B)^{-1}\Delta$ as the population quantity for our estimation procedure.

Our next goal is to perform a variable selection, which in the discriminant analysis corresponds to having zeros in the matrix of canonical vectors. An ℓ_1 penalty $\|V\|_1 = \sum_{i=1}^{p} \sum_{j=1}^{G-1} |v_{ij}|$ is commonly used for variable selection as it induces the element-wise sparsity. We are not using this penalty for two reasons. First, the element-wise sparsity leads to variable selection within each canonical vector, however the number of variables used by all vectors may be very large. Secondly, the element-wise sparsity is not preserved under the orthogonal transformation, leading to different sparsity patterns in $(\Sigma_W + \Sigma_B)^{-1}\Delta$ and $(\Sigma_W + \Sigma_B)^{-1}\Delta P_2$.

Instead of an ℓ_1 penalty, we consider the row-wise ℓ_2 penalty $\sum_{i=1}^{p} \|v_i\|_2$ which induces row-sparsity. Unlike the element-wise sparsity, the row-sparsity

eliminates the variables from all canonical vectors and is preserved under the orthogonal transformation. Alternative penalties include group SCAD and group MCP, we refer the reader to [31] for an overview. Our choice of $\sum_{i=1}^{p} \|v_i\|_2$ is motivated by the fact that it preserves convexity of the underlying optimization problem. Combining (3.7) with this penalty suggests an estimator $\hat{V}(\lambda)$, defined as

$$\hat{V}(\lambda) = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \operatorname{Tr}\left(V^\top W V\right) + \frac{1}{2}\|D^\top V - I\|_F^2 + \lambda \sum_{i=1}^{p} \|v_i\|_2, \tag{3.8}$$

where the objective function is convex and bounded below by zero. When W is nonsingular and $\lambda = 0$, $\hat{V}(\lambda) = (W+B)^{-1}D$, which according to Proposition 3 is the matrix of sample canonical vectors up to an orthogonal rotation. The three components of the objective function in (3.8) attempt to minimize the within-group variability, control the level of the between-group variability and provide regularization by inducing sparsity respectively.

3.2.3 Connection with other sparse discriminant analysis methods when $G = 2$

The motivation for our method is based on the eigenstructure of the discriminant analysis problem in the multi-group setting, however it has a direct connection with the two-group methods previously proposed in the literature. When $G = 2$, V is a vector in $\mathbb{R}^p$ and (3.8) takes the form

$$\hat{V}(\lambda) = \arg \min_{V \in \mathbb{R}^p} \frac{1}{2} V^\top W V + \frac{1}{2}(D^\top V - 1)^2 + \lambda \|V\|_1.$$

Proposition 5. *Consider $\hat{V}_{DSDA}(\lambda)$ [37], defined as*

$$\hat{V}_{DSDA}(\lambda) = \arg \min_{\beta_0 \in \mathbb{R}, V \in \mathbb{R}^p} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - X_i^\top V)^2 + \lambda \|V\|_1,$$

where $y_i = -\frac{N}{n_1}$ *if the* i*th subject is in group 1 and* $y_i = \frac{N}{n_2}$ *otherwise. Then*

$$\hat{V}(\lambda) = \frac{N}{\sqrt{n_1 n_2}} \hat{V}_{DSDA}\left(\frac{N}{\sqrt{n_1 n_2}}\lambda\right).$$

Proof. By definition $D = \frac{\sqrt{n_1 n_2}}{N}(\bar{X}_1 - \bar{X}_2)$. Therefore

$$\begin{aligned}\hat{V}_{DSDA}(\lambda) &= \arg\min_{V\in\mathbb{R}^p} \frac{1}{2} V^\top (W + DD^\top)V - \frac{N}{\sqrt{n_1 n_2}} D^\top V + \lambda \|V\|_1 \\ &= \arg\min_{V\in\mathbb{R}^p} \frac{1}{2}\frac{\sqrt{n_1 n_2}}{N} V^\top (W + DD^\top)V - D^\top V + \frac{\lambda\sqrt{n_1 n_2}}{N}\|V\|_1.\end{aligned}$$

Define

$$f_{DSDA}(V,\lambda) = \frac{1}{2}\frac{\sqrt{n_1 n_2}}{N} V^\top (W + DD^\top)V - D^\top V + \frac{\lambda\sqrt{n_1 n_2}}{N}\|V\|_1.$$

Similarly, $\hat{V}(\lambda) = \arg\min_{V\in\mathbb{R}^p} f(V,\lambda)$, where

$$f(V,\lambda) = \frac{1}{2} V^\top (W + DD^\top)V - D^\top V + \lambda\|V\|_1.$$

Note that

$$\begin{aligned} f\left(\frac{\sqrt{n_1 n_2}}{N}V, \lambda\right) &= \frac{\sqrt{n_1 n_2}}{N}\left(\frac{1}{2}\frac{\sqrt{n_1 n_2}}{N} V^\top (W + DD^\top)V - D^\top V + \lambda\|V\|_1\right) \\ &= \frac{\sqrt{n_1 n_2}}{N} f_{DSDA}\left(V, \frac{N}{\sqrt{n_1 n_2}}\lambda\right).\end{aligned}$$

It follows that $\hat{V}(\lambda) = \frac{N}{\sqrt{n_1 n_2}}\hat{V}_{DSDA}\left(\frac{N}{\sqrt{n_1 n_2}}\lambda\right)$. □

Furthermore, [36] show an equivalence between the three methods for sparse discriminant analysis in the two-group setting: [71], [16] and [37]. It follows that our method belongs to the same class, however it can be applied to any number of groups. Thus, it can be viewed as a multi-group generalization of this class of methods.

The optimization problem in (3.8) corresponds to the choice of $\rho = 1$ in Proposition 1. In general, any $\rho > 0$ leads to

$$\hat{V}(\lambda,\rho) = \arg\min_{V\in\mathbb{R}^{p\times G-1}} \frac{1}{2}\operatorname{Tr}(V^\top W V) + \frac{\rho}{2}\|D^\top V - I\|_F^2 + \lambda\sum_{i=1}^{p}\|v_i\|_2. \qquad (3.9)$$

When $\rho \to \infty$, (3.9) is equivalent to

$$\hat{V}(\lambda, \rho = \infty) = \arg \min_{D^\top V = I} \frac{1}{2} \operatorname{Tr}(V^\top W V) + \lambda \sum_{i=1}^{p} \|v_i\|_2, \tag{3.10}$$

hence the optimization problem (3.9) can be considered a convex relaxation to (3.10) for large values of ρ. When $G = 2$, the optimization problem (3.9) is equivalent to the proposal of [23], who also observe the connection between (3.9) and (3.10). They perform a simulation study to assess the effect of the tuning parameter ρ and note that its value doesn't significantly affect the classification results as long as the best λ is chosen for each ρ. They keep the value of ρ at a fixed level $\rho = 10$.

3.2.4 Optimization algorithm

The optimization problem in (3.8) is convex with respect to V and therefore can be solved efficiently using a block-coordinate descent algorithm. The convexity of the problem guarantees the convergence to the global optimum from any initial starting value. We refer the reader to [5] for the overview of convex optimization with sparsity-inducing norms, including alternative algorithms like proximal gradient and interior-point methods. Alternative algorithms include proximal gradient methods and interior-point methods, we refer the reader to [5] for the overview of convex optimization with sparsity-inducing norms. We chose to use the block-coordinate descent algorithm as it takes advantage of warm starts when solving for a range of tuning parameters and is one of the fastest algorithms for smooth losses with separable regularizers [5, 46]. Define the usual sample covariance matrix as

$$T = W + B = W + DD^\top. \tag{3.11}$$

By convexity, the solution to (3.8) satisfies the KKT conditions [11, Chapter 5.5]. Differentiating (3.8) with respect to the $(G-1)\times 1$ vector v_j formed by the jth row of V leads to

$$V^\top T_j - d_j + \lambda u_j = 0, \tag{3.12}$$

where T_j is the jth column of matrix T in (3.11), d_j is a $(G-1)\times 1$ vector formed by the jth row of matrix D in (3.4) and u_j is the subgradient of $\|v_j\|_2$:

$$u_j = \begin{cases} \frac{v_j}{\|v_j\|_2}, & \text{if } v_j \neq 0; \\ \in \{u : \|u\|_2 \leq 1\}, & \text{if } v_j = 0. \end{cases}$$

Solving (3.12) with respect to v_j leads to $v_j = \left(d_j - \sum_{i\neq j} t_{ij} v_i - \lambda u_j\right)/t_{jj}$, where t_{ij} are the elements of matrix T. This leads to the block-coordinate descent algorithm.

Algorithm 2 Block-coordinate descent algorithm.

Given: $k = 1$, $V^{(0)}$, ε

repeat

$\bar{V} \leftarrow V^{(k-1)}$

for $j = 1$ **to** p **do**

$v_j^{(k)} \leftarrow \left(1 - \frac{\lambda}{\|d_j - \sum_{i\neq j} t_{ij}\bar{v}_i\|_2}\right)_+ \left(d_j - \sum_{i\neq j} t_{ij}\bar{v}_i\right)/t_{jj}$

end for

$k \leftarrow k + 1$

until $k = k_{\max}$ or $V^{(k)}$ satisfies $\max_{i,j} |v_{ij}^{(k)} - v_{ij}^{(k-1)}| < \varepsilon$

More implementation details, including the choice of $V^{(0)}$ and tuning parameter λ, are described in Section 3.4.3.

If T is non-singular, by applying the vectorization operator (3.8) can be rewritten as

$$\hat{V}(\lambda) = \arg\min_{V\in\mathbb{R}^{p\times G-1}} \frac{1}{2}\left\|\mathrm{vec}(D^\top T^{-1/2}) - (T^{1/2}\otimes I_{G-1})\,\mathrm{vec}(V^\top)\right\|_2^2 + \lambda\sum_{i=1}^{p}\|v_i\|_2.$$

This formulation corresponds to a group lasso optimization problem [73] with the response vector $\mathrm{vec}(D^\top T^{-1/2})$ and the design matrix $T^{1/2} \otimes I_{G-1}$. Due to the form of the design matrix, each block subproblem can be solved in a closed form, making the implementation of block-coordinate descent algorithm straightforward.

3.3 Theoretical guarantees

In this section we analyze the variable selection and classification performance of the estimator $\hat{V}(\lambda)$ defined in (3.8). In Section 3.2.3 we established an equivalence between our proposal and the proposal of [37] for the two-group case. We use this connection to extend the variable selection consistency results of [37] to the multi-group case. In particular, to prove Theorem 1, we derive Lemmas 2 and 3 that serve as a multi-group version of Lemma A1 in [37]. We also show that the variable selection consistency implies classification consistency.

Let $\Sigma = \Sigma_W + \Sigma_B$, $\Psi' = \Sigma^{-1}\Delta$ and denote the support of Ψ' by $A = \{j : \|\tilde{\psi}_j\|_2 \neq 0\}$, where ψ'_j is the jth row of Ψ'. Denote the support of $\hat{V}(\lambda)$ by $\hat{A} = \{j : \|\hat{v}_j(\lambda)\|_2 \neq 0\}$. Furthermore, let $s = \mathrm{card}(A)$, $\kappa = \|\Sigma_{A^cA}\Sigma_{AA}^{-1}\|_\infty$, $\phi = \|\Sigma_{AA}^{-1}\|_\infty$, $\Psi_{\min} = \min_{i\in A}\|\psi'_i\|_2$ and $\Delta = \|\Delta\|_{\infty,2}$, where Σ_{AA} is the sub-matrix of Σ formed by the intersection of the rows and columns in A. In Theorem 1 we establish lower bounds on $P(A = \hat{A})$ and $P\left(\|\hat{V}(\lambda)_A - \Psi'_A\|_{\infty,2} \leq 2\phi\lambda\right)$.

Theorem 1. *Assume $\kappa < 1$ and $(X_i|Y_i = g) \sim N(\mu_g, \Sigma_W)$, $i = 1, .., N$. Then*

1. *For any $\lambda > 0$ and positive $\epsilon \leq \frac{\lambda(1-\kappa)}{(\kappa+1)(\phi\Delta+1)+2\phi\lambda}$, $\hat{V}(\lambda)_{A^C} = 0$ with a probability*

of at least $1 - t_1$*, where*

$$t_1 = c_1 ps \exp(-c_2 N s^{-2} \epsilon^2) + 2(G-1)p \exp(-c_3 N \epsilon^2).$$

2. *For any* $\lambda < \frac{\Psi_{\min}}{\phi}$ *and* $\epsilon < \frac{\Psi_{\min} - \lambda\phi}{\phi(1+\phi\Delta+\Psi_{\min})}$ *none of the elements of* $\hat{V}(\lambda)_A$ *are zero with a probability of at least* $1 - t_2$*, where*

$$t_2 = c_1 s^2 \exp(-c_2 N s^{-2} \epsilon^2) + 2(G-1)s \exp(-c_3 N \epsilon^2).$$

3. *For any positive* $\epsilon < \frac{\lambda}{1+\phi\Delta+2\phi\lambda}$

$$P\left(\|\hat{V}(\lambda)_A - \Psi'_A\|_{\infty,2} \leq 2\phi\lambda\right) \geq 1 - c_1 s^2 \exp(-c_2 N s^{-2} \epsilon^2) + 2(G-1)s \exp(-c_3 N \epsilon^2).$$

While the motivation for the proposed optimization problem doesn't rely on the normality assumption, the normality assumption does simplify the proof. We discuss possible extensions to the non-normal case in the online supplement in Section S8. We further use Theorem 1 to establish variable selection consistency of the estimator $\hat{V}(\lambda)$ defined in (3.8). Specifically, Theorem 1 implies the asymptotic conditions under which $P(A = \hat{A}) \to 1$, which coincide with asymptotic conditions for the two-group case [37]:

(C1) $N \to \infty$, $p \to \infty$, $G = O(1)$ and $\frac{\log(ps)s^2}{N} \to 0$.

(C2) $\sqrt{\frac{\log(ps)s^2}{N}} << \lambda_N << \Psi_{\min}$.

Corollary 3. *If (C1) and (C2) hold, then* $P(A = \hat{A}) \to 1$.

We also show that under the same asymptotic conditions the sample classification rule based on $\hat{V}$ coincides with the population classification rule h_Ψ defined in (3.1). Let $X \in \mathbb{R}^p$ be a new observation with a value $x \in \mathbb{R}^p$.

Corollary 4. *If (C1) and (C2) hold, then* $P(\|\hat{V}(\lambda)_A - \Psi'_A\|_{\infty,2} \leq 2\phi\lambda_N) \to 1$*. Moreover, if* $\lambda_N \to 0$*, then* $P\left(\hat{h}_{\hat{V}}(x) = h_\Psi(x)\right) \to 1$.

3.4 Simulation Results

In this section we evaluate the performance of the estimator $\hat{V}(\lambda)$ defined in (3.8) against the alternative methods proposed in the literature. We refer to our proposal as MGSDA for Multi-Group Sparse Discriminant Analysis. The test datasets are the same size as the training datasets and are generated independently. The following structures for Σ_W are considered in all the simulations:

1. **Identity:** $\Sigma_W = I$.
2. **Autoregressive:** $\Sigma_W = (\Sigma_{ij})_{p\times p}$ with $\Sigma_{ij} = 0.8^{|i-j|}$ for $1 \leq i, j \leq p$.
3. **Data Based:** $\Sigma_W = (1-\alpha)S + \alpha I$, where $\alpha = 0.01$ and S is a sample correlation matrix estimated from the most variable $p = 800$ features of Ramaswamy dataset [47]. The dataset is available from `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.

Structures 1-2 have been used in simulation studies in LDA literature [12, 69, 37]. We view the Data Based structure as an approximation to a covariance structure that is more realistic in practical settings.

3.4.1 The two-group case

This simulation scenario considers the classification between the two groups with $\mu_1 = 0_p$ and $\mu_2 = (1_{10}, 0_{p-10})$ for covariance structures 1-2. For the Data Based structure, we take $\mu_2 = (d_{10}, 0_{p-10})$ with d ranging from 0.1 to 0.5 since in this case the Bayes error is almost zero for $\mu_2 = (1_{10}, 0_{p-10})$. The sample size

for each group is $n = 100$. Conditional on group g, the samples are drawn independently from the multivariate normal distribution $N(\mu_g, \Sigma_W)$.

[37] perform extensive simulations to compare their proposal with the methods of [71], [69], [58] and [22]. In all the settings, the method of [37] performs the best in terms of misclassification error. Given Proposition 5, we do not compare MGSDA with any of these methods. On the other hand, [12] also show that their proposal performs the best when compared to [52], [22] and [58]. To our knowledge, no comparison was performed between the methods of [37] and [12], therefore in this section we compare our results to the results of [12]. We follow the terminology of [12] and refer to their method as Linear Programming Discriminant (LPD). We also evaluate the performance of $\tilde{\Psi} = \Sigma_W^{-1}\Delta$. We refer to $\tilde{\Psi}$ as the Oracle.

We note that the LPD requires additional regularization of the within-group sample covariance matrix: $\tilde{W} = W + \rho I$. This regularization is needed to generate a feasible starting point for the optimization algorithm. [12] suggest taking $\rho \leq \sqrt{\log p/N}$. In our simulations $N = 200$ and therefore $\rho = 0.15$ satisfies this requirement for both $p = 100$ and $p = 800$. We also try $\rho = 2$ to examine how the choice of ρ affects the misclassification rate.

The misclassification error rates are reported in Table 3.8 and the corresponding number of selected features in Table 3.9. The methods have similar error rates, with MGSDA performing better on the Autoregressive covariance structure. They also select comparable numbers of features in all scenarios. Note that the matrix of population canonical vector Ψ is truly sparse only in the Identity case, it is only approximately sparse in other scenarios.

Table 3.1: Mean misclassification error rates as percentages over 100 replications, $G = 2$, standard deviation is given in parentheses.

Covariance	p	MGSDA	LPD, $\rho = 0.15$	LPD, $\rho = 2$	Oracle
Identity	100	6.65(2.07)	6.75(2.04)	6.17(1.94)	5.58(1.89)
	800	7.32(2.09)	6.84(1.97)	6.44(1.73)	5.75(1.56)
Autoregressive	100	19.02(2.91)	20.83(3.17)	23.88(2.99)	16.65(2.48)
	800	22.29(3.26)	23.59(3.35)	24.5(3.06)	16.05(2.59)

Table 3.2: Mean number of selected features over 100 replications, $G = 2$, standard deviation is given in parentheses.

Covariance	p	MGSDA	LPD, $\rho = 0.15$	LPD, $\rho = 2$
Identity	100	20(7)	19(14)	19(15)
	800	29(16)	25(26)	24(29)
Autoregressive	100	19(6)	20(14)	30(21)
	800	32(15)	23(21)	27(31)

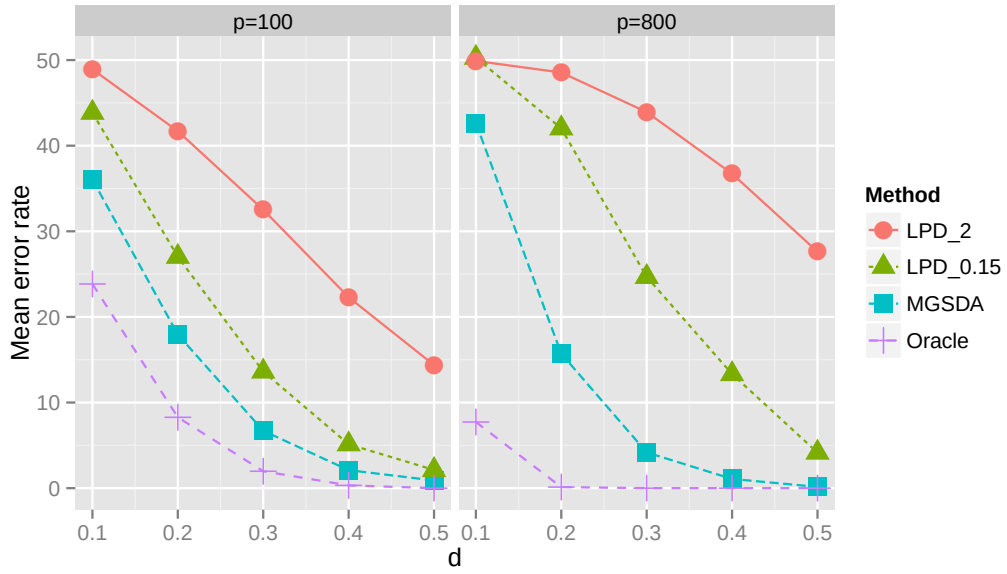
The mean misclassification rates for the Data Based covariance structure are reported in Figure 3.1. In this case MGSDA performs significantly better than LPD regardless of the choice of ρ. The difference in misclassification rates is especially noticeable when the difference in means d is small.

The error rates of LPD with $\rho = 0.15$ and $\rho = 2$ are comparable except for the Data Based structure. This suggests that the choice of ρ can significantly affect the performance of the LPD, with smaller values of ρ likely to result in smaller misclassification error. Unfortunately it remains unclear how to choose the optimal ρ in practical settings.

3.4.2 The multi-group case

This simulation scenario considers the classification between the five groups with $\mu_1 = 0_p$, $\mu_2 = (d_5, -d_5, 0_{p-10})$, $\mu_3 = (-d_5, d_5, 0_{p-10})$, $\mu_4 = (d_{10}, 0_{p-10})$ and $\mu_5 = (d, -d, d, -d, ..., 0_{p-10})$ with $d = 1.5$ for Identity and Autoregressive covariance structures, and $d = 0.3$ for Data Based covariance structure. The sample

Figure 3.1: Mean misclassification error rate in percentage over 25 replications for Data Based covariance structure as a function of difference in means d, $G = 2$.



size for each group is $n = 50$. We consider both multivariate normal and multivariate t distribution with 5 degrees of freedom.

The LPD method of [12] is developed for the two-group setting. Though it can be generalized to the multi-group case, this generalization is not unique. Among the popular methods are "one versus one" and "one versus all" approaches [29, p. 658]. In addition to requiring the computation of a larger number of discriminant vectors ($G(G-1)/2$ and G correspondingly), these approaches can disagree in their classification rules as well as in selected features. Given this ambiguity, we do not compare our method to the LPD in the multi-group case.

We were able to find only two methods in the literature that specifically consider sparse discriminant analysis in the multi-group case: penalizedLDA by [69] and sparseLDA by [16]. Therefore, we compare the performance of MGSDA

with these two methods.

For all three methods, the misclassification errors are higher when the data comes from multivariate t_5 distribution. This is not surprising since t_5 distribution has heavier tails than normal distribution leading to higher oracle misclassification error rate. The comparative performance of MGSDA, sparseLDA and penalizedLDA is the same for each distribution. The misclassification error rates are reported in Table 3.3 and the number of selected features in Table 3.4.

All three methods have similar misclassification rates for Identity case. For the Autoregressive and Data Based structures, penalizedLDA performs the worst. MGSDA and sparseLDA have comparable error rates for the Identity and Autoregressive structure, with MGSDA selecting a smaller number of features. Hence, MGSDA achieves the best tradeoff between the misclassification error and sparsity of the solution.

For the Data Based structure, MGSDA has significantly smaller error rate than sparseLDA, however it selects a larger number of features. Note that sparseLDA is restricted to have at most 160 features as in our simulations sparseLDA package produced errors otherwise. To make the comparison between MGSDA and sparseLDA for Data Based covariance structure clearer, we restricted the range of the tuning parameter for MGSDA to select the comparable number of features. The results are reported in Table 3.5. While the misclassification rate of MGSDA is worse with smaller number of features, it is still better than the misclassification rate for sparseLDA.

Table 3.3: Mean misclassification error rates as percentages over 100 replications, $G = 5$, standard deviation is given in parentheses.

		t_5 distribution		
Covariance	p	penLDA	sparseLDA	MGSDA
Identity	100	6.39(1.35)	7.2(1.6)	6.47(1.62)
	800	6.34(1.8)	7.84(1.83)	6.78(1.58)
Autoregressive	100	16.02(2.49)	8.76(2.01)	9.04(2.02)
	800	15.51(2.25)	10.41(2.08)	10.84(1.89)
Data Based	100	72.89(4.09)	20.34(3.04)	15.64(2.64)
	800	78.79(2.2)	37.44(5.39)	7.54(1.76)
		normal distribution		
Covariance	p	penLDA	sparseLDA	MGSDA
Identity	100	1.93(0.91)	2.45(1.18)	1.89(0.84)
	800	1.7(0.85)	2.98(1.21)	2.06(0.94)
Autoregressive	100	11.14(2.21)	5.11(1.56)	5.63(1.55)
	800	10.72(1.98)	6.42(1.82)	7.13(1.78)
Data Based	100	70.47(5.39)	14.24(2.73)	9.33(2.21)
	800	78.7(2.36)	26.47(4.06)	3.4(1.45)

Table 3.4: Mean number of selected features over 100 replications, $G = 5$, standard deviation is given in parentheses.

		t_5 distribution		
Covariance	p	penLDA	sparseLDA	MGSDA
Identity	100	24(24)	36(20)	16(11)
	800	41(39)	47(32)	16(14)
Autoregressive	100	12(9)	41(15)	29(18)
	800	25(26)	70(43)	29(33)
Data Based	100	59(33)	80(4)	93(8)
	800	458(340)	139(6)*	330(29)*
		normal distribution		
Covariance	p	penLDA	sparseLDA	MGSDA
Identity	100	34(32)	31(19)	13(10)
	800	26(29)	33(23)	11(3)
Autoregressive	100	15(15)	43(17)	31(20)
	800	19(26)	57(36)	15(11)
Data Based	100	48(34)	80(5)	92(9)
	800	411(354)	138(5)*	302(29)*

Table 3.5: Comparison of MGSDA and sparseLDA on Data Based covariance structure, $G = 5$ and $p = 800$, the tuning parameter for MGSDA is restricted to allow for comparable number of features with sparseLDA, standard deviation is given in parentheses.

	t_5 distribution		normal distribution	
	sparseLDA	MGSDA	sparseLDA	MGSDA
Error	37.44(5.39)	23.75(4.15)	26.47(4.06)	16.46(3.72)
Features	139(6)	152(14)	138(5)	143(13)

3.4.3 Implementation Details

The method of [12] is implemented using linprogPD function from the package CLIME from CRAN. Note that linprogPD almost never returns a sparse solution. However, all the values below the precision level should be treated as zeroes [**?**]. We used the default value of 10^{-3} for precision. The grid for the tuning parameter is chosen from 0.01 to 0.5 by 0.01. The method of [69] is implemented using the package penalizedLDA from CRAN. The grid for tuning parameter is chosen from 0 to 1 by 0.01. The method of [16] is implemented using the package sparseLDA from CRAN. Each canonical vector is constrained to have between 3 and $0.8n$ features. This is quite a restrictive range for tuning, however the sparseLDA package produced errors when we used a wider range of features. MGSDA is implemented using the R package MGSDA. The grid for the tuning parameter $\lambda_1 \leq ... \leq \lambda_{max}$ is chosen adaptively for each dataset with $\lambda_{max} = \max_j \|d_j\|_2$, which corresponds to zero selected features. For each $\lambda_l < \lambda_{\max}$ we set $V^{(0)} = \hat{V}(\lambda_{l+1})$. For all the methods, the final tuning parameter is chosen from the respective grid through 5-fold cross-validation to minimize the error rate.

Witten and Tibshirani's penalizedLDA has significantly faster running time than all other methods since penalizedLDA assumes that the covariance matrix

has diagonal structure. This assumption results in a simplified optimization algorithm, for details we refer to [69]. The running time of penalizedLDA is followed by MGSDA and sparseLDA. Surprisingly, LPD has the slowest performance. We suspect that this is not due to the method itself, but due to the use of linprogPD function in its implementation. A different linear program solver is likely to result in much faster running time, however the use of a general solver makes the method implementation less straightforward.

3.5 Real Data

3.5.1 Metabolomics Dataset

Metabolomics is the global study of all metabolites in a biological system under a given set of conditions. Metabolites are the final products of enzymes and enzyme networks whose substrates and products often cannot be deduced from genetic information and whose levels reflect the integrated product of the genome, proteome and environment. Metabolomic readouts thus represent the most direct (or phenotypic) readout of a cells physiologic state. From a technical standpoint, analytical studies of metabolism have been historically limited to one or a limited set of metabolites. However, advances in liquid chromatography and mass spectrometry have recently made it possible to measure hundreds of metabolites and with enough biomass well over 1000, in parallel. Such technologies have thus opened the door to obtaining global biochemical readouts of a cells physiologic state and response to perturbation. Cornell researchers have developed and applied a state-of-the-art metabolomic platform to track the

intrabacterial pharmacokinetic fates and pharmacodynamic actions of a given compound within Mycobacterium tuberculosis [44, 20, 19, 13]. These studies demonstrate the highly unpredictable nature and identities of these properties even for well-studied antibiotics.

We investigate a (currently unpublished) metabolomics data obtained from Dr. Kyu Rhee, which seeks to systematically elucidate the intrabacterial pharmacokinetic and pharmacodynamic fates and actions of antimycobacterial hit or lead compound series identified in high throughput screens against replicating and non- or slowly replicating forms of Mycobacterium tuberculosis.

The data contains measurements of 171 metabolic responses of 68 patients to 25 antibiotics that are administered at different dosage levels. Each measurement is an average of three replicates, normalized to the vehicle control and log2 transformed. 14 out of 25 antibiotics can be divided into the following 5 groups: STREP_AMI(strep, ami), FLQ(lev, moxi), DHFR(nitd2, sri8210, sri 8710, sri 8857), DHPS(smx, snl, aps) and InhA(eta, isoxyl, gsk93). These antibiotics are administered to 35 patients out of 68. In the subsequent analysis we only focus on these 5 groups of antibiotics and do not consider the dosage levels.

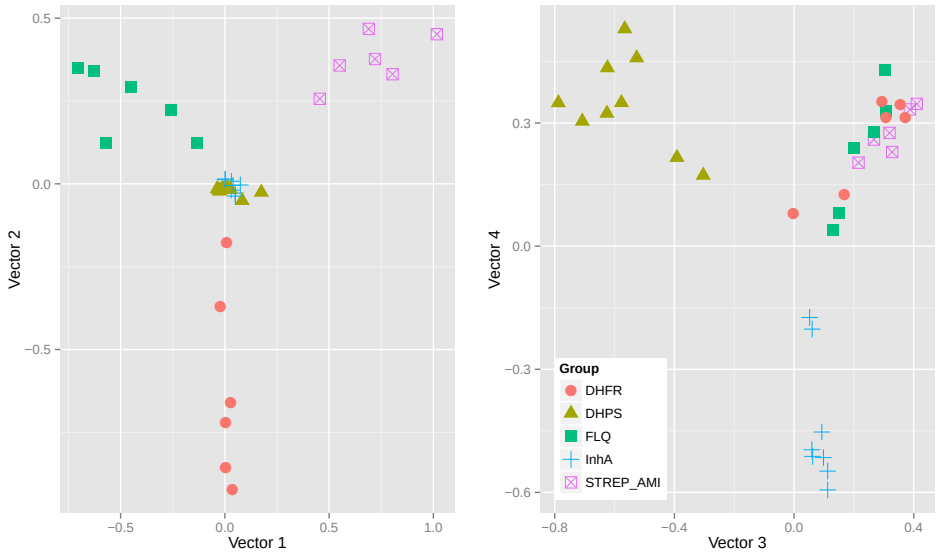
We compare the performance of MGSDA, penalizedLDA [69] and sparseLDA [16] on this dataset using the following measures: the mean number of misclassified samples and the mean number of selected features over 100 replications of 5-fold cross-validation. We do not perform random splits into the training and test set due to the small sample size. The results are reported in Table 3.6.

The results show that all three methods perform very well in terms of mis-

Table 3.6: Mean number of misclassified samples and mean number of selected features over 100 replications on metabolomics dataset, standard deviation is given in brackets.

	MGSDA	penalizedLDA	sparseLDA
CV error	0.074(0.11)	0.072(0.11)	0.014(0.05)
Features	19(8)	165(13)	34(12)

Figure 3.2: Metabolomics dataset projected onto 4 column vectors of V, $k = 8$ metabolite features are used.



classification error; the mean number of misclassified samples is significantly less than one indicating that all three methods lead to almost perfect classification performance. Such a good performance suggests that there is a significant difference in the metabolic responses between the 5 groups of antibiotics. However, penalizedLDA achieves this performance by selecting almost all of the metabolites, whereas MGSDA and sparseLDA use less than 20% of the original features. Note that there is a substantial variation between the replications due to the small sample size of the data.

We further estimate four canonical vectors using MGSDA with $\lambda = 0.57$

and illustrate the projected data in Figure 3.2. Note that 8 selected metabolites provide perfect linear separation between the groups. $\lambda = 0.57$ is chosen as one of the hundred tuning parameters from above replications of cross-validation splits. We have tried the other values of λ as well, however they all provided perfect linear separation between the groups with projected data being very similar to Figure 3.2. Though there is a variation between the cross-validation replications due to the small sample size of the data, this variation has negligible effect on the final projection.

3.5.2 14 Cancer Dataset

In this section we compare the performance of MGSDA, penalizedLDA [69] and sparseLDA [16] on the 14 cancer dataset by [47]. This dataset contains 16063 gene expression measurements collected on 198 samples. Each sample belongs to one of the 14 cancer classes. The dataset can be obtained from `http://statweb.stanford.edu/~tibs/ElemStatLearn/`. We selected this dataset as it is publicly available and has been previously analyzed by a number of authors including [69].

Following the recommendation of [29, p. 654], we first standardize the data to have mean zero and standard deviation one for each patient. To reduce the overall computational cost, we restrict the analysis to 3000 genes. We select these genes following the novel model-free feature screening procedure for discriminant analysis of [17]. Following the approach taken by [69], we perform 100 independent splits of the data set into the training set containing 75% of the samples and the test set containing 25% of the samples. The tuning parame-

Table 3.7: Mean number of misclassified samples and mean number of selected features over 100 splits on 14 cancer dataset, standard deviation is given in brackets.

	MGSDA	penalizedLDA	sparseLDA
Error	7.76(2.35)	13.41(2.49)	9.29(2.40)
Features	295(78)	2962(34)	293(55)

ter for all methods is selected using 5-fold cross-validation on the training set. The mean number of misclassified samples on the test set and the mean number of selected features over 100 splits are reported in Table 3.7. MGSDA and sparseLDA perform better than penalizedLDA in terms of the misclassification error and select much smaller number of features. MGSDA and sparseLDA select comparable number of features, with the misclassification error of MGSDA being the smallest.

3.6 Discussion

This paper introduces a novel procedure that estimates population canonical vectors in the multi-group setting, and a corresponding R package MGSDA is available on CRAN. The proposed method is a natural generalization of the two-group methods that were previously studied in the literature. In addition to being computationally tractable, the method performs feature selection which results in sparse canonical vectors. The group penalty eliminates features from all canonical vectors at once with the remaining non-zero features being the same for all the vectors.

One possible extension of the proposed method is to allow canonical vectors

to have different sparsity patterns. This goal can be achieved through the addition of the within-row penalty term to the objective function (3.8). Such an estimation procedure has already been considered in the regression context; for example, [54] propose the following optimization problem:

$$\hat{\beta}(\lambda, \alpha) = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|Y - X\beta\|_2^2 + (1-\alpha)\lambda \sum_{g=1}^{G} \sqrt{p_g} \|\beta^{(g)}\|_2 + \alpha\lambda\|\beta\|_1,$$

where p_g is the size of group g. In our case this approach results in

$$\hat{V}(\lambda, \alpha) = \arg\min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \operatorname{Tr}\left(V^\top W V\right) + \frac{1}{2}\|D^\top V - I\|_F^2 + (1-\alpha)\lambda \sum_{i=1}^{p} \|v_i\|_2 + \alpha\lambda\|V\|_1.$$

Another possible extension is to perform canonical vectors selection in addition to feature selection, which will enhance the interpretability, especially when the number of groups G is large. This goal can be achieved through the addition of the nuclear norm penalty term to the objective function (3.8):

$$\hat{V}(\lambda, \alpha) = \arg\min_{V \in \mathbb{R}^{p \times (G-1)}} \frac{1}{2} \operatorname{Tr}\left(V^\top W V\right) + \frac{1}{2}\|D^\top V - I\|_F^2 + \lambda \sum_{i=1}^{p} \|v_i\|_2 + \alpha\|V\|_*.$$

Depending on the value of $\alpha > 0$, the resulting matrix $\hat{V}$ has rank that is less than $G - 1$, effectively resulting in a lower-dimensional eigenspace.

Both extensions result in convex optimization problems, but require additional modifications to the optimization algorithm. An interesting direction for future research is to examine how these extensions compare to the original method in different scenarios.

An underlying assumption of MGSDA is the equality of covariance structures between groups. A small simulation study of the effect of covariance misspecification is given in the supplement. It indicates that MGSDA is robust with respect to this assumption, outperforming the traditional quadratic discrimi-

nant analysis (QDA) in the $p < n$ setting. This result is supported by previous research, which demonstrated the superiority of LDA over QDA for small samples and moderate values of p (see Chapter 6.3.2 in [51] and the references therein). It is of interest to investigate whether a direct estimation procedure can be applied to QDA to improve its performance in high-dimensional settings.

We established the variable selection and classification consistency of proposed estimator in the regime where $\frac{\log(ps)s^2}{N} \to 0$. While preparing this manuscript, we became aware of the work of [32], who show variable selection consistency of the sparse discriminant analysis under the conditions that $G = 2$ and $N \geq Cs\log((p-s)\log(N))$ for some constant $C > 0$. These improved rates directly apply to our proposal in the case $G = 2$, however the extension of these results to the case $G > 2$ is not clear. This is another direction for future research.

3.7 Additional simulation results

3.7.1 Simulation results when $G = 2$

In this section we present additional simulation results for the two-group case when $\mu_1 = 0_p$ and $\mu_2 = (1_s, 0_{p-s})$. We consider both $s = 10$ and $s = 30$ with the following covariance structures:

- **Equicorrelation:** $\Sigma_W = (\Sigma_{ij})_{p\times p}$ with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.5$ for $i \neq j$.
- **Bernoulli:**. $\Sigma_W = \Omega^{-1}$ with $\Omega = (B + \delta I)/(1 + \delta)$. Here $B = (b_{ij})_{p\times p}$ with $b_{ii} = 1$ for $1 \leq i \leq p$, $b_{ij} = b_{ji} = 0.5 \times Ber(1, 0.2)$ for $1 \leq i \leq s_0$,

$i < j \leq p$ and $b_{ij} = b_{ji} = 0.5$ for $s_0 + 1 \leq i \leq p$, $i < j \leq p$. δ is taken as $\delta = \max(-\lambda_{min}(B), 0) + 0.05$ to ensure that Ω is positive definite.

The Bernoulli structure has been previously used by [12]. The results are reported in Tables 3.8 and 3.9. MGSDA and LPD have similar performance for the Equicorrelation structure, with LPD being significantly better for Bernoulli structure at the expense of selecting a larger number of features. Note that for the Bernoulli structure the value of ρ has a significant effect on both the misclassification error rate and the sparsity of the solution.

Table 3.8: Mean misclassification error rates as percentages over 100 replications, $G = 2$, standard deviation is given in brackets.

Covariance	s	p	MGSDA	LPD, $\rho = 0.15$	LPD, $\rho = 2$	Oracle
Equicorrelation	10	100	3.32(1.25)	3.38(1.7)	3.02(1.58)	1.51(0.89)
	10	800	3.11(1.25)	2.98(1.39)	2.79(1.17)	1.45(0.81)
	30	100	0.55(0.53)	0.55(0.67)	0.52(0.73)	0.06(0.2)
	30	800	0.27(0.38)	0.5(0.61)	0.56(0.77)	0(0)
Bernoulli	10	100	6.12(1.69)	5.88(1.48)	5.75(1.62)	4.37(1.35)
	10	800	37.14(6.04)	17.03(3.4)	28.62(3.59)	4.6(1.49)
	30	100	0.35(0.42)	0.3(0.38)	0.22(0.34)	0.05(0.15)
	30	800	8.27(2.81)	3.29(1.4)	7.17(2.74)	0.04(0.14)

Table 3.9: Mean number of selected features over 100 replications, $G = 2$, standard deviation is given in brackets.

Covariance	s	p	MGSDA	LPD, $\rho = 0.15$	LPD, $\rho = 2$
Equicorrelation	10	100	51(5)	55(12)	73(10)
	10	800	84(13)	90(54)	128(48)
	30	100	77(4)	78(9)	93(6)
	30	800	147(13)	112(54)	177(40)
Bernoulli	10	100	24(9)	18(11)	20(18)
	10	800	43(33)	70(83)	19(14)
	30	100	43(8)	39(14)	33(11)
	30	800	116(32)	216(117)	43(18)

3.7.2 Simulation results when $G = 3$

This simulation scenario considers the classification between the three groups with $\mu_1 = 0_p$, $\mu_2 = (1_{s/2}, -1_{s/2}, 0_{p-s})$ and $\mu_3 = (-1_{s/2}, 1_{s/2}, 0_{p-s})$. The simulations are performed for the values of $s = 10$ and $s = 30$. The sample size for each group is $n = 100$

The results are reported in Tables 3.10 and 3.11. The population canonical vectors matrix Ψ is truly row sparse only in the Identity case, it is only approximately row sparse in other scenarios. The results suggest that all three methods are comparable in terms of misclassification rate except for the Autoregressive covariance structure. In this scenario, both MGSDA and sparseLDA outperform the penalizedLDA. In terms of the number of features, MGSDA tends to select fewer than its competitors.

3.7.3 Simulation results when $p < n$

In this section we compare the performance of MGSDA, penalizedLDA, sparseLDA and traditional LDA in the scenario where $p = 50$ and $p = 90$. The sample size for each group is $n = 100$. The group means are $\mu_1 = 0_p$, $\mu_2 = (1_5, -1_5, 0_{p-10})$ and $\mu_3 = (-1_5, 1_5, 0_{p-10})$. We consider three covariance structures: Identity, Equicorrelation ($\rho = 0.5$) and Autoregressive ($\rho = 0.8$). The results are reported in Figure 3.3. The misclassification errors of MGSDA and sparseLDA are comparable in all the scenarios, with the number of false positive features being consistently lower for MGSDA. The misclassification error of penalizedLDA is the best when covariance structure is Identity, which is not surprising since the diagonal covariance structure is an underlying assumption

Table 3.10: Mean misclassification error rates as percentages over 100 replications, $G = 3$, standard deviation is given in brackets.

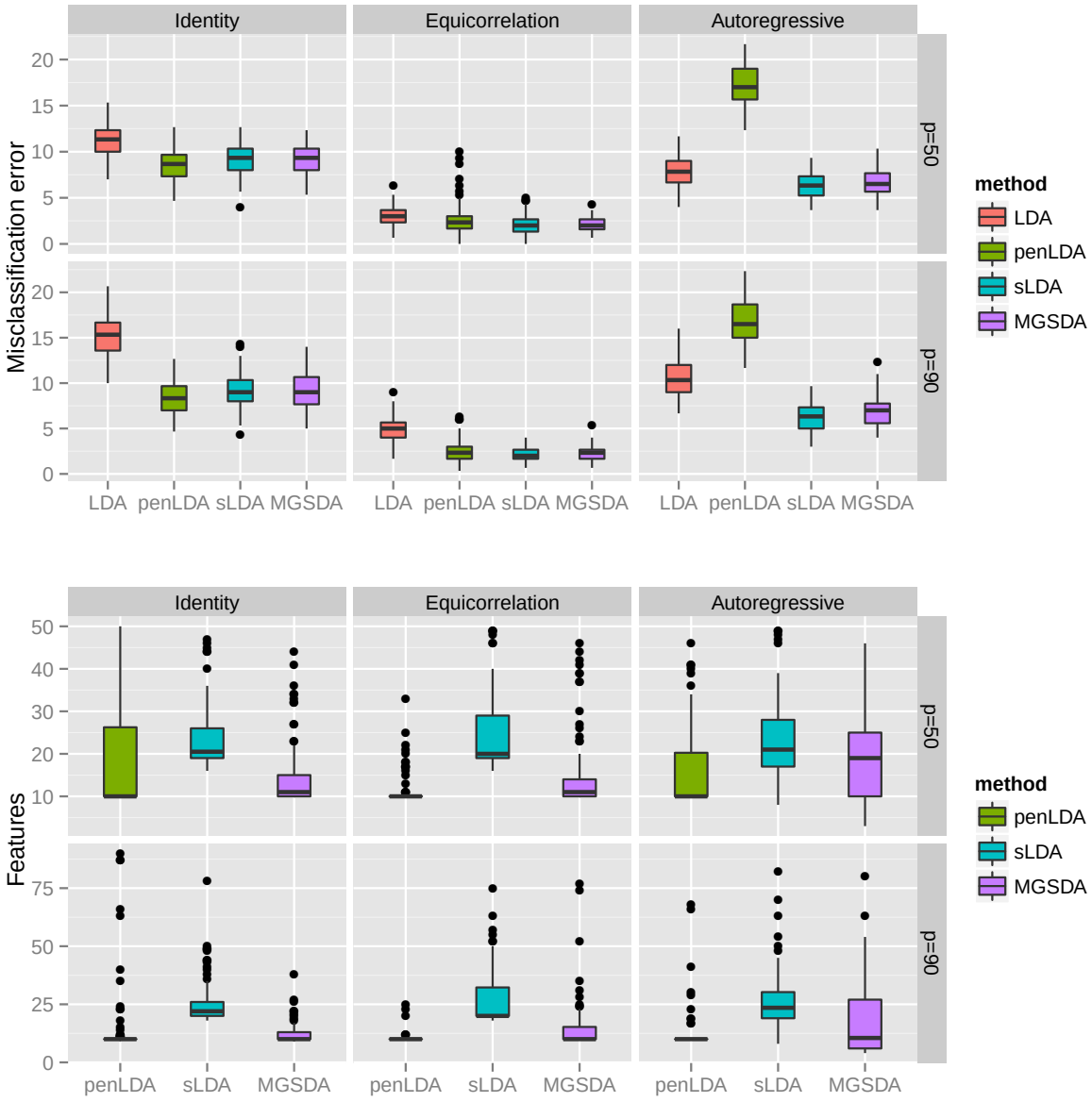
Covariance	s	p	MGSDA	penalizedLDA	sparseLDA	Oracle
Identity	10	100	9.11(1.52)	8.4(1.6)	9.39(1.68)	7.83(1.41)
	10	800	9.22(1.73)	7.92(1.5)	9.58(1.87)	7.67(1.43)
	30	100	1.06(0.67)	0.5(0.42)	0.93(0.63)	0.42(0.37)
	30	800	1.43(0.81)	0.52(0.39)	1.14(0.72)	0.43(0.35)
Equicorrelation	10	100	2.15(0.97)	2.34(1.18)	2.03(0.94)	1.65(0.86)
	10	800	2.19(0.89)	2.12(1.1)	2.15(0.85)	1.68(0.84)
	30	100	0.23(0.38)	0.3(1.11)	0.26(0.44)	0.01(0.05)
	30	800	0.31(0.43)	0.04(0.11)	0.39(0.52)	0.01(0.06)
Autoregressive	10	100	6.83(1.4)	16.87(2.16)	6.34(1.32)	4.87(1)
	10	800	7.29(1.77)	16.53(2.44)	7.47(2.54)	4.9(1.17)
	30	100	5.45(1.48)	16(1.95)	4.86(1.44)	3.57(1.05)
	30	800	5.89(1.53)	15.46(2.33)	5.98(2.03)	3.65(1.05)
Bernoulli	10	100	11.15(1.91)	10.56(1.73)	11.35(1.88)	8.51(1.62)
	10	800	43.13(3.06)	44.84(4)	41.72(3.19)	8.56(1.67)
	30	100	1.54(0.72)	1.05(0.56)	1.37(0.74)	0.59(0.47)
	30	800	20.59(3.08)	22.66(4.11)	16.42(2.57)	0.63(0.46)

Table 3.11: Mean number of selected features over 100 replications, $G = 3$, standard deviation is given in brackets.

Covariance	s	p	MGSDA	penalizedLDA	sparseLDA
Identity	10	100	13(7)	15(15)	26(11)
	10	800	11(2)	15(8)	24(9)
	30	100	46(18)	61(19)	58(11)
	30	800	37(11)	51(93)	67(18)
Equicorrelation	10	100	14(8)	10(1)	27(12)
	10	800	12(6)	12(3)	28(19)
	30	100	29(11)	38(12)	47(8)
	30	800	29(14)	30(0)	50(7)
Autoregressive	10	100	21(16)	12(6)	27(12)
	10	800	7(3)	15(11)	29(19)
	30	100	28(16)	54(25)	39(11)
	30	800	14(5)	36(40)	49(24)
Bernoulli	10	100	14(10)	16(15)	30(13)
	10	800	118(115)	33(61)	100(46)
	30	100	51(21)	59(22)	61(11)
	30	800	42(33)	48(14)	108(33)

of penalizedLDA. Usual LDA performs the worst except when the covariance structure is Autoregressive. In this case, it outperforms penalizedLDA.

Figure 3.3: Misclassification error (in percentage) and the number of selected features over 100 replications, $p < n$.

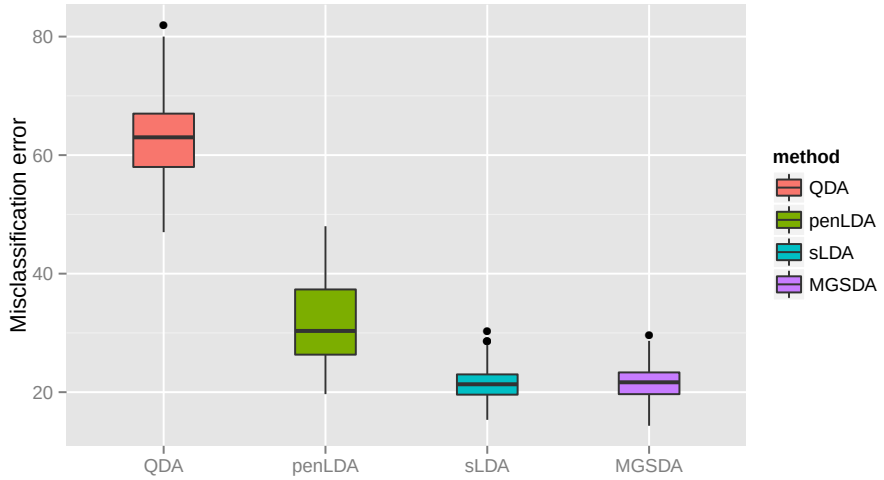


3.7.4 Robustness with respect to the assumption of equal covariance matrices

In Section 2.2, we assume that observations $X_i \in \mathbb{R}^p$ come from G groups with the same within-group covariance matrix Σ_W. In practice, this assumption is likely to be violated, requiring separate estimation of each group covariance structure. The task of covariance estimation is extremely difficult when the number of groups is large and the number of samples is small. Therefore, in this section we investigate whether the traditional quadratic discriminant analysis (QDA) outperforms the high-dimensional LDA methods when the assumption of equal within-group covariance matrices is violated.

We consider the case of three groups with the following covariance structures: equicorrelation with $\rho = 0.5$, equicorrelation with $\rho = 0.8$ and data based (as described in Section 4). To make the comparison between high-dimensional LDA methods and QDA fair, we consider the case $p < n$ with $p = 50$ and the sample size for each group $n = 100$. The group means are $\mu_1 = 0_p$, $\mu_2 = (0.4_5, -0.4_5, 0_{p-10})$ and $\mu_3 = (-0.4_5, 0.4_5, 0_{p-10})$. Figure 3.4 indicates that even when the covariance matrices are not the same, the high-dimensional LDA methods outperform traditional quadratic discriminant analysis. Again, the performance of sparseLDA and MGSDA is almost identical.

Figure 3.4: Misclassification error (in percentage) over 100 replications, unequal covariance matrices, $p = 50$ and $n_g = 100$.



3.8 Technical proofs

3.8.1 Proofs of auxillary lemmas for Theorem 1.

Lemma 1. $\|AB\|_{\infty,2} \leq \|A\|_\infty \|B\|_{\infty,2}$

Proof. This inequality is a special case of Lemma 8 in [42]. Note that

$$\|A\|_{\infty,2} = \max_i \|a_i\|_2 = \max_i \max_{\|y_i\|_2 \leq 1} |y^t a_i| = \max_{\|y\|_2 \leq 1} \max_i |y^t a_i| = \max_{\|y\|_2 \leq 1} \|Ay\|_\infty.$$

It follows that

$$\|AB\|_{\infty,2} = \max_{\|y\|_2 \leq 1} \|ABy\|_\infty \leq \max_{\|y\|_2 \leq 1} \|A\|_\infty \|By\|_\infty = \|A\|_\infty \max_{\|y\|_2 \leq 1} \|By\|_\infty = \|A\|_\infty \|B\|_{\infty,2}.$$

□

Lemma 2. *Define $F = D - \Delta$. Then there exists constant $c_3 > 0$ such that*

$$P\left(\|F\|_{\infty,2} \geq \epsilon\right) \leq 2p(G-1)\exp(-c_3 N\epsilon^2).$$

Proof. From the definition of Δ and under the assumption $\pi_g = 1/G$, its rth column has the form $\Delta_r = \sum_{i=1}^{r}(\mu_i - \mu_{r+1})/\sqrt{Gr(r+1)}$. Similarly, $D_r = \sum_{i=1}^{r}(\bar{X}_i - \bar{X}_{r+1})/\sqrt{Gr(r+1)}$. Therefore,

$$F_r = \frac{1}{\sqrt{Gr(r+1)}} \sum_{i=1}^{r} \big((\bar{X}_i - \bar{X}_{r+1}) - (\mu_i - \mu_{r+1})\big).$$

Since the groups are independent and $(\bar{X}_g)_j \sim N\left((\mu_g)_j, \frac{\sigma_j^2}{n}\right)$ for all $g \in \{1, ..., G\}$ and $j \in \{1, ..., p\}$, then for all r:

$$\sum_{i=1}^{r}(\bar{X}_i - \bar{X}_{r+1})_j \sim N\left(\sum_{i=1}^{r}\big(\mu_i - \mu_{r+1}\big)_j, \frac{r(r+1)\sigma_j^2}{n}\right),$$

or equivalently

$$d_{jr} \sim N\left(\Delta_{jr}, \frac{\sigma_j^2}{Gn}\right),$$

where d_{jr} are the elements of matrix D and Δ_{jr} are the elements of matrix Δ. It follows that for all $r \in \{1, ..., G-1\}$ and for all $j \in \{1, ..., p\}$

$$P(|f_{jr}| \geq \epsilon) = P(|d_{jr} - \Delta_{jr}| \geq \epsilon) \leq 2\exp\left(-\frac{N\epsilon^2}{2\sigma_j^2}\right) \leq 2\exp(-cN\epsilon^2).$$

Therefore

$$\begin{aligned} P(\|f_j\|_2 \geq \epsilon) &= P\left(\sqrt{f_{j1}^2 + ... + f_{(G-1)j}^2} \geq \epsilon\right) \leq P\left(\sqrt{G-1}\max_r |f_{jr}| \geq \epsilon\right) \\ &\leq P\left(\cup_r \left\{|f_{jr}| \geq \frac{\epsilon}{\sqrt{G-1}}\right\}\right) \leq (G-1)P\left(|f_{jr}| \geq \frac{\epsilon}{\sqrt{G-1}}\right) \\ &\leq 2(G-1)\exp(-c_3 N\epsilon^2). \end{aligned}$$

The result follows by applying the union bound over $j \in \{1, ..., p\}$. □

Lemma 3. *Let $T = W + B$ and $\Sigma = \Sigma_W + \Sigma_B$. Then there exist constants $c_1 > 0$ and $c_2 > 0$ such that*

$$P\left(\|T_{AA} - \Sigma_{AA}\|_\infty \geq \epsilon\right) \leq c_1 s^2 \exp(-c_2 N s^{-2}\epsilon^2);$$

$$P\left(\|T_{A^cA} - \Sigma_{A^cA}\|_\infty \geq \epsilon\right) \leq c_1 s(p-s)\exp(-c_2 N s^{-2}\epsilon^2).$$

Proof. First, we show that $P(|\Sigma_{ij} - T_{ij}| > \epsilon)] \leq c_1 \exp(-c_2 N\epsilon^2)$. By definition,

$$\begin{aligned}
\Sigma_{ij} - T_{ij} &= \Sigma_{W_{ij}} + \Sigma_{B_{ij}} - \frac{1}{N}\sum_{k=1}^{N}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j) \\
&= \Sigma_{W_{ij}} + \sum_{g=1}^{G}\pi_g(\mu_{gi} - \mu_i)(\mu_{gj} - \mu_j) - \frac{1}{N}\sum_{k=1}^{N}X_{ki}X_{kj} + \bar{X}_i\bar{X}_j \\
&= \Sigma_{W_{ij}} + \sum_{g=1}^{G}\pi_g\mu_{gi}\mu_{gj} + \bar{X}_i\bar{X}_j - \mu_i\mu_j - \frac{1}{N}\sum_{g=1}^{G}\sum_{k\in I_g}X_{ki}X_{kj}.
\end{aligned}$$

Furthermore

$$\frac{1}{n_g}\sum_{k\in I_g}X_{ki}X_{kj} = \frac{1}{n_g}\sum_{k\in I_g}(X_{ki}-\mu_{gi})(X_{kj}-\mu_{gj})+\mu_{gi}(\bar{X}_{gj}-\mu_{gj})+\mu_{gj}(\bar{X}_{gi}-\mu_{gi})+\mu_{gj}\mu_{gi}.$$

Therefore

$$\begin{aligned}
\Sigma_{ij} - T_{ij} =& \Sigma_{W_{ij}} + \sum_{g=1}^{G}\pi_g\mu_{gi}\mu_{gj} + \bar{X}_i\bar{X}_j - \mu_i\mu_j - \frac{1}{N}\sum_{g=1}^{G}n_g\times \\
&\times\left(\frac{1}{n_g}\sum_{k\in I_g}(X_{ki} - \mu_{gi})(X_{kj} - \mu_{gj}) + \mu_{gi}(\bar{X}_{gj} - \mu_{gj}) + \mu_{gj}(\bar{X}_{gi} - \mu_{gi}) + \mu_{gj}\mu_{gi}\right) \\
=& \sum_{g=1}^{G}\frac{n_g}{N}\left(\Sigma_{W_{ij}} - \frac{1}{n_g}\sum_{k\in I_g}(X_{ki} - \mu_{gi})(X_{kj} - \mu_{gj})\right) \\
&+ \sum_{g=1}^{G}\frac{n_g}{N}\left(\mu_{gi}(\mu_{gj} - \bar{X}_{gj}) + \mu_{gj}(\mu_{gi} - \bar{X}_{gi})\right) + \sum_{g=1}^{G}\left(\pi_g - \frac{n_g}{N}\right)\mu_{gi}\mu_{gj} \\
&+ (\bar{X}_i\bar{X}_j - \mu_i\mu_j).
\end{aligned}$$

Under the assumption $\pi_g = \frac{1}{G}$ and $n_g = \frac{1}{G}$, the above expression is further simplified as

$$\begin{aligned}
\Sigma_{ij} - T_{ij} =& \frac{1}{G}\sum_{g=1}^{G}\left(\Sigma_{W_{ij}} - \frac{1}{n_g}\sum_{k\in I_g}(X_{ki} - \mu_{gi})(X_{kj} - \mu_{gj})\right) \\
&+ \frac{1}{G}\sum_{g=1}^{G}\left(\mu_{gi}(\mu_{gj} - \bar{X}_{gj}) + \mu_{gj}(\mu_{gi} - \bar{X}_{gi})\right) + (\bar{X}_i\bar{X}_j - \mu_i\mu_j) \\
=& I_1 + I_2 + I_3.
\end{aligned}$$

For the final bound it remains to show that for each I_j there exist constants $c_{1j} > 0$ and $c_{2j} > 0$ such that $P(|I_j| \geq \epsilon) \leq c_{1j} \exp(-c_{2j} N\epsilon^2)$.

Analysis of I_1. From Lemma A.3 in [8], there exist constants $C_1 > 0$ and $C_2 > 0$ such that for $\epsilon < \epsilon_0$

$$P\left(\left|\frac{1}{n}\sum_{k=1}^{n}(Z_{ik}Z_{jk} - \sigma_{jk})\right| \geq \epsilon\right) \leq C_1 \exp(-C_2 n\epsilon^2),$$

where Z_i are i.i.d $N(0, \Sigma)$ and σ_{ij} are elements of Σ. Let $\tilde{Z}_i = X_i - \mu_i$, where $\mu_i = \mu_g$ if observation i belongs to group g. By definition $\mathbb{E}(\tilde{Z}_i) = 0$ and $\tilde{Z}_i$ are i.i.d $N(0, \Sigma_W)$. Note that I_1 can be rewritten as

$$I_1 = \frac{1}{N}\sum_{g=1}^{G}\sum_{k\in I_g}\left(\Sigma_{W_{ij}} - (X_{ki} - \mu_{gi})(X_{kj} - \mu_{gj})\right) = \frac{1}{N}\sum_{l=1}^{N}(\Sigma_{W_{ij}} - \tilde{Z}_{li}\tilde{Z}_{lj}).$$

Therefore

$$P(|I_1| \geq \epsilon) \leq P\left(\left|\frac{1}{N}\sum_{l=1}^{N}(\Sigma_{W_{ij}} - \tilde{Z}_{li}\tilde{Z}_{lj})\right| \geq \epsilon\right) \leq C_1 \exp(-C_2 N\epsilon^2).$$

Analysis of I_2. Dy definition, $I_2 = \frac{1}{G}\sum_{g=1}^{G}\left(\mu_{gi}(\mu_{gj} - \bar{X}_{gj}) + \mu_{gj}(\mu_{gi} - \bar{X}_{gi})\right)$. It follows that

$$|I_2| \leq 2\max\left(\left|\frac{1}{G}\sum_{g=1}^{G}\mu_{gi}(\mu_{gj} - \bar{X}_{gj})\right|, \left|\frac{1}{G}\sum_{g=1}^{G}\mu_{gj}(\mu_{gi} - \bar{X}_{gi})\right|\right).$$

Since the groups are independent,

$$\frac{1}{G}\sum_{g=1}^{G}\mu_{gi}(\mu_{gj} - \bar{X}_{gj}) \sim N\left(0, \frac{\sum_{g=1}^{G}\mu_{gi}^2}{NG}\sigma_j^2\right).$$

Therefore,

$$P\left(\left|\frac{1}{G}\sum_{g=1}^{G}\mu_{gi}(\mu_{gj} - \bar{X}_{gj})\right| \geq \epsilon\right) \leq 2\exp(-cN\epsilon^2),$$

hence

$$P(|I_2| \geq \epsilon) \leq 4\exp\left(-\frac{c}{4}N\epsilon^2\right).$$

Analysis of I_3. Note that

$$\begin{aligned} I_3 = \bar{X}_i\bar{X}_j - \mu_i\mu_j &= \sum_{g=1}^{G}\frac{1}{G}\bar{X}_{gi}\sum_{l=1}^{G}\frac{1}{G}\bar{X}_{lj} - \sum_{g=1}^{G}\frac{1}{G}\mu_{gi}\sum_{l=1}^{G}\frac{1}{G}\mu_{lj} \\ &= \sum_{g=1}^{G}\frac{1}{G}\left(\bar{X}_{gi}-\mu_{gi}\right)\sum_{l=1}^{G}\frac{1}{G}\left(\bar{X}_{lj}-\mu_{lj}\right) \\ &\quad + \mu_i\sum_{l=1}^{G}\frac{1}{G}(\bar{X}_{lj}-\mu_{lj}) + \mu_j\sum_{g=1}^{G}\frac{1}{G}(\bar{X}_{gi}-\mu_{gi}). \end{aligned}$$

Therefore

$$|I_3| \leq \max_{t\in\{i,j\}}\left|\frac{1}{G}\sum_{g=1}^{G}(\bar{X}_{gt}-\mu_{gt})\right|^2 + 2\max_g|\mu_g|\max_{t\in\{i,j\}}\left|\frac{1}{G}\sum_{g=1}^{G}(\bar{X}_{gt}-\mu_{gt})\right|.$$

Since the groups are independent,

$$\frac{1}{G}\sum_{g=1}^{G}\bar{X}_{gj} \sim N\left(\frac{1}{G}\sum_{g=1}^{G}\mu_{gj}, \frac{\sigma_j^2}{nG}\right).$$

Therefore,

$$P\left(\max_{t\in\{i,j\}}\left|\frac{1}{G}\sum_{g=1}^{G}(\bar{X}_{gt}-\mu_{gt})\right| \geq \epsilon\right) \leq 2\cdot 2\exp(-c_1N\epsilon^2).$$

Note that if $\max_{t\in\{i,j\}}\left|\frac{1}{G}\sum_{g=1}^{G}(\bar{X}_{gt}-\mu_{gt})\right| \leq k\epsilon$, then $|I_3| \leq k^2\epsilon^2 + 2k\epsilon\max_g|\mu_g|$. Choosing $k \leq \frac{1}{\max(\sqrt{2},2\max_g\|\mu_g\|)}$ leads to $|I_3| \leq \frac{\epsilon^2+\epsilon}{2} \leq \epsilon$ for small values of ϵ. Hence,

$$P(|I_3| \geq \epsilon) \leq 4\exp(-cN\epsilon^2).$$

Combining the results for I_1-I_3 leads to $c_1 = C_1 + 4 + 4$. The results of lemma follow from the definition of $\|\cdot\|_\infty$ and the union bound. □

Lemma 4. *Let $F_T = T_{A^cA}(T_{AA})^{-1} - \Sigma_{A^cA}(\Sigma_{AA})^{-1}$. Then there exists constant $c_3 > 0$ such that*

$$P\left(\|F_T\|_\infty \geq \epsilon\phi(\kappa+1)(1-\phi\epsilon)^{-1}\right) \leq 6(G+1)ps\exp(-c_3Ns^{-2}\epsilon^2).$$

Proof. This result is a multi-group version of Lemma A2 in [37]. The proof follows the proof of Lemma A2 in [37] and uses the results of Lemma 3. □

3.8.2 Proof of Theorem 1

Proof. The proof follows the proof of Theorem 1 in [37] using the results of auxillary lemmas. For simplicity of illustration, we consider the case $\pi_i = \frac{1}{G}$ and $n_i = \frac{N}{G}$. **Part 1.** First we derive the conditions under which $\hat{V}(\lambda)_{A^c} = 0$. Let $T = W + DD^t$. From KKT conditions on $V = \hat{V}(\lambda)$, $TV - D + \lambda u = 0$, where u is the subgradient of $\sum_{i=1}^{p} \|v_i\|_2$ such that

$$u_j = \begin{cases} \frac{v_j}{\|v_j\|_2}, & \text{if } v_j \neq 0; \\ \in \{u : \|u\|_2 \leq 1\}, & \text{if } v_j = 0. \end{cases}$$

Partition T as

$$T = \begin{pmatrix} T_{AA} & T_{AA^c} \\ T_{A^cA} & T_{A^cA^c} \end{pmatrix}.$$

Then

$$T_{AA}V_A + T_{AA^c}V_{A^c} - D_A + \lambda u_A = 0;$$

$$T_{A^cA}V_A + T_{A^cA^c}V_A^c - D_{A^c} + \lambda u_{A^c} = 0.$$

It follows that if $\|T_{A^cA}V_A - D_{A^c}\|_{\infty,2} \leq \lambda$, then $\hat{V}(\lambda)_{A^c} = 0$.

Next we derive the bounds for $\|T_{A^cA}V_A - D_{A^c}\|_{\infty,2}$. From KKT conditions

$$V_A = (T_{AA})^{-1}(D_A - \lambda u_A)$$

and therefore

$$\begin{aligned}
T_{A^cA}V_A - D_{A^c} =& T_{A^cA}(T_{AA})^{-1}\left(D_A - \lambda u_A\right) - D_{A^c} \\
=& T_{A^cA}(T_{AA})^{-1}(D_A - \Delta_A) + T_{A^cA}(T_{AA})^{-1}\Delta_A - T_{A^cA}(T_{AA})^{-1}\lambda u_A \\
& - D_{A^c} + \Delta_{A^c} - \Delta_{A^c} \\
=& T_{A^cA}(T_{AA})^{-1}(D_A - \Delta_A) + T_{A^cA}(T_{AA})^{-1}\Delta_A \\
& - T_{A^cA}(T_{AA})^{-1}\lambda u_A + (\Delta_{A^c} - D_{A^c}) - \Sigma_{A^cA}(\Sigma_{AA})^{-1}\Delta_A \\
=& \left(T_{A^cA}(T_{AA})^{-1} - \Sigma_{A^cA}(\Sigma_{AA})^{-1}\right)\Delta_A + (\Delta_{A^c} - D_{A^c}) \\
& + \left(T_{A^cA}(T_{AA})^{-1} - \Sigma_{A^cA}(\Sigma_{AA})^{-1}\right)(D_A - \Delta_A) + \Sigma_{A^cA}(\Sigma_{AA})^{-1}(D_A - \Delta_A) \\
& - \left(T_{A^cA}(T_{AA})^{-1} - \Sigma_{A^cA}(\Sigma_{AA})^{-1}\right)\lambda u_A - \Sigma_{A^cA}(\Sigma_{AA})^{-1}\lambda u_A.
\end{aligned}$$

Here we used the fact that $\Sigma_{AA}\tilde{\Psi}'_A = \Delta_A$ and $\Sigma_{A^cA}\tilde{\Psi}'_A = \Sigma_{A^cA}(\Sigma_{AA})^{-1}\Delta_A = \Delta_{A^c}$. Denote $K = \Sigma_{A^cA}(\Sigma_{AA})^{-1}$, $F_T = T_{A^cA}(T_{AA})^{-1} - \Sigma_{A^cA}(\Sigma_{AA})^{-1}$, $F_{A^c} = \Delta_{A^c} - D_{A^c}$ and $F_A = \Delta_A - D_A$. Then

$$\begin{aligned}
\|T_{A^cA}V_A - D_{A^c}\|_{\infty,2} &= \|F_T\Delta_A + F_{A^c} - F_TF_A - KF_A - F_T\lambda u_A - K\lambda u_A\|_{\infty,2} \\
&\leq \|F_T\|_{\infty,2}(\Delta + \lambda) + \|F_{A^c}\|_{\infty,2} + \|F_T\|_\infty\|F_A\|_{\infty,2} + \|F_A\|_{\infty,2}\kappa + \lambda\kappa \\
&\leq \|F_T\|_\infty(\Delta + \lambda) + \|F\|_{\infty,2}(1 + \|F_T\|_\infty + \kappa) + \lambda\kappa,
\end{aligned}$$

where we used Lemma 1 and $\|F\|_{\infty,2} = \max(\|F_A\|_{\infty,2}, \|F_{A^c}\|_{\infty,2})$. If $\|F_T\|_\infty \leq (\kappa+1)\epsilon\phi(1-\phi\epsilon)^{-1}$ and $\|F\|_{\infty,2} \leq \epsilon$ for some $\epsilon < \frac{1}{\phi}$ then

$$\|T_{A^cA}V_A - D_{A^c}\|_{\infty,2} \leq (\kappa+1)\epsilon\phi(1-\phi\epsilon)^{-1}(\Delta+\lambda) + \epsilon(1+(\kappa+1)\epsilon\phi(1-\phi\epsilon)^{-1}+\kappa) + \lambda\kappa.$$

Therefore $\|T_{A^cA}V_A - D_{A^c}\|_{\infty,2} \leq \lambda$ if

$$(\kappa+1)\epsilon\phi(1-\phi\epsilon)^{-1}(\Delta+\lambda) + \epsilon(1+(\kappa+1)\epsilon\phi(1-\phi\epsilon)^{-1}+\kappa) + \lambda\kappa \leq \lambda,$$

which is equivalent to

$$\epsilon \leq \frac{\lambda(1-\kappa)}{(\kappa+1)(\phi\Delta+1) + 2\phi\lambda}.$$

Applying Lemma 2 and Lemma 3 leads to

$$P\left(\|T_{A^cA}V_A - D_{A^c}\|_\infty \le \lambda\right) \ge 1 - c_1 ps\exp(-c_2 N s^{-2}\epsilon^2) - 2p(G-1)\exp(-c_3 N\epsilon^2).$$

Part 2. From KKT conditions,

$$\begin{aligned}\hat{V}(\lambda)_A = T_{AA}^{-1}\left(D_A - \lambda u_A\right) =& \Sigma_{AA}^{-1}\Delta + (T_{AA}^{-1} - \Sigma_{AA}^{-1} + \Sigma_{AA}^{-1})(D-\Delta) \\ &+ (T_{AA}^{-1} - \Sigma_{AA}^{-1})\Delta - \lambda(T_{AA}^{-1} - \Sigma_{AA}^{-1})u_A + \lambda\Sigma_{AA}^{-1}u_A.\end{aligned}$$

Denote $\nu_1 = \|T_{AA}^{-1} - \Sigma_{AA}^{-1}\|_\infty$. Then for any $j \in A$

$$\|\hat{v}_j\|_2 \ge \min_{i\in A}\|\tilde{\Psi}_i\|_2 - (\nu_1+\phi)(\lambda + \|F_A\|_{\infty,2}) - \nu_1\Delta.$$

Let $\nu_2 = \|T_{AA} - \Sigma_{AA}\|_\infty$. From the proof of Lemma A.2 in [37]:

$$\nu_1 < \phi^2\nu_2(1-\phi\nu_2)^{-1}.$$

If $\nu_2 \le \epsilon$ and $\|F_A\|_{\infty,2} \le \epsilon$, then $\|\hat{v}_j\|_2 > 0$ if

$$\epsilon < \frac{\tilde{\Psi}_{\min} - \lambda\phi}{\phi(1+\phi\Delta+\tilde{\Psi}_{\min})}.$$

It follows that

$$P\left(\|v_j\|_2 > 0 \text{ for all } j \in A\right) \ge 1 - c_1 s^2\exp(-c_2 N s^{-2}\epsilon^2) - 2s(G-1)\exp(-c_3 N\epsilon^2).$$

Part 3. From parts 1 and 2,

$$\|\hat{V}(\lambda)_A - \tilde{V}\|_{\infty,2} \le (\nu_1+\phi)(\|F_A\|_{\infty,2}+\lambda)+\nu_1\Delta \le (1-\phi\nu_2)^{-1}\phi(\|F_A\|_{\infty,2}+\lambda+\phi\nu_2\Delta).$$

If $\nu_2 \le \epsilon$ and $\|F_A\|_{\infty,2} \le \epsilon$, then $\|\hat{V}(\lambda)_A - \tilde{V}\|_{\infty,2} \le 2\phi\lambda$ if

$$\epsilon \le \frac{\lambda}{1+\phi\Delta+2\phi\lambda}.$$

□

3.8.3 Proof of Corollary 1

Proof. Follows directly from parts 1 and 2 of Theorem 1. □

3.8.4 Proof of Corollary 2

Proof. The first result follows directly from part 3 of Theorem 1. To show the second result, we consider the events

$$\begin{aligned}
\mathcal{E}_1 &= \cap_g \left\{ \left| \log \pi_g - \log \frac{n_i}{N} \right| \leq C_1 \frac{1}{\sqrt{N}} \right\}; \\
\mathcal{E}_2 &= \left\{ \|\hat{V}_A - \Psi_A\|_{\infty,2} \leq C_2 \sqrt{\frac{\log(ps)s^2}{N}} \right\}; \\
\mathcal{E}_3 &= \left\{ \|W_A - \Sigma_{WAA}\|_\infty \leq C_3 \sqrt{\frac{\log(s^2)s^2}{N}} \right\}; \\
\mathcal{E}_4 &= \cap_g \left\{ \|\mu_{gA} - \bar{x}_{gA}\|_\infty \leq C_4 \sqrt{\frac{\log(s)}{N}} \right\}; \\
\mathcal{E}_5 &= \{A = \hat{A}\},
\end{aligned}$$

where C_i are constants independent of n, p and s and let $\mathcal{E} = \cap_i \mathcal{E}_i$. Given a new observation $X \in \mathbb{R}^p$ with a value x, define for each $g \in \{1, ..., G\}$

$$\begin{aligned}
h^g &= h^g(x) = (x - \mu_g)^t \Psi' (\Psi'^t \Sigma_W \Psi')^{-1} \Psi'^t (x - \mu_g) - 2 \log \pi_g; \\
\hat{h}^g &= \hat{h}^g(x) = (x - \bar{x}_g)^t \hat{V} (\hat{V}^t W \hat{V})^{-1} \hat{V}^t (x - \bar{x}_g) - 2 \log \frac{n_g}{N}.
\end{aligned}$$

Since the classification rule is invariant to scaling and orthogonal rotation, it follows that population classification rule $h_\Psi(x) = \arg\min_g h^g(x)$ and the sample classification rule $\hat{h}_{\hat{V}}(x) = \arg\min_g \hat{h}^g(x)$. We first prove $\hat{h}^g \xrightarrow{P} h^g$: there exists constant C such that on $\mathcal{E}$

$$|h^g - \hat{h}^g| \leq C \sqrt{\frac{\log(ps)s^2}{N}},$$

and $P(\mathcal{E}) \to 1$ under (C1) and (C2). Let $a = \Psi_A^t(x_A - \mu_{gA})$, $\hat{a} = \hat{V}_A^t(x_A - \bar{x}_{gA})$, $\Lambda^{-1} = (\Psi_A \Sigma_{WAA} \Psi_A)^{-1}$ and $\hat{\Lambda}^{-1} = (\hat{V}_A^t W_{AA} \hat{V}_A)^{-1}$. Consider

$$
\begin{aligned}
|h^g - \hat{h}^g| &= |a^t \Lambda^{-1} a - \hat{a} \hat{\Lambda}^{-1} \hat{a}| \\
&= |(\hat{a} - a)^t (\hat{\Lambda}^{-1} - \Lambda^{-1})(\hat{a} - a) + 2a^t (\hat{\Lambda}^{-1} - \Lambda^{-1})(\hat{a} - a) \\
&\quad + 2a^t \Lambda^{-1} (\hat{a} - a) + a^t (\hat{\Lambda}^{-1} - \Lambda^{-1}) a| \\
&\leq (\|\hat{a} - a\|_2^2 + 2\|a\|_2 \|_2 \|\hat{a} - a\|_2 + \|a\|_2) \|\hat{\Lambda}^{-1} - \Lambda^{-1}\|_2 + 2\|a\|_2 \|\Lambda^{-1}\|_2 \|\hat{a} - a\|_2.
\end{aligned}
$$

By definition of $\hat{a}$ and a, on $\mathcal{E}$

$$
\begin{aligned}
\|\hat{a} - a\|_2 &= \|\Psi_A^t (x_A - \mu_{gA}) - \hat{V}_A^t (x_A - \bar{x}_{gA})\|_2 \\
&\leq \|x_A^t (\Psi_A - \hat{V}_A)\|_2 + \|(\mu_{gA} - \bar{x}_{gA})^t (\Psi_A - \hat{V}_A)\|_2 \\
&\leq \|\Psi_A - \hat{V}_A\|_{\infty,2} (\|x_A\|_\infty + \|\mu_{gA} - \bar{x}_{gA}\|_\infty).
\end{aligned}
$$

Therefore, there exists constant C' such that on the event $\mathcal{E}$

$$
\|\hat{a} - a\|_2 \leq C' \sqrt{\frac{\log(ps) s^2}{N}}.
$$

By definition of $\hat{\Lambda}$ and Λ, $\hat{\Lambda} = \Lambda + \Upsilon$, where

$$
\begin{aligned}
\Upsilon &= (\hat{V}_A - \Psi_A)^t (W_{AA} - \Sigma_{WAA})(\hat{V}_A - \Psi_A) \\
&\quad + \Psi_A^t (W_{AA} - \Sigma_{WAA})(\hat{V}_A - \Psi_A) + (\hat{V}_A - \Psi_A)^t \Sigma_{WAA} (\hat{V}_A - \Psi_A) \\
&\quad + \Psi_A^t \Sigma_{WAA} (\hat{V}_A - \Psi_A) + (\hat{V}_A - \Psi_A)^t (W_{AA} - \Sigma_{WAA}) \Psi_A \\
&\quad + \Psi_A^t (W_{AA} - \Sigma_{WAA}) \Psi + (\hat{V}_A - \Psi_A)^t \Sigma_{WAA} \Psi_A.
\end{aligned}
$$

Moreover,

$$
\|\hat{\Lambda}^{-1} - \Lambda^{-1}\|_2 \leq \|\hat{\Lambda}^{-1}\|_2 \|\hat{\Lambda} - \Lambda\|_2 \|\Lambda^{-1}\|_2 \leq (\|\Lambda^{-1}\|_2 + \|\hat{\Lambda}^{-1} - \Lambda^{-1}\|_2) \|\Upsilon\|_2 \|\Lambda^{-1}\|_2.
$$

By triangle inequality

$$
\begin{aligned}
\|\Upsilon\|_2 &\leq \|(\hat{V}_A - \Psi_A)^t (W_{AA} - \Sigma_{WAA})(\hat{V}_A - \Psi_A)\|_2 + 2\|\Psi_A^t (W_{AA} - \Sigma_{WAA})(\hat{V}_A - \Psi_A)\|_2 \\
&\quad + \|(\hat{V}_A - \Psi_A)^t \Sigma_{WAA} (\hat{V}_A - \Psi_A)\|_2 + 2\|\Psi_A^t \Sigma_{WAA} (\hat{V}_A - \Psi_A)\|_2 + \|\Psi_A^t (W_{AA} - \Sigma_{WAA}) \Psi_A\|_2.
\end{aligned}
$$

Since $\|A\|_2 \leq \|A\|_{\infty,2}$ and $\|AB\|_{\infty,2} \leq \|A\|_\infty \|B\|_{\infty,2}$, it follows that on $\mathcal{E}$

$$\begin{aligned}
\|\Upsilon\|_2 &\leq \|\hat{V}_A - \Psi_A\|_{\infty,2}^2 \|W_{AA} - \Sigma_{WAA}\|_\infty + 2\|\Psi_A\|_{\infty,2}\|\hat{V}_A - \Psi_A\|_{\infty,2}\|W_{AA} - \Sigma_{WAA}\|_\infty \\
&\quad + \|\hat{V}_A - \Psi_A\|_{\infty,2}^2 \|\Sigma_{WAA}\|_\infty + 2\|\Psi_A\|_{\infty,2}\|\Sigma_{WAA}\|_\infty \|\hat{V}_A - \Psi_A\|_{\infty,2} \\
&\quad + \|\Psi_A\|_{\infty,2}^2 \|W_{AA} - \Sigma_{WAA}\|_\infty \\
&\leq \|W_{AA} - \Sigma_{WAA}\|_\infty \left(\|\hat{V}_A - \Psi_A\|_{\infty,2}^2 + 2\|\Psi_A\|_{\infty,2}\|\hat{V}_A - \Psi_A\|_{\infty,2} + \|\Psi_A\|_{\infty,2}^2 \right) \\
&\quad + 2\|\Psi_A\|_{\infty,2}\|\Sigma_{WAA}\|_\infty \|\hat{V}_A - \Psi_A\|_{\infty,2}.
\end{aligned}$$

Therefore, there exists constant C'' such that on the event $\mathcal{E}$

$$\|\Upsilon\|_2 \leq C'' \sqrt{\frac{\log(ps)s^2}{N}}.$$

If $\|\Upsilon\|_2 \|\Lambda^{-1}\|_2 < 1$, then

$$\|\hat{\Lambda}^{-1} - \Lambda^{-1}\|_2 \leq \|\Lambda^{-1}\|_2^2 \|\Upsilon\|_2^2 (1 - \|\Upsilon\|_2 \|\Lambda^{-1}\|_2)^{-1}.$$

It follows that on the event $\mathcal{E}$, for each $g \in \{1, ..., G\}$ there exists constant C such that

$$|h^g - \hat{h}^g| \leq C \sqrt{\frac{\log(ps)s^2}{N}}.$$

Under (C1) and (C2), $P(\mathcal{E}_1) \to 1$ by Hoeffding inequality, $P(\mathcal{E}_2) \to 1$ from the first part of Corollary 2, $P(\mathcal{E}_3) \to 1$ and $P(\mathcal{E}_4) \to 1$ from the proofs of Lemmas 2 and 3, $P(\mathcal{E}_5) \to 1$ from Corollary 1. Therefore, $P(\mathcal{E}) \to 1$, hence

$$\hat{h}^g \xrightarrow{P} h^g.$$

By the Continuous Mapping Theorem [60, Theorem 2.3], this implies that for any g_1, g_2

$$\hat{h}^{g_1} - \hat{h}^{g_2} \xrightarrow{P} h^{g_1} - h^{g_2}.$$

Without loss of generality, assume $h_\Psi(x) = 1$. In other words, $h^1 < h^g$ for all $g \in \{2, ..., G\}$. Then

$$\begin{aligned} P\left(\hat{h}_{\hat{V}}(x) = h_\Psi(x)\right) &= P\left(\hat{h}^1 \leq \hat{h}^g \text{ for all } g \neq 1\right) \\ &= P\left(h_1 - h_g + \left(\hat{h}^1 - \hat{h}^g - (h^1 - h^g)\right) \leq 0 \text{ for all } g \neq 1\right) \\ &= P\left(\left(\hat{h}^1 - \hat{h}^g - (h^1 - h^g)\right) \leq h^g - h^1 \text{ for all } g \neq 1\right). \end{aligned}$$

Since $\hat{h}^1 - \hat{h}^g \xrightarrow{P} h^1 - h^g$, it follows that $P\left(\hat{h}_{\hat{V}}(x) = h_\Psi(x)\right) \to 1$.

□

3.8.5 Extension of Theorem 1 to general $\pi_i > 0$ and n_i

Given that $n_i \sim Bin(N, \pi_i)$ for all $i = 1, ..., G$, by Hoeffding inequality

$$P\left(\left|\pi_i - \frac{n_i}{N}\right| \geq \epsilon\right) \leq 2\exp(-2N\epsilon^2).$$

Hence

$$P\left(\left|\pi_i - \frac{n_i}{N}\right| \geq \frac{\pi_i}{2}\right) \leq 2\exp\left(-\frac{N\pi_i^2}{2}\right).$$

Consider event

$$\mathbf{A}_N = \bigcap_{i=1}^{G} \left\{\frac{\pi_i N}{2} \leq n_i \leq \frac{3\pi_i N}{2}\right\}.$$

From above it follows that

$$P(\mathbf{A}_N) \geq 1 - 2G\exp\left(-\frac{N\min_i \pi_i^2}{2}\right).$$

The results of auxillary lemmas can be obtained by conditioning on Y and considering the probabilities on the event $\mathbf{A}_N$ and its compliment. Since $P(\mathbf{A}_N)$ doesn't depend on p, s or ϵ, this conditioning does not affect the final rates.

3.8.6 Extension of Theorem 1 to sub-Gaussian sase

The normality assumption is employed to use the following concentration inequalities:

1. For any independent $X_1, ..., X_N \sim N(\mu, \sigma^2)$

$$P\left(|\bar{X} - \mu| \geq \epsilon\right) \leq 2\exp(-cN\epsilon^2).$$

2. Lemma A.3 in [8]. For any independent $X_1, ..., X_N \sim N_p(\mu, \Sigma)$ and $\epsilon \leq \epsilon_0$

$$P\left(\left|\frac{1}{N}\sum_{k=1}^{N}(X_{ki} - \mu_i)(X_{kj} - \mu_g) - \Sigma_{ij}\right| \geq \epsilon\right) \leq C_1\exp(-C_2N\epsilon^2).$$

Note that the first inequality remains true with the relaxation $X_1, ..., X_N$ are sub-Gaussian (see for example [62] and references therein). The second inequality remains true for sub-Gaussian random vectors $X_1, ..., X_N$ as long as

$$P\left(\left|\frac{1}{N}X^tY - \mathbb{E}\left(\frac{1}{N}X^tY\right)\right| > \epsilon\right) \leq C_1\exp(-C_2N\epsilon^2) \tag{3.13}$$

holds for zero-mean sub-Gaussian X and Y with independent individual components and some small $\epsilon < \epsilon_0$. If $X = Y$, then (3.13) is a special case of the Hanson-Wright inequality [48]. Therefore the generalization of Hanson-Wright inequality for the case $X \neq Y$ is needed. If such a generalization can be obtained, then the results of Theorem 1 remain true with the relaxation $(X_i|Y_i = g)$ are sub-Gaussian.

CHAPTER 4

OPTIMAL VARIABLE SELECTION IN MULTI-GROUP SPARSE DISCRIMINANT ANALYSIS

joint with Mladen Kolar

4.1 Introduction

The focus of this chapter is on establishing optimal conditions under which the Multi-Group Sparse Discriminant Analysis (MGSDA) procedure, described in Chapter 3, consistently recovers the relevant variables for classification. Consistent variable selection is an important property, since many domain scientist use the selected variables for hypothesis generation, downstream analysis and scientific discovery. In Chapter 3 we established the equivalence between MGSDA and sparse discriminant analysis [37] in the two group case and then extended the proof technique of [37] to the multi-group case. This strategy, however, does not lead to optimal sample size scaling for consistent variable selection in the two group case [32]. In this chapter, we use a refined proof strategy that allows us to establish consistent variable selection in the multi-group (with $G = \mathcal{O}(1)$) case under the same sample size scaling as in the two-group case. In particular, we establish that the sample size n needs to satisfy

$$n \geq K \|\Sigma_{AA}^{-1}\|_2 \left(\max_{j \in A^c} \sigma_{jj \cdot A} \right) (G-1) s \log((p-s)\log(n))$$

in order for MGSDA to recover the correct variables. Here K is a fixed constant independent from n, p, s and G, and $\sigma_{jj \cdot A} = \Sigma_{jj} - \Sigma_{jA}\Sigma_{AA}^{-1}\Sigma_{Aj}$. At a high-level, we will follow the primal-dual strategy used in [32], however, there are a

number of details that require careful dealing in order to establish the desired scaling. In particular, in Chapter 3 it was shown that the solution to (1.4) is matrix $V = W^{-1}DR$, where R is a $(G-1)$-dimensional orthogonal matrix. Furthermore, at the optima $\{v_g\}_{g=1}^{G-1}$, the objective values in (1.4) are equal to the non-zero eigenvalues of $D^\top W^{-1} D$. However, in Chapter 3 we separately consider the deviations of W^{-1} and D from their population counterparts, which is not sufficient to establish the optimal scaling of (n, p, s) for consistent variable selection. In contrast, here we consider these quantities jointly. In the two-group case, $W^{-1}D$ is a vector and $D^\top W^{-1} D$ is a scalar, which allows [32] to use concentration inequalities for χ^2 distributed random variables to achieve the optimal rate. In the multi-group case, one needs to characterize the joint distribution of the columns of $W^{-1}D$ and the behavior of the $\|\!| D^\top W^{-1} D |\!\|_2$, hence an analysis different from [32] is required. In particular, we use the distributional results of [10] to characterize $W^{-1}D$ and the results from random matrix theory [62, 65] for $\|\!| D^\top W^{-1} D |\!\|_2$.

The rest of the chapter is organized as follows. In §4.2, we briefly review the MGSDA procedure. In §4.3, we study the population version of the MGSDA estimator. Our main result is stated in §4.4. Illustrative simulation studies, which corroborate our theoretical findings, are provided in §4.5. Technical proofs are given in §4.7.

4.2 Preliminaries

The MGSDA estimator is found as the solution to the following convex optimization problem

$$\hat{V} = \arg \min_{V \in \mathbb{R}^{p \times (G-1)}} \left\{ \frac{1}{2} \operatorname{Tr}(V^\top W V) + \frac{1}{2} \|D^\top V - I\|_F^2 + \lambda \sum_{i=1}^{p} \|v_i\|_2 \right\}, \qquad (4.1)$$

where W and D are defined in Chapter 3. The sparsity of the estimated canonical vectors $\hat{V}$ is controlled by the user specified parameter $\lambda > 0$. Note that the ℓ_2-norm penalty encourages the rows of $\hat{V}$ to be sparse leading to the variable selection. When $\lambda = 0$ and W is nonsingular, $\hat{V} = (W + DD^\top)^{-1} D$ spans the $(G-1)$-dimensional eigenspace of $W^{-1} DD^t$. Since the classification rule (1.5) is invariant with respect to linear transformations, the MGSDA coincides with classical sample canonical correlation analysis. Intuitively, the three components of the objective function in (4.1) minimize the within-class variability, control the level of between-class variability and provide regularization by inducing sparsity respectively.

In the next two sections, we study conditions under which the MGSDA consistently recovers the correct set of discriminant variables.

4.3 Variable Selection in the Population Setting

In this section, we develop understanding of the MGSDA in the limit of infinite amount of data. We will develop understanding of limitations of the procedure for the purpose of consistent variable selection.

Let π_g be the prior group probabilities, $P(Y_i = g) = \pi_g$. Let μ_g be the pop-

ulation within-group mean, $\mu_g = \mathbb{E}(X_i \mid Y_i = g)$. Let Σ be the population within-group covariance matrix, $\mathrm{Cov}(X_i \mid Y_i = g) = \Sigma$, and $\Delta \in \mathbb{R}^{p\times(G-1)}$ be the matrix of population mean contrasts between G groups with rth column

$$\Delta_r = \frac{\sqrt{\pi_{r+1}}\sum_{g=1}^{r}\pi_g(\mu_g - \mu_{r+1})}{\sqrt{\sum_{g=1}^{r}\pi_g \sum_{g=1}^{r+1}\pi_g}}.$$

The population canonical vectors are eigenvectors of matrix $\Sigma^{-1}\Delta\Delta^\top$. The column vectors of matrix $\Psi = \Sigma^{-1}\Delta$ define the $(G-1)$-dimensional eigenspace of $\Sigma^{-1}\Delta\Delta^t$. Since the canonical vectors determine the variables that are relevant for the classification rule, in the high-dimensional setting we assume that the matrix Ψ is row sparse. Let A be the support of Ψ, $A = \{i \mid \|\Psi_i\|_2 \neq 0\}$, and s be the cardinality of A, $s = |A|$.

The population version of MGSDA optimization problem is

$$\hat{\Psi} = \arg\min_{V\in\mathbb{R}^{p\times(G-1)}} \left\{ \frac{1}{2}\mathrm{Tr}(V^\top\Sigma V) + \frac{1}{2}\|V^\top\Delta - I\|_F^2 + \lambda\sum_{i=1}^{p}\|v_i\|_2 \right\}. \tag{4.2}$$

Compared to the optimization program in (4.1), in (4.2) we assume access to the population covariance Σ and mean contrasts Δ. Theorem 1 characterizes conditions under which $\hat{\Psi} = (\hat{\Psi}_A^\top, 0_{p-s}^\top)^\top$ and $||e_j^\top\hat{\Psi}_A||_2 \neq 0$ for all $j \in A$.

Theorem 1. *Suppose that*

$$\|\Sigma_{A^cA}\Sigma_{AA}^{-1}s_A\|_{\infty,2} < 1 \tag{4.3}$$

and the tuning parameter λ in (4.2) *satisfies*

$$\lambda < \frac{\Psi_{\min}}{\|(\Sigma_{AA} + \Delta_A\Delta_A^\top)^{-1}\|_\infty \left(1 + \|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A\|_2\right)}, \tag{4.4}$$

where $\Psi_{\min} = \min_{j\in A}\|e_j^\top\Psi_A\|_2 = \min_{j\in A}\|e_j^\top\Sigma^{-1}\Delta\|_2$. Then the solution $\hat{\Psi}$ to (4.2) *is of the form $\hat{\Psi} = (\hat{\Psi}_A^\top, 0_{p-s}^\top)^\top$, where*

$$\hat{\Psi}_A = \Psi_A(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1} - \lambda(\Sigma_{AA} + \Delta_A\Delta_A^\top)^{-1}s_A, \tag{4.5}$$

and s_A is the sub-gradient of $\sum_{i\in A}\|\hat{\psi}_i\|_2$. Furthermore, we have that $||e_j^\top \hat{\Psi}_A||_2 \neq 0$ for all $j \in A$.

Theorem 1 provides sufficient conditions (4.3) and (4.4) under which the solution to (4.2) recovers the true support A. The condition (4.3) is of the same form as the irrepresentable condition in a multi-task regression [42]. The condition (4.4) relates the tuning parameter λ and the minimal signal strength $\Psi_{\min}$. The tuning parameter λ should not be too large, so that the relevant variables in A are not shrank to zero. The upper bound depends on the minimal signal strength $\Psi_{\min}$ and the classification difficulty characterized by $\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2$. Note that $\|(\Sigma_{AA}+\Delta_A\Delta_A^\top)^{-1}\|_\infty \leq \sqrt{s}\|(\Sigma_{AA}+\Delta_A\Delta_A^\top)^{-1}\|_2$, therefore it is sufficient for λ to satisfy

$$\lambda < \frac{\Psi_{\min}}{\sqrt{s}\|(\Sigma_{AA}+\Delta_A\Delta_A^\top)^{-1}\|_2\left(1+\|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A\|_2\right)}. \tag{4.6}$$

Equation (4.5) provides an explicit form for the solution $\hat{\Psi}$. Note that it estimates Ψ_A up to the linear transformation $(I+\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1}$ and the bias term due to the penalty. The linear transformation has no effect on the support or the classification assignment due to invariance of classification rule (1.5). The bias term has no effect on the support as long as λ satisfies (4.4). Note that Theorem 1 of [32] is a special case of our result in the two-group case.

4.4 Consistent Variable Selection of MGSDA

In this section, we establish our main result on the sample complexity needed for the variable selection consistency of the MGSDA.

We require the following assumptions.

(C1) Irrepresentability. There exists a constant $\alpha \in (0, 1]$ such that

$$\|\Sigma_{A^cA}\Sigma_{AA}^{-1}s_A\|_{\infty,2} \leq 1 - \alpha.$$

(C2) Minimal signal strength. There exists a constant $K_\psi > 0$ such that

$$\begin{aligned}\Psi_{\min} &= \min_{j\in A} ||e_j^\top \Psi_A|| \\ &\geq \lambda\sqrt{s}\|(\Sigma_{AA} + \Delta_A\Delta_A^\top)^{-1}\|_2 \times \\ &\times \left(1 + K_\psi \left[\|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A\|_2 \vee 1\right] \left(1 + \sqrt{\max_{j\in A}(\Sigma_{AA}^{-1})_{jj}\frac{(G-1)\log(s\log(n))}{n}}\right)\right).\end{aligned}$$

Irrepresentable condition is commonly used in the high-dimensional literature as a way to ensure exact variable selection of lasso like procedures [75, 65, 42, 32]. The second condition is commonly known as a beta-min condition and it states that the relevant variables should have sufficiently large signal in order for the procedure to distinguish them from noise.

Let $\hat{A}$ be the support of $\hat{V}$ defined in (4.1), $\hat{A} = \{i : \|\hat{v}_i\|_2 \neq 0\}$.

Theorem 2. *Assume that the conditions 4.4 and 4.4 are satisfied. Furthermore, suppose that the sample size satisfies*

$$n \geq K\left(\max_{j\in A^c}\sigma_{jj\cdot A}\right)\|\Sigma_{AA}^{-1}\|_2(G-1)s\log((p-s)\log(n))$$

for some absolute constant $K > 0$. If the tuning parameter λ is selected as

$$\lambda \geq K_\lambda(1 + \|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A\|_2)^{-1}\sqrt{\left(\max_{j\in A^c}\sigma_{jj\cdot A}\right)\frac{(G-1)\log((p-s)\log(n))}{n}},$$

where K_λ is an absolute constant that does not depend on the problem parameters, then the MGSDA procedure defined in (4.1) *satisfies*

$$\hat{A} = A,$$

with probability at least $1 - \mathcal{O}(\log^{-1}(n))$.

Theorem 2 is the finite sample version of Theorem 1. The main result states that the set of relevant variables will be recovered with high probability when the sample size n is of the order $\mathcal{O}(s\log(p))$ and the minimal signal strength is of the order $\mathcal{O}\left(\sqrt{n^{-1}s\log(p)}\right)$. The $\sqrt{s}$ term in the minimal signal strength condition comes from the substitutions of $\|\!|(\Sigma_{AA}+\Delta_A\Delta_A^\top)^{-1}|\!\|_\infty$ by $\|\!|(\Sigma_{AA}+\Delta_A\Delta_A^\top)^{-1}|\!\|_2$. Theorem 2 significantly improves on the result in Chapter 3 which requires n to be of the order $\mathcal{O}(s^2\log(ps))$ and $\Psi_{\min}$ to be of the order $\mathcal{O}\left(\sqrt{n^{-1}s^2\log(ps)}\right)$. These improvements are achieved through the joint characterization of the distribution of $W_{AA}^{-1}D_A$ and deviations of $\|\!|D_A^\top W_{AA}^{-1}D_A|\!\|_2$ from $\|\!|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A|\!\|_2$. When $G=2$, Theorem 2 reduces to the result established in [32] up to the condition on the tuning parameter λ. In [32] there is an additional factor $\sqrt{\left[\|\!|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A|\!\|_2\vee 1\right]}$, which we avoid due to the use of a different proof technique.

4.4.1 Outline of the proof

The proof of Theorem 2 is based on the primal-dual witness technique [65]. In the course of the proof, one proposes a solution $\hat{V}$ to (4.1) and verifies that the optimality conditions are satisfied.

We will verify that the vector $(\tilde{V}_A^\top, 0^\top)^\top$, where $\tilde{V}_A$ is the solution to the following oracle optimization program

$$\tilde{V}_A = \arg\min_{V\in\mathbb{R}^{s\times(G-1)}} \frac{1}{2}\operatorname{Tr}(V^\top W_{AA}V) + \frac{1}{2}\|D_A^\top V - I\|_F^2 + \lambda\sum_{i\in A}\|v_i\|_2,$$

satisfies the Karush-Kuhn-Tucker conditions for (4.1). The next lemma characterizes the form of the oracle solution $\tilde{V}_A$.

Lemma 5. *The oracle solution satisfies*

$$\tilde{V}_A = W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1} - \lambda (W_{AA} + D_A D_A^\top)^{-1} s_A,$$

where s_A is sub-gradient of $\sum_{i \in A} \|\tilde{v}_i\|_2$.

Lemma 6 provides the sufficient conditions for the estimator $(\tilde{V}_A^\top, 0^\top)^\top$ to be the oracle solution.

Lemma 6. *If*

$$\|\!|(W_{A^c A} + D_{A^c} D_A^\top) \tilde{V}_A - D_{A^c}|\!\|_{\infty,2} \leq \lambda; \tag{4.7}$$

$$\min_{j \in A} \|e_j^\top W_{AA}^{-1} D_A\|_2 > \lambda \|\!|(W_{AA} + D_A D_A^\top)^{-1}|\!\|_\infty (1 + \|\!|D_A^\top W_{AA}^{-1} D_A|\!\|_2), \tag{4.8}$$

then $\hat{V} = (\tilde{V}_A^\top, 0^\top)^\top$ and $\|e_j^\top \tilde{V}_A\|_2 \neq 0$ for all $j \in A$.

Lemma 6 is deterministic in nature. We proceed to show that (4.7) and (4.8) are satisfied with high probability under conditions of Theorem 2. In particular, next theorem established that the correct variables j, $j \in A$, are estimated as nonzero by $\tilde{V}_A$

Theorem 3. *Under conditions of Theorem 2, with probability at least* $1 - \mathcal{O}(\log^{-1}(n))$

$$\min_{j \in A} \|e_j^\top W_{AA}^{-1} D_A\|_2 > \lambda \|\!|(W_{AA} + D_A D_A^\top)^{-1}|\!\|_\infty (1 + \|\!|D_A^\top W_{AA}^{-1} D_A|\!\|_2).$$

To complete the proof, in the following theorem we establish that the wrong variables j, $j \in A^c$, are zero in $\hat{V}$.

Theorem 4. *Under conditions of Theorem 2, with probability at least* $1 - \mathcal{O}(\log^{-1}(n))$

$$\|\!|(W_{A^c A} + D_{A^c} D_A^\top) \tilde{V}_A - D_{A^c}|\!\|_{\infty,2} \leq \lambda.$$

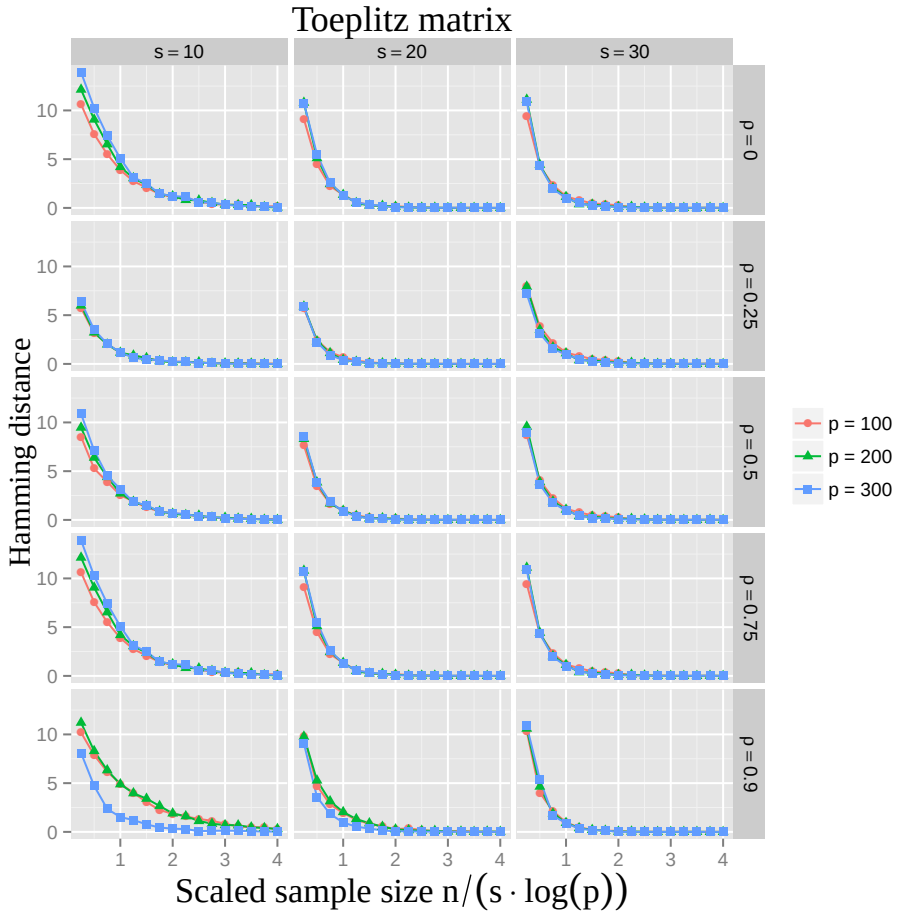


Figure 4.1: Performance of the MGSDA estimator averaged over 100 simulation runs. Plots of the rescaled sample size $n/(s\log(p))$ versus the Hamming distance between $\hat{A}$ and A for the Toeplitz matrix (see main text for details). Columns correspond to the size of A, $s \in \{10, 20, 30\}$, and rows correspond to different correlation strengths $\rho \in \{0, 0.25, 0.5, 0.75, 0.9\}$. Each subfigure shows three curves, corresponding to the problem sizes $p \in \{100, 200, 300\}$.

4.5 Simulation Results

We conduct several simulations to numerically illustrate finite sample properties of the MGSDA for the task of variable selection. The number of groups $G = 3$ and we change the size of the set A, $s \in \{10, 20, 30\}$, and the ambient dimension $p \in \{100, 200, 300\}$. The sample size is set as $n = \theta s \log(p)$ where θ is a control parameter that is varied. We report how well the MGSDA estimator

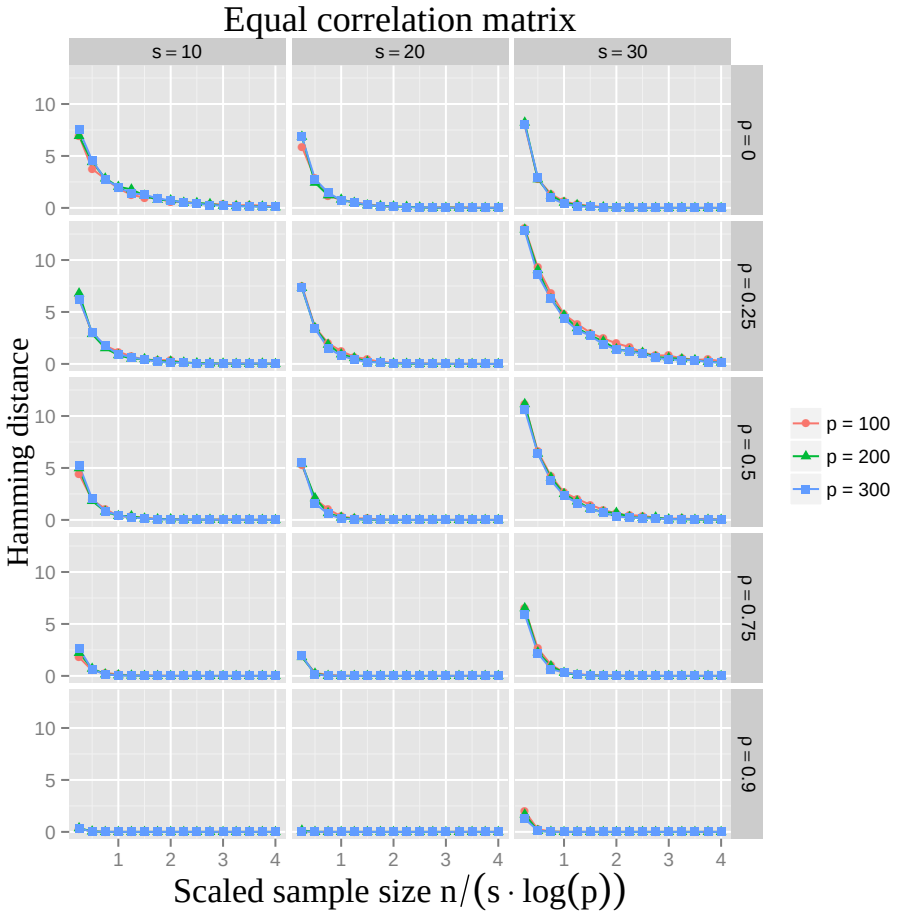


Figure 4.2: Performance of the MGSDA estimator averaged over 100 simulation runs. Plots of the rescaled sample size $n/(s\log(p))$ versus the Hamming distance between $\hat{A}$ and A for equal correlation matrix (see main text for details). Columns correspond to the size of A, $s \in \{10, 20, 30\}$, and rows correspond to different correlation strengths $\rho \in \{0, 0.25, 0.5, 0.75, 0.9\}$. Each subfigure shows three curves, corresponding to the problem sizes $p \in \{100, 200, 300\}$.

recovers the set of variables A as the control parameter θ varies. According to Theorem 2, the MGSDA recovers the correct variables when $n = Ks\log(p)$ for some $K > 0$ and this will be illustrated in our simulations.

Next, we describe the data generating model. We set $\mathbb{P}(Y = g) = \frac{1}{3}$ for

$g \in \{1, 2, 3\}$ and $X \mid Y = g \sim \mathcal{N}(\mu_g, \Sigma)$ with

$$\mu_1 = 0, \quad \mu_2 = (\underbrace{1, \ldots, 1}_{s}, \underbrace{0, \ldots, 0}_{p-s})^\top \quad \text{and} \quad \mu_3 = (\underbrace{1, \ldots, 1}_{s/2}, \underbrace{-1, \ldots, -1}_{s/2}, \underbrace{0, \ldots, 0}_{p-s})^\top.$$

We specify the covariance matrix Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{AA} & 0_{s \times p-s} \\ 0_{p-s \times s} & I_{p-s} \end{pmatrix}$$

and consider two cases for the component Σ_{AA}:

1. Toeplitz matrix, where $\Sigma_{TT} = [\Sigma_{ab}]_{a,b \in T}$ and $\Sigma_{ab} = \rho^{|a-b|}$ with $\rho \in \{0, 0.25, 0.5, 0.75, 0.9\}$, and
2. equal correlation matrix, where $\Sigma_{ab} = \rho$ when $a \neq b$ and $\sigma_{aa} = 1$, $\rho \in \{0, 0.25, 0.5, 0.75, 0.9\}$.

Finally, we set the penalty parameter as

$$\lambda = 0.5 \times \left(1 + \|\Delta_A^\top \Sigma_{AA}^{-1} \Delta_A^\top\|_2\right)^{-1} \sqrt{\frac{\log(p-s)}{n}}$$

for all cases, as suggested by Theorem 2. For each setting, we report the Hamming distance between the estimated set $\hat{A}$ and the true set A averaged over 200 independent simulation runs.

Figure 4.1 and Figure 4.2 illustrate finite sample performance of the MGSDA procedure. The Hamming distance is plotted against the control parameter θ, which represents the rescaled number of samples. Each figure contains a number of subfigures, which correspond to different simulation settings. Columns correspond to different number of relevant variables, $|A| = s \in \{10, 20, 30\}$, and rows correspond to different values of ρ, $\rho \in \{0, 0.25, 0.5, 0.75, 0.9\}$. Each subfigure contains three curves for different problem sizes $p \in \{100, 200, 300\}$. We

observe that as the control parameter θ increases the MGSDA procedure starts to recover the true set of variables, A, irrespective of the problem size, therefore, illustrating that our theoretical results describe well the finite sample performance of the procedure.

4.6 Discussion

In this chapter we consider the problem of variable selection in discriminant analysis. This is the first time that the consistent variable selection in the multi-class setting has been established under the same conditions as in the two-class setting. Throughout the chapter we have assumed that the number of classes G does not increase with the sample size n, however this condition is not necessary for consistent variable selection and is used for the simplicity of exposition. We hope to address this issue in future work.

4.7 Technical Proofs

4.7.1 Proof of Theorem 1

Using the Karush-Kuhn-Tucker conditions, we have that any solution $\hat{\Psi}$ of (4.2) satisfies

$$(\Sigma_{AA} + \Delta_A \Delta_A^\top)\hat{\Psi}_A + (\Sigma_{AA^c} + \Delta_A \Delta_{A^c}^\top)\hat{\Psi}_{A^c} - \Delta_A = -\lambda s_A; \tag{4.9}$$

$$(\Sigma_{A^cA} + \Delta_{A^c} \Delta_A^\top)\hat{\Psi}_A + (\Sigma_{A^cA^c} + \Delta_{A^c} \Delta_{A^c}^\top)\hat{\Psi}_{A^c} - \Delta_{A^c} = -\lambda s_{A^c}. \tag{4.10}$$

We proceed to verify that these conditions are satisfied by $\hat{\Psi} = (\hat{\Psi}_A^\top, 0^\top)^\top$ where $\hat{\Psi}_A^\top$ is given in (4.5). It is immediately clear that (4.9) is satisfied. We proceed to show that (4.10) is also satisfied. In particular, we show that

$$\|\!|(\Sigma_{A^cA} + \Delta_{A^c}\Delta_A^\top)\hat{\Psi}_A - \Delta_{A^c}\|\!|_{\infty,2} < \lambda.$$

Since $\Sigma\Sigma^{-1}\Delta = \Delta$, it follows that $\Sigma_{A^cA}\Sigma_{AA}^{-1}\Delta_A = \Delta_{A^c}$. Therefore,

$$\begin{aligned}
&\left(\Sigma_{A^cA} + \Delta_{A^c}\Delta_A^\top\right)\hat{\Psi}_A \\
&= (\Sigma_{A^cA} + \Delta_{A^c}\Delta_A^\top)(\Psi_A(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1} - \lambda(\Sigma_{AA} + \Delta_A\Delta_A^\top)^{-1}s_A) \\
&= \Sigma_{A^cA}\Sigma_{AA}^{-1}\Delta_A(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1} + \Delta_{A^c}\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1} \\
&\quad - \lambda\Sigma_{A^cA}(\Sigma_A + \Delta_A\Delta_A^\top)^{-1}s_A - \lambda\Delta_{A^c}\Delta_A^\top(\Sigma_A + \Delta_A\Delta_A^\top)^{-1}s_A \\
&= \Delta_{A^c}(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1} + \Delta_{A^c}(I - (I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1}) \\
&\quad - \lambda\Sigma_{A^cA}(\Sigma_{AA}^{-1} - \Sigma_{AA}^{-1}\Delta_A(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1}\Delta_A^\top\Sigma_{AA}^{-1})s_A \\
&\quad - \lambda\Delta_{A^c}\Delta_A^\top(\Sigma_{AA}^{-1} - \Sigma_{AA}^{-1}\Delta_A(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1}\Delta_A^\top\Sigma_{AA}^{-1})s_A \\
&= \Delta_{A^c} - \lambda\Sigma_{A^cA}\Sigma_{AA}^{-1}s_A + \lambda\Delta_{A^c}(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1}\Delta_A^\top\Sigma_{AA}^{-1}s_A \\
&\quad - \lambda\Delta_{A^c}\Delta_A^\top\Sigma_{AA}^{-1}s_A + \lambda\Delta_{A^c}\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1}\Delta_A^\top\Sigma_{AA}^{-1}s_A \\
&= \Delta_{A^c} - \lambda\Sigma_{A^cA}\Sigma_{AA}^{-1}s_A + \lambda\Delta_{A^c}\Delta_A^\top\Sigma_{AA}^{-1}s_A - \lambda\Delta_{A^c}\Delta_A^\top\Sigma_{AA}^{-1}s_A \\
&= \Delta_{A^c} - \lambda\Sigma_{A^cA}\Sigma_{AA}^{-1}s_A.
\end{aligned}$$

By assumption (4.3),

$$\|\!|(\Sigma_{A^cA} + \Delta_{A^c}\Delta_A^\top)\hat{\Psi}_A - \Delta_{A^c}\|\!|_{\infty,2} = \lambda\|\!|\Sigma_{A^cA}\Sigma_{AA}^{-1}s_A\|\!|_{\infty,2} < \lambda,$$

which verifies that $\hat{\Psi}$ also satisfies (4.10) .

To complete the proof, we show that no component of $\hat{\Psi}_A$ is set to zero. From (4.5),

$$e_j^\top\hat{V}_A = e_j^\top\Psi_A(I + \Delta_A^\top\Sigma_{AA}^{-1}\Delta_A)^{-1} - \lambda e_j^\top(\Sigma_{AA} + \Delta_A\Delta_A^\top)^{-1}s_A.$$

Since

$$\|e_j^\top \Psi_A (I + \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A)^{-1}\|_2 \geq \frac{1}{\|\!| I + \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A |\!\|_2} \|e_j^\top \Psi_A\|_2 \geq \frac{\Psi_{\min}}{1 + \|\!| \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A |\!\|_2}$$

and

$$\|\lambda e_j^\top (\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1} s_A\|_2 \leq \lambda \|\!| (\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1} |\!\|_\infty,$$

the result follows.

Proof of Lemma 5 and 6. The proof follows the proof of Theorem 1. □

Proof of Theorem 3. From Lemma 10, with probability at least $1 - \mathcal{O}(\log^{-1}(n))$

$$\|\!| (W_{AA} + D_A D_A^\top)^{-1} |\!\|_\infty \leq \sqrt{s} \|\!| (\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1} |\!\|_2 \left(1 + \mathcal{O}\left(\sqrt{\frac{s \log(\log(n))}{n}}\right)\right).$$

From Lemma 13, with probability at least $1 - \mathcal{O}(\log^{-1}(n))$

$$\begin{aligned}\|\!| D_A^\top W_{AA}^{-1} D_A |\!\|_2 &\leq C \|\!| \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A |\!\|_2 \\ &\quad + \mathcal{O}\left(\frac{(G-1) s \log(\log(n))}{n} \vee \sqrt{\|\!| \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A |\!\|_2 \frac{(G-1)\log(\log(n))}{n}}\right).\end{aligned}$$

Therefore, with probability at least $1 - \mathcal{O}(\log^{-1}(n))$

$$\begin{aligned}&\lambda \|\!| (W_{AA} + D_A D_A^\top)^{-1} |\!\|_\infty (1 + \|\!| D_A^\top W_{AA}^{-1} D_A |\!\|_2) \\ &\quad \leq \lambda \sqrt{s} \|\!| (\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1} |\!\|_2 \left(1 + C\left[\|\!| \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A |\!\|_2 \vee 1\right] \left(1 + \sqrt{\frac{(G-1)\log(\log(n))}{n}}\right)\right).\end{aligned}$$

On the other hand, from Lemma 7, with probability at least $1 - \mathcal{O}(\log^{-1}(n))$

$$\begin{aligned}&\min_{j \in A} ||e_j^\top W_{AA}^{-1} D_A||_2 \\ &\geq \min_{j \in A} \|e_j^\top \Sigma_{AA}^{-1} \Delta_A\|_2 \left(1 - \mathcal{O}\left(\sqrt{\left[\|\!| \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A |\!\|_2 \vee 1\right] \max_{j \in A} (\Sigma_{AA}^{-1})_{jj} \frac{(G-1)\log(s\log(n))}{n}}\right)\right) \\ &\geq \Psi_{\min} \left(1 - \mathcal{O}\left(\sqrt{\left[\|\!| \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A |\!\|_2 \vee 1\right] \max_{j \in A} (\Sigma_{AA}^{-1})_{jj} \frac{(G-1)\log(s\log(n))}{n}}\right)\right).\end{aligned}$$

The final result follows from the condition on the sample size n and 4.4. □

Lemma 7. *With probability at least* $1 - \log^{-1}(n)$, $\forall j \in A$

$$||e_j^\top W_{AA}^{-1} D_A||_2$$
$$\geq \|e_j^\top \Sigma_{AA}^{-1} \Delta_A\|_2 \left(1 - \mathcal{O}\left(\sqrt{\left[\|\Delta_A^\top \Sigma_{AA}^{-1} \Delta_A\|_2 \vee 1\right] (\Sigma_{AA}^{-1})_{jj} \frac{(G-1)\log(s\log(n))}{n}}\right)\right).$$

Proof of Lemma 7. By triangle inequality

$$\|e_j^\top W_{AA}^{-1} D_A - e_j^\top \Sigma_{AA}^{-1} \Delta_A\|_2 \leq \|e_j^\top W_{AA}^{-1} D_A - e_j^\top \Sigma_{AA}^{-1} D_A\|_2 + \|e_j^\top \Sigma_{AA}^{-1} D_A - e_j^\top \Sigma_{AA}^{-1} \Delta_A\|_2.$$

Consider the first term,

$$\|e_j^\top W_{AA}^{-1} D_A - e_j^\top \Sigma_{AA}^{-1} D_A\|_2$$
$$\leq e_j^\top W_{AA}^{-1} e_j \left\| \frac{D_A^\top W_{AA}^{-1} e_j}{e_j^\top W_{AA}^{-1} e_j} - \frac{D_A^\top \Sigma_{AA}^{-1} e_j}{e_j^\top \Sigma_{AA}^{-1} e_j} \right\|_2 + \|e_j^\top \Sigma_{AA}^{-1} D_A\|_2 \left| \frac{e_j^\top \Sigma_{AA}^{-1} e_j}{e_j^\top W_{AA}^{-1} e_j} - 1 \right|.$$

From [32, Lemma 14], $\forall j \in A$

$$\left| \frac{e_j^\top \Sigma_{AA}^{-1} e_j}{e_j^\top W_{AA}^{-1} e_j} - 1 \right| \leq C_2 \sqrt{\frac{\log(s\log(n))}{n}}$$

with probability at least $1 - (\log(n))^{-1}$. Further, using Lemma 9

$$\left\| \frac{D_A^\top W_{AA}^{-1} e_j}{e_j^\top W_{AA}^{-1} e_j} - \frac{D_A^\top \Sigma_{AA}^{-1} e_j}{e_j^\top \Sigma_{AA}^{-1} e_j} \right\|_2 = \|\hat{H}_{12}\hat{H}_{22}^{-1} - H_{12}H_{22}^{-1}\|_2 = \|\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_h\|_2,$$

where

$$\hat{H}_{12}\hat{H}_{22}^{-1} | D_A \sim t_{G-1}(d_H, \mu_H, \Gamma_H)$$

with degrees of freedom $d_H = n - s - G + 2$, mean $\mu_H = H_{12}H_{22}^{-1}$ and scale parameter $\Gamma_H = \frac{1}{d_H}(D_A^\top R D_A)/(e_j^\top \Sigma_{AA}^{-1} e_j)$ with $R = \Sigma_{AA}^{-1} - \frac{\Sigma_{AA}^{-1} e_j e_j^\top \Sigma_{AA}^{-1}}{e_j^\top \Sigma_{AA}^{-1} e_j}$. Hence,

$$\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_H = \frac{\Gamma_H^{1/2} y_H}{\sqrt{Z_H/d_H}} \quad \text{and} \quad \|\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_H\|_2^2 = \frac{y_H^\top \Gamma_H y_H}{Z_H/d_H},$$

where $y_H \sim \mathcal{N}(0, I_{G-1})$ and $z_H \sim \chi^2_{d_H}$ are independent. Therefore,

$$P\left(\|\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_H\|_2 \leq \sqrt{\frac{\epsilon_1}{\epsilon_2}}\right) = P\left(\|\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_H\|_2^2 \leq \frac{\epsilon_1}{\epsilon_2}\right)$$
$$= P\left(\frac{y_H^\top \Gamma_H y_H}{Z_H/d_H} \leq \frac{\epsilon_1}{\epsilon_2}\right) \geq P(y_H^\top \Gamma_H y_H \leq \epsilon_1, Z_H/d_H \geq \epsilon_2)$$
$$\geq P(y_H^\top \Gamma_H y_H \leq \epsilon_1) P(Z_H/d_H \geq \epsilon_2).$$

Since $Z_H \sim \chi^2_{d_H}$, by Lemma 1 in [34] for all $y \geq 0$

$$P(Z_H/d_H \geq 1 - y) \geq 1 - \exp\left(-d_H \frac{y^2}{4}\right).$$

Since $y_H \sim \mathcal{N}(0, I_{G-1})$, using Proposition 1.1 in [30]

$$P(y_H^\top \Gamma_H y_H \geq \mathrm{Tr}(\Gamma_H) + 2\sqrt{\mathrm{Tr}(\Gamma_H^2)t} + 2|||\Gamma_H|||_2 t) \leq \exp(-t).$$

Combining the above displays,

$$\|\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_H\|_2 \leq \sqrt{\frac{\mathrm{Tr}(\Gamma_H) + 2\sqrt{\mathrm{Tr}(\Gamma_H^2)t} + 2|||\Gamma_H|||_2 t}{1-y}}$$

with probability at least

$$(1-\exp(-t))(1-\exp(-d_H\frac{y^2}{4})) = 1-(\exp(-t)+\exp(-d_H y^2/4)-\exp(-t)\exp(-d_H y^2/4)).$$

Setting it to be $1 - \mathcal{O}(\log^{-1}(n))$ for all $j \in A$, we get $t = \log(s\log(n))$, $y = 2\sqrt{\frac{\log(s\log(n))}{n-s-G+2}}$ and

$$\|\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_H\|_2 \leq \sqrt{\frac{\mathrm{Tr}(\Gamma_H) + 2\sqrt{\mathrm{Tr}(\Gamma_H^2)\log(s\log(n))} + 2|||\Gamma_H|||_2 \log(s\log(n))}{1 - 2\sqrt{\frac{\log(s\log(n))}{n-s-G+2}}}}.$$

Since $\mathrm{Tr}(\Gamma_H) \leq (G-1)|||\Gamma_H|||_2$ and $\mathrm{Tr}(\Gamma_H^2) \leq (G-1)^2|||\Gamma_H|||_2^2$, the above display can be rewritten as

$$\|\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_H\|_2$$
$$\leq \sqrt{|||\Gamma_H|||_2((G-1) + 2(G-1)\sqrt{\log(s\log(n))} + 2\log(s\log(n))\left(1 + O\left(\sqrt{\frac{\log(s\log(n))}{n}}\right)\right)}.$$

Hence, there exists constant $C > 0$ such that with probability at least $1 - O(\log^{-1}(n))$

$$\|\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_H\|_2 \leq C\sqrt{|||\Gamma_H|||_2(G-1)\log(s\log(n))}.$$

Using the definition of R,

$$|||\Gamma_H|||_2 = \frac{1}{n-s-G-2}\frac{1}{(\Sigma_{AA}^{-1})_{jj}}|||D_A^\top R D_A|||_2 \leq \frac{1}{n-s-G-2}\frac{1}{(\Sigma_{AA}^{-1})_{jj}}|||D_A^\top \Sigma_{AA}^{-1} D_A|||_2.$$

Applying Lemma 12, with probability at least $1 - \log^{-1}(n)$

$$\|\Gamma_H\|_2 \leq C\frac{1}{n-s-G-2}\frac{1}{(\Sigma_{AA}^{-1})_{jj}}\left[\|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A\|_2 \vee \sqrt{\|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A\|_2}\right].$$

Therefore, with probability at least $1 - \log^{-1}(n)$

$$\|\hat{H}_{12}\hat{H}_{22}^{-1} - \mu_H\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\left[\|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A\|_2 \vee 1\right]}{(\Sigma_{AA}^{-1})_{jj}}\frac{(G-1)\log(s\log(n))}{n}}\right).$$

Consider $\|e_j^\top\Sigma_{AA}^{-1}D_A - e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_2$. From Lemma 8, $\Sigma_{AA}^{-1}D_A \sim \mathcal{N}(\Sigma_{AA}^{-1}\Delta_A, \frac{\Sigma_{AA}^{-1}}{n} \otimes I_{G-1})$. Hence,

$$\begin{aligned}P\left(\|e_j^\top\Sigma_{AA}^{-1}D_A - e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_2 \geq \epsilon\right) &\leq P\left(\sqrt{G-1}\|e_j^\top\Sigma_{AA}^{-1}D_A - e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_\infty \geq \epsilon\right)\\ &\leq 2(G-1)\exp\left(-\frac{n\epsilon^2}{2(\Sigma_{AA}^{-1})_{jj}(G-1)}\right).\end{aligned}$$

Let $\epsilon = \sqrt{2(\Sigma_{AA}^{-1})_{jj}(G-1)\frac{\log(2(G-1)s\log(n))}{n}}$. Then for all $j \in A$

$$\|e_j^\top\Sigma_{AA}^{-1}D_A - e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_2 \leq \sqrt{2(\Sigma_{AA}^{-1})_{jj}(G-1)\frac{\log(2(G-1)s\log(n))}{n}}$$

with probability at least $1 - \log^{-1}(n)$. Also,

$$\begin{aligned}\|e_j^\top\Sigma_{AA}^{-1}D_A\|_2 &\leq \|e_j^\top\Sigma_{AA}^{-1}D_A - e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_2 + \|e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_2\\ &\leq \|e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_2 + \sqrt{2(\Sigma_{AA}^{-1})_{jj}(G-1)\frac{\log(2(G-1)s\log(n))}{n}}.\end{aligned}$$

Combining the above displays, with probability at least $1 - (\log(n))^{-1}$, for all $j \in A$

$$\begin{aligned}&\|e_j^\top W_{AA}^{-1}D_A - e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_2\\ &\quad\leq C_1(\Sigma_{AA}^{-1})_{jj}\sqrt{\frac{\left[\|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A\|_2 \vee 1\right]}{(\Sigma_{AA}^{-1})_{jj}}\frac{(G-1)\log(s\log(n))}{n}}\\ &\qquad+ \|e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_2 C_2\sqrt{\frac{\log(s\log(n))}{n}} + C_3\sqrt{(\Sigma_{AA}^{-1})_{jj}(G-1)\frac{\log(s\log(n))}{n}}\\ &\quad\leq C\|e_j^\top\Sigma_{AA}^{-1}\Delta_A\|_2\sqrt{\left[\|\Delta_A^\top\Sigma_{AA}^{-1}\Delta_A\|_2 \vee 1\right](\Sigma_{AA}^{-1})_{jj}\frac{(G-1)\log(s\log(n))}{n}}.\end{aligned}$$

The final result follows form triangle inequality. □

Proof of Theorem 4. Since $(n-G)W \sim W_p(n-G,\Sigma)$, then $(n-G)W = UU^\top$, where $U \in \mathbb{R}^{p\times(n-G)}$ with columns $u_i \overset{iid}{\sim} \mathcal{N}(0,\Sigma)$. Let

$$E_D = D_{A^c} - \Sigma_{A^cA}\Sigma_{AA}^{-1}D_A;$$
$$E_U = U_{A^c} - \Sigma_{A^cA}\Sigma_{AA}^{-1}U_A \text{ with } (n-G)W_{A^cA} = U_{A^c}U_A^\top.$$

Then,

$$D_{A^c} = \Sigma_{A^cA}\Sigma_{AA}^{-1}D_A + E_D;$$
$$(n-G)W_{A^cA} = U_{A^c}U_A^\top = (\Sigma_{A^cA}\Sigma_{AA}^{-1}U_A + E_U)U_A^\top = \Sigma_{A^cA}\Sigma_{AA}^{-1}(n-G)W_{AA} + E_UU_A^\top.$$

and therefore

$$\begin{aligned}
&(W_{A^cA} + D_{A^c}D_A^\top)\tilde{V}_A - D_{A^c}\\
=&(\Sigma_{A^cA}\Sigma_{AA}^{-1}W_{AA} + (n-G)^{-1}E_UU_A^\top\\
&\quad + (\Sigma_{A^cA}\Sigma_{AA}^{-1}D_A + E_D)D_A^\top)(W_{AA}^{-1}D_A(I + D_A^\top W_{AA}^{-1}D_A)^{-1} - \lambda(W_{AA}+D_AD_A^\top)^{-1}s_A)\\
&\quad - \Sigma_{A^cA}\Sigma_{AA}^{-1}D_A - E_D\\
=&\Sigma_{A^cA}\Sigma_{AA}^{-1}(D_A(I + D_A^\top W_{AA}^{-1}D_A)^{-1} + D_AD_A^\top W_{AA}^{-1}D_A(I + D_A^\top W_{AA}^{-1}D_A)^{-1} - D_A)\\
&\quad + \Sigma_{A^cA}\Sigma_{AA}^{-1}(-\lambda W_{AA}(W_{AA}+D_AD_A^\top)^{-1}s_A - \lambda D_AD_A^\top(W_{AA}+D_AD_A^\top)^{-1}s_A)\\
&\quad + (n-G)^{-1}E_UU_A^\top(W_{AA}^{-1}D_A(I + D_A^\top W_{AA}^{-1}D_A)^{-1} - \lambda(W_{AA}+D_AD_A^\top)^{-1}s_A)\\
&\quad + E_D(D_A^\top W_{AA}^{-1}D_A(I + D_A^\top W_{AA}^{-1}D_A)^{-1} - \lambda D_A^\top(W_{AA}+D_AD_A^\top)^{-1}s_A - I)\\
=&-\lambda\Sigma_{A^cA}\Sigma_{AA}^{-1}s_A + (n-G)^{-1}E_UU_A^\top(W_{AA}+D_AD_A^\top)^{-1}(D_A - \lambda s_A)\\
&\quad - E_D(\lambda D_A^\top(W_{AA}+D_AD_A^\top)^{-1}s_A + (I + D_A^\top W_{AA}^{-1}D_A)^{-1})\\
=&-\lambda\Sigma_{A^cA}\Sigma_{AA}^{-1}s_A + (n-G)^{-1}E_UU_A^\top(W_{AA}+D_AD_A^\top)^{-1}(D_A - \lambda s_A)\\
&\quad - E_D(I + D_A^\top W_{AA}^{-1}D_A)^{-1}(\lambda D_A^\top W_{AA}^{-1}s_A + I)
\end{aligned}$$

We would like to establish the following:

$$\lambda \|\Sigma_{A^cA}\Sigma_{AA}^{-1}s_A\|_{\infty,2} < \lambda(1-\alpha) \tag{4.11}$$

$$\|(n-G)^{-1}E_U U_A^\top W_{AA}^{-1}D_A(I + D_A^\top W_{AA}^{-1}D_A)^{-1}\|_{\infty,2} < \lambda\alpha/4 \tag{4.12}$$

$$\lambda\|(n-G)^{-1}E_U U_A^\top W_{AA}^{-1}(I + W_{AA}^{-1}D_A D_A^\top)^{-1}s_A\|_{\infty,2} \leq \lambda\alpha/4 \tag{4.13}$$

$$\lambda\|E_D(I + D_A^\top W_{AA}^{-1}D_A)^{-1}D_A^\top W_{AA}^{-1}s_A\|_{\infty,2} \leq \lambda\alpha/4 \tag{4.14}$$

$$\|E_D(I + D_A^\top W_{AA}^{-1}D_A)^{-1}\|_{\infty,2} \leq \lambda\alpha/4. \tag{4.15}$$

1. Show $\|E_D(I + D_A^\top W_{AA}^{-1}D_A)^{-1}\|_{\infty,2} \leq \lambda\alpha/4$.

Consider $E_D = \Sigma_{A^cA}\Sigma_{AA}^{-1}D_A - D_{A^c}$. Since $\Sigma\Sigma^{-1}\Delta = \Delta$, it follows that $\Sigma_{A^cA}\Sigma_{AA}^{-1}\Delta_A = \Delta_{A^c}$. Hence $\mathbb{E}(D_{A^c}) = \Delta_{A^c} = \Sigma_{A^cA}\Sigma_{AA}^{-1}\Delta_A$. Therefore $\mathbb{E}(E_D) = 0$. Moreover,

$$\begin{aligned}
\operatorname{Cov}(E_D, D_A) &= \operatorname{Cov}(\Sigma_{A^cA}\Sigma_{AA}^{-1}D_A - D_{A^c}, D_A) = \operatorname{Cov}(\Sigma_{A^cA}\Sigma_{AA}^{-1}D_A, D_A) - \operatorname{Cov}(D_{A^c}, D_A) \\
&= \Sigma_{A^cA}\Sigma_{AA}^{-1}\operatorname{Cov}(D_A) - \operatorname{Cov}(D_{A^c}, D_A) \\
&= \Sigma_{A^cA}\Sigma_{AA}^{-1}\operatorname{Cov}(D_A) - \Sigma_{A^cA}\Sigma_{AA}^{-1}\operatorname{Cov}(D_A) \\
&= 0.
\end{aligned}$$

From Lemma 8, for all $j \in A^c$

$$e_j^\top E_D \sim \mathcal{N}\left(0, \frac{1}{n}\sigma_{jj\cdot A}I_{G-1}\right)$$

where $\sigma_{jj\cdot A} = \Sigma_{jj} - \Sigma_{jA}\Sigma_{AA}^{-1}\Sigma_{Aj}$ and $e_j^\top E_D$ is independent of D_A. Note that

$$\begin{aligned}
\|E_D(I + D_A^\top W_{AA}^{-1}D_A)^{-1}\|_{\infty,2} &= \max_{j\in A^c}\|e_j^\top E_D(I + D_A^\top W_{AA}^{-1}D_A)^{-1}\|_2 \\
&\leq \frac{\max_{j\in A^c}\|e_j^\top E_D\|_2}{1 + \sigma_{\min}(D_A^\top W_{AA}^{-1}D_A)} \\
&\leq \max_{j\in A^c}\|e_j^\top E_D\|_2
\end{aligned}$$

Using Proposition 1.1 in [30]

$$\bigcap_{j\in A^C}\left\{\frac{||e_j^\top E_D||_2^2}{\sigma_{jj\cdot A}} \leq \frac{(G-1)}{n} + 2\frac{\sqrt{(G-1)\log((p-s)\log(n))}}{n} + 2\frac{\log((p-s)\log(n))}{n}\right\}$$

with probability at least $1-\log^{-1}(n)$. Hence, with probability at least $1-\log^{-1}(n)$

$$\max_{j\in A^c}\frac{\|e_j^\top E_D\|_2^2}{\sigma_{jj\cdot A}}\leq\mathcal{O}\left(\frac{(G-1)\log((p-s)\log(n))}{n}\right),$$

or equivalently

$$\max_{j\in A^c}\|e_j^\top E_D\|_2\leq\mathcal{O}\left(\sqrt{\max_{j\in A^c}\sigma_{jj\cdot A}\frac{(G-1)\log((p-s)\log(n))}{n}}\right).$$

2. Show $\lambda\|\!|E_D(I+D_A^\top W_{AA}^{-1}D_A)^{-1}D_A^\top W_{AA}^{-1}s_A\|\!|_{\infty,2}\leq\lambda\alpha/4$.

Since $e_j^\top E_D\sim\mathcal{N}(0,n^{-1}\sigma_{jj\cdot A}I_{G-1})$, it follows that

$$e_j^\top E_D(I+D_A^\top W_{AA}^{-1}D_A)^{-1}D_A^\top W_{AA}^{-1}s_A\sim\mathcal{N}\left(0,\frac{\sigma_{jj\cdot A}}{n}s_A^\top W_{AA}^{-1}D_A(I+D_A^\top W_{AA}^{-1}D_A)^{-2}D_A^\top W_{AA}^{-1}s_A\right).$$

Following the above arguments, the following event has probability at least $1-\log^{-1}(n)$

$$\bigcap_{j\in A^C}\left\{\frac{||e_j^\top E_D(I+D_A^\top W_{AA}^{-1}D_A)^{-1}D_A^\top W_{AA}^{-1}s_AL^{-1/2}||_2^2}{\sigma_{jj\cdot A}}\leq\mathcal{O}\left(\frac{(G-1)\log((p-s)\log(n))}{n}\right)\right\},$$

where $L=s_A^\top W_{AA}^{-1}D_A(I+D_A^\top W_{AA}^{-1}D_A)^{-2}D_A^\top W_{AA}^{-1}s_A$. This implies that with probability at least $1-\log^{-1}(n)$

$$\max_{j\in A^c}\frac{||e_j^\top E_D(I+D_A^\top W_{AA}^{-1}D_A)^{-1}D_A^\top W_{AA}^{-1}s_A||_2^2}{\sigma_{jj\cdot A}}\leq\|\!|L|\!\|_2\mathcal{O}\left(\frac{(G-1)\log((p-s)\log(n))}{n}\right)$$

By triangle inequality

$$\begin{aligned}
\|\!|L|\!\|_2&=\|\!|s_A^\top W_{AA}^{-1}D_A(I+D_A^\top W_{AA}^{-1}D_A)^{-2}D_A^\top W_{AA}^{-1}s_A\|\!|_2\\
&\leq\|\!|s_A^\top W_{AA}^{-1}s_A\|\!|_2\|\!|(I+D_A^\top W_{AA}^{-1}D_A)^{-1}D_A^\top W_{AA}^{-1/2}\|\!|_2^2\\
&\leq\|\!|s_A^\top W_{AA}^{-1}s_A\|\!|_2\\
&\leq s\|\!|s_A\|\!|_{\infty,2}^2\|\!|W_{AA}^{-1}\|\!|_2\\
&\leq s\|\!|W_{AA}^{-1}\|\!|_2.
\end{aligned}$$

From Lemma 9 in [65], with probability at least $1 - \log^{-1}(n)$

$$\|\!|W_{AA}^{-1}|\!\|_2 \leq \|\!|\Sigma_{AA}^{-1}|\!\|_2 \left(1 + \mathcal{O}\left(\sqrt{\frac{s\log(\log(n))}{n}}\right)\right).$$

Combining the above displays, with probability at least $1 - \mathcal{O}(\log^{-1}(n))$

$$\max_{j \in A^c} \frac{||e_j^\top E_D (I + D_A^\top W_{AA}^{-1} D_A)^{-1} D_A^\top W_{AA}^{-1} s_A||_2^2}{\sigma_{jj \cdot A}} \leq \|\!|\Sigma_{AA}^{-1}|\!\|_2 \mathcal{O}\left(\frac{(G-1)s\log((p-s)\log(n))}{n}\right),$$

or equivalently

$$\max_{j \in A^c} ||e_j^\top E_D (I + D_A^\top W_{AA}^{-1} D_A)^{-1} D_A^\top W_{AA}^{-1} s_A||_2 \leq \mathcal{O}\left(\sqrt{\|\!|\Sigma_{AA}^{-1}|\!\|_2 \max_{j \in A^c} \sigma_{jj \cdot A} \frac{(G-1)s\log((p-s)\log(n))}{n}}\right).$$

3. Show $\|\!|(n-G)^{-1} E_U U_A^\top W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1}|\!\|_{\infty,2} < \lambda\alpha/4$.

By definition $E_U = U_{A^c} - \Sigma_{A^c A} \Sigma_{AA}^{-1} U_A$, hence

$$\mathrm{vec}(E_U) \sim \mathcal{N}(0, \Sigma_{A^c A^c \cdot A} \otimes I_{n-G})$$

and is independent of U_A. Therefore

$$(n-G)^{-1} e_j^\top E_U \sim \mathcal{N}\left(0, \frac{1}{(n-G)^2} \sigma_{jj \cdot A} I_{n-G}\right),$$

where $\sigma_{jj \cdot A} = \Sigma_{jj} - \Sigma_{jA} \Sigma_{AA}^{-1} \Sigma_{Aj}$. Conditional on X_A,

$$\frac{1}{n-G} e_j^\top E_U U_A^\top W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1} \sim \mathcal{N}\left(0, \frac{\sigma_{jj \cdot A}}{n-G} (I + D_A^\top W_{AA}^{-1} D_A)^{-1} D_A^\top W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1}\right).$$

Let $(I + D_A^\top W_{AA}^{-1} D_A)^{-1} D_A^\top W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1} = L$. Then by Proposition 1.1 in [30]

$$\bigcap_{j \in A^C} \left\{ \frac{||(n-G)^{-1} e_j^\top E_U U_A^\top W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1} L^{-1/2}||_2^2}{\sigma_{jj \cdot A}} \leq \frac{(G-1)}{n-G} + 2\frac{\sqrt{(G-1)\log((p-s)\log(n))}}{n-G} + 2\frac{\log((p-s)\log(n))}{n-G} \right\}$$

with probability at least $1 - \log^{-1}(n)$. Therefore,

$$\bigcap_{j \in A^C} \left\{ \frac{||(n-G)^{-1} e_j^\top E_U U_A^\top W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1}||_2^2}{\sigma_{jj \cdot A}} \right.$$
$$\left. \leq \|\!| L |\!\|_2 \mathcal{O}\left(\frac{(G-1)\log((p-s)\log(n))}{n-G} \right) \right\}$$

with probability at least $1 - \log^{-1}(n)$. Since

$$\begin{aligned} \|\!| L |\!\|_2 &= \|\!| (I + D_A^\top W_{AA}^{-1} D_A)^{-1} D_A^\top W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1} |\!\|_2 \\ &= \|\!| (I + D_A^\top W_{AA}^{-1} D_A)^{-2} D_A^\top W_{AA}^{-1} D_A |\!\|_2 < 1, \end{aligned}$$

with probability at least $1 - \log^{-1}(n)$

$$\max_{j \in A^c} \frac{||(n-G)^{-1} e_j^\top E_U U_A^\top W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1}||_2^2}{\sigma_{jj \cdot A}} \leq \mathcal{O}\left(\frac{(G-1)\log((p-s)\log(n))}{n-G} \right),$$

or equivalently

$$\max_{j \in A^c} ||(n-G)^{-1} e_j^\top E_U U_A^\top W_{AA}^{-1} D_A (I + D_A^\top W_{AA}^{-1} D_A)^{-1}||_2$$
$$\leq \mathcal{O}\left(\sqrt{\max_{j \in A^c} \sigma_{jj \cdot A} \frac{(G-1)\log((p-s)\log(n))}{n-G}} \right),$$

4. Show $\lambda \|\!| (n-G)^{-1} E_U U_A^\top W_{AA}^{-1} (I + W_{AA}^{-1} D_A D_A^\top)^{-1} s_A |\!\|_{\infty,2} \leq \lambda\alpha/4$.

Since $(n-G)^{-1} e_j^\top E_U \sim \mathcal{N}(0, (n-G)^{-2} \sigma_{jj \cdot A} I_{n-G})$, it follows that

$$\frac{1}{n-G} e_j^\top E_U U_A^\top (W_{AA} + D_A D_A^\top)^{-1} s_A$$
$$\sim \mathcal{N}\left(0, \frac{\sigma_{jj \cdot A}}{n-G} s_A^\top (W_{AA} + D_A D_A^\top)^{-1} W_{AA} (W_{AA} + D_A D_A^\top)^{-1} s_A \right).$$

Similar to parts 2 and 3, with probability at least $1 - \log^{-1}(n)$

$$\max_{j \in A^c} || \frac{1}{n-G} e_j^\top E_U U_A^\top (W_{AA} + D_A D_A^\top)^{-1} s_A ||_2$$
$$\leq \mathcal{O}\left(\sqrt{\|\!| L |\!\|_2 \max_{j \in A^c} \sigma_{jj \cdot A} \frac{(G-1)\log((p-s)\log(n))}{n}} \right),$$

where

$$\begin{aligned}
\|\!| L |\!\|_2 &= \|\!| s_A^\top (W_{AA} + D_A D_A^\top)^{-1} W_{AA} (W_{AA} + D_A D_A^\top)^{-1} s_A |\!\|_2 \\
&= \|\!| W_{AA}^{1/2} (W_{AA} + D_A D_A^\top)^{-1} s_A |\!\|_2^2 \\
&\leq s \|\!| W_{AA}^{1/2} W_{AA}^{-1/2} (I + W_{AA}^{-1/2} D_A D_A^\top W_{AA}^{-1/2})^{-1} W_{AA}^{-1/2} |\!\|_2^2 \\
&\leq s \|\!| (I + W_{AA}^{-1/2} D_A D_A^\top W_{AA}^{-1/2})^{-1} |\!\|_2^2 \|\!| W_{AA}^{-1/2} |\!\|_2^2 \\
&\leq s \|\!| W_{AA}^{-1} |\!\|_2.
\end{aligned}$$

Following the same argument as in part 2, with probability at least $1 - \mathcal{O}(\log^{-1} n)$

$$\begin{aligned}
&\max_{j \in A^c} || \frac{1}{n-G} e_j^\top E_U U_A^\top (W_{AA} + D_A D_A^\top)^{-1} s_A ||_2 \\
&\qquad \leq \mathcal{O}\left(\sqrt{\|\!| \Sigma_{AA}^{-1} |\!\|_2 \max_{j \in A^c} \sigma_{jj \cdot A} \frac{(G-1) s \log((p-s)\log(n))}{n}} \right).
\end{aligned}$$

Combining 1-4. The equations (4.12)-(4.15) are satisfied with probability at least $1 - \mathcal{O}(\log^{-1}(n))$ if for some constants $C_1 \geq 0$ and $C_2 \geq 0$

$$\alpha \geq C_1 \sqrt{\|\!| \Sigma_{AA}^{-1} |\!\|_2 \max_{j \in A^c} \sigma_{jj \cdot A} \frac{(G-1) s \log((p-s)\log(n))}{n}}$$

and

$$\lambda \geq \frac{1}{\alpha} C_2 \sqrt{\max_{j \in A^c} \sigma_{jj \cdot A} \frac{(G-1)\log((p-s)\log(n))}{n-G}}.$$

These inequalities are satisfied by 4.4 and the conditions on sample size n and tuning parameter λ from Theorem 2. □

4.7.2 Auxillary Technical Results

Lemma 8. *If $X_i | Y_i = g \sim \mathcal{N}(\mu_g, \Sigma)$ for $i = 1, ..., n$, then*

$$D \sim \mathcal{N}(\Delta + o(1), \Sigma/n \otimes I + o(1)); \quad (n-G) W_p \sim W(\Sigma, n-G).$$

Remark 1. *The bias term $o(1)$ does not depend on either s or p, and therefore we don't consider this term in the remaining analysis.*

Proof of Lemma 8. The result for W is trivial. The definition of D and the multivariate normality assumption on X_i imply $D \sim \mathcal{N}(\mu_D, \Sigma_{D1} \otimes \Sigma_{D2})$. It remains to show $\mu_D = \Delta + o(1)$, $\Sigma_{D1} = \Sigma/n$ and $\Sigma_{D2} = I$. Consider the rth column of D,

$$D_r = \frac{\sqrt{n_{r+1}} \sum_{g=1}^{r} n_g (\bar{X}_g - \bar{X}_{r+1})}{\sqrt{n} \sqrt{\sum_{g=1}^{r} n_g \sum_{g=1}^{r+1} n_g}},$$

and the rth column of Δ,

$$\Delta_r = \frac{\sqrt{\pi_{r+1}} \sum_{g=1}^{r} \pi_g (\mu_g - \mu_{r+1})}{\sqrt{\sum_{g=1}^{r} \pi_g \sum_{g=1}^{r+1} \pi_g}}.$$

Note that $\mathbb{E}(\bar{X}_i - \bar{X}_j) = \mu_i - \mu_j$ for all $i, j \in \{1, ..., G\}$. Moreover, $(n_1, ..., n_G) \sim Mult(n, (\pi_1, ..., \pi_G))$, and therefore $\mathbb{E}(n_i/n) = \pi_i$ and $\mathrm{Cov}(n_i/n, n_j/n) = \pi_i \pi_j / n$ for all $i, j \in \{1, ..., G\}$. Hence,

$$\begin{aligned}
\mathbb{E}(D_r) &= \mathbb{E}(\mathbb{E}(D_r | n_1, ..., n_G)) = \mathbb{E}\left(\frac{\sqrt{n_{r+1}} \sum_{g=1}^{r} n_g \mathbb{E}((\bar{X}_g - \bar{X}_{r+1}) | n_1, ..., n_G)}{\sqrt{n} \sqrt{\sum_{g=1}^{r} n_g \sum_{g=1}^{r+1} n_g}} \right) \\
&= \mathbb{E}\left(\frac{\sqrt{n_{r+1}} \sum_{g=1}^{r} n_g (\mu_g - \mu_{r+1})}{\sqrt{n} \sqrt{\sum_{g=1}^{r} n_g \sum_{g=1}^{r+1} n_g}} \right) \\
&= \Delta_r + o(1).
\end{aligned}$$

First, consider the case $n_g/n = \pi_g$ for all $g \in \{1, ..., G\}$. Since the groups are

independent,

$$
\begin{aligned}
&\mathrm{Cov}(D_r) \\
&= \mathbb{E}\left\{(D_r - \Delta_r)(D_r - \Delta_r)^\top\right\} \\
&= \frac{1}{Gr(r+1)}\mathbb{E}\left\{\left(\sum_{i=1}^{r}(\bar{x}_i - \mu_i) - r(\bar{x}_{r+1} - \mu_{r+1})\right)\left(\sum_{i=1}^{r}(\bar{x}_i - \mu_i) - r(\bar{x}_{r+1} - \mu_{r+1})\right)^\top\right\} \\
&= \frac{1}{Gr(r+1)}\left\{\sum_{i=1}^{r}\mathbb{E}\left\{(\bar{x}_i - \mu_i)(\bar{x}_i - \mu_i)^\top\right\} + r^2\mathbb{E}\left\{(\bar{x}_{r+1} - \mu_{r+1})(\bar{x}_{r+1} - \mu_{r+1})^\top\right\}\right\} \\
&= \frac{1}{Gr(r+1)}(r + r^2)\frac{\Sigma}{n/G} = \frac{\Sigma}{n},
\end{aligned}
$$

and for $s > r$

$$
\begin{aligned}
&\mathrm{Cov}(D_r, D_s) \\
&= \mathbb{E}\left\{(D_r - \Delta_r)(D_s - \Delta_s)^\top\right\} \\
&= \frac{1}{G\sqrt{r(r+1)s(s+1)}}\times \\
&\qquad \times \mathbb{E}\left\{\left(\sum_{i=1}^{r}(\bar{x}_i - \mu_i) - r(\bar{x}_{r+1} - \mu_{r+1})\right)\left(\sum_{i=1}^{s}(\bar{x}_i - \mu_i) - s(\bar{x}_{s+1} - \mu_{s+1})\right)^\top\right\} \\
&= \frac{1}{G\sqrt{r(r+1)s(s+1)}}\left\{\sum_{i=1}^{r}\mathbb{E}\left\{(\bar{x}_i - \mu_i)(\bar{x}_i - \mu_i)^\top\right\} - r\mathbb{E}\left\{(\bar{x}_{r+1} - \mu_{r+1})(\bar{x}_{r+1} - \mu_{r+1})^\top\right\}\right\} \\
&= \frac{1}{G\sqrt{r(r+1)s(s+1)}}(r - r)\frac{\Sigma}{n/G} = 0.
\end{aligned}
$$

The final result follows since $|n_i/n - \pi_i| = o(1)$. □

Lemma 9.

$$\frac{D_A^\top W_{AA}^{-1} e_j}{e_j^\top W_{AA}^{-1} e_j} | D_A \sim t_{G-1}(d_H, \mu_H, \Gamma_H)$$

with degrees of freedom $d_H = n - s - G + 2$, *mean* $\mu_H = D_A^\top \Sigma_{AA}^{-1} e_j / (e_j^\top \Sigma_{AA}^{-1} e_j)$ *and scale parameter* $\Gamma_H = \frac{1}{d_H}(D_A^\top R D_A)/(e_j^\top \Sigma_{AA}^{-1} e_j)$ *with* $R = \Sigma_{AA}^{-1} - \frac{\Sigma_{AA}^{-1} e_j e_j^\top \Sigma_{AA}^{-1}}{e_j^\top \Sigma_{AA}^{-1} e_j}$.

Proof of Lemma 9. Let

$$H = \begin{pmatrix} D_A^\top \Sigma_{AA}^{-1} D_A & D_A^\top \Sigma_{AA}^{-1} e_j \\ e_j^\top \Sigma_{AA}^{-1} D_A & e_j^\top \Sigma_{AA}^{-1} e_j \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} \\ H_{12}^\top & H_{22} \end{pmatrix},$$

and

$$\hat{H} = \begin{pmatrix} D_A^\top W_{AA}^{-1} D_A & D_A^\top W_{AA}^{-1} e_j \\ e_j^\top W_{AA}^{-1} D_A & e_j^\top W_{AA}^{-1} e_j \end{pmatrix} = \begin{pmatrix} \hat{H}_{11} & \hat{H}_{12} \\ \hat{H}_{12}^\top & \hat{H}_{22} \end{pmatrix}.$$

By definition, $\frac{D_A^\top W_{AA}^{-1} e_j}{e_j^\top W_{AA}^{-1} e_j} = \hat{H}_{12}\hat{H}_{22}^{-1}$. Let $M = (D_A\ e_j)^\top \in \mathbb{R}^{G\times s}$. Then H can be rewritten as $H = M\Sigma_{AA}^{-1}M^\top$ and $\hat{H}$ as $\hat{H} = MW_{AA}^{-1}M^\top$. Since $(n-G)W_{AA} \sim W_s(n-G, \Sigma_{AA})$ and $\mathrm{rank}(M) = G$, by [41, Theorem 3.2.11]

$$(n-G)\hat{H}^{-1} \sim W_G(n-s, H^{-1}),$$

or equivalently

$$\frac{1}{n-G}\hat{H} \sim W_G^{-1}(n-s+G+1, H).$$

By definition of R, $H_{11\cdot 2} = D_A^\top R D_A$. Using [10, Theorem 3], $\hat{H}_{12}\hat{H}_{22}^{-1}$ has density

$$\begin{aligned} f_{\hat{H}_{12}\hat{H}_{22}^{-1}}(X) =& \frac{|D_A^\top R D_A|^{-\frac{1}{2}} |e_j^\top \Sigma_{AA}^{-1} e_j|^{\frac{G-1}{2}}}{\pi^{(G-1)/2}} \frac{\Gamma(\frac{n-s+1}{2})}{\Gamma(\frac{n-s-G+2}{2})} \\ & \times |I + e_j^\top \Sigma_{AA}^{-1} e_j (D_A^\top R D_A)^{-1}(X - H_{12}H_{22}^{-1})(X - H_{12}H_{22}^{-1})^\top|^{-\frac{1}{2}(n-s+1)}. \end{aligned}$$

Since $|I + uv^\top| = 1 + u^\top v$,

$$\begin{aligned} f_{\hat{H}_{12}\hat{H}_{22}^{-1}}(X) =& \frac{|D_A^\top R D_A|^{-\frac{1}{2}} |e_j^\top \Sigma_{AA}^{-1} e_j|^{\frac{G-1}{2}}}{\pi^{(G-1)/2}} \frac{\Gamma(\frac{n-s+1}{2})}{\Gamma(\frac{n-s-G+2}{2})} \\ & \times \left(1 + e_j^\top \Sigma_{AA}^{-1} e_j (X - H_{12}H_{22}^{-1})^\top (D_A^\top R D_A)^{-1}(X - H_{12}H_{22}^{-1})\right)^{-\frac{1}{2}(n-s+1)}. \end{aligned}$$

This density corresponds to a $(G-1)$-dimensional elliptical t-distribution with $n-s-G+2$ degrees of freedom, mean $\mathbb{E}(\hat{H}_{12}\hat{H}_{22}^{-1}) = H_{12}H_{22}^{-1}$ and $\mathrm{Cov}(\hat{H}_{12}\hat{H}_{22}^{-1}) = \frac{1}{n-s-G}\frac{D_A^\top R D_A}{e_j^\top \Sigma_{AA}^{-1} e_j}$. □

Lemma 10. *With probability at least* $1 - \mathcal{O}(\log^{-1}(n))$

$$\|\!|(W_{AA} + D_A D_A^\top)^{-1}|\!\|_\infty \leq \sqrt{s}\|\!|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1}|\!\|_2 \left(1 + \mathcal{O}\left(\sqrt{\frac{s\log(\log(n))}{n}}\right)\right).$$

Proof of Lemma 10. First, we prove that unconditional distribution of $X_{Ai} \in \mathbb{R}^s$, $i = 1, ..., n$, is sub-gaussian: for all $x \in \mathbb{R}^s$, $< X_{Ai}, x >$ is sub-gaussian. Since $X_{Ai}|Y_i = g \sim \mathcal{N}(\mu_{gA}, \Sigma_{AA})$, X_{Ai} can be expressed as

$$X_{Ai} = C_{Ai} + Z_{Ai},$$

where $Z_{Ai} \sim \mathcal{N}(0, \Sigma_{AA})$ and $P(C_{Ai} = \mu_{gA}) = \pi_g$ for $g = 1, ..., G$. Let $\tilde{x} =< X_{Ai}, x >$, $\tilde{c} =< C_{Ai}, x >$ and $\tilde{z} =< Z_{Ai}, x >$. Then $\tilde{x} = \tilde{c} + \tilde{z}$. Consider the sub-gaussian norm of $\tilde{x}$ [62, Definition 5.7]

$$\|\tilde{x}\|_{\psi_2} = \sup_{d \geq 1} d^{-1/2} \left(\mathbb{E}|\tilde{x}|^d\right)^{1/d}.$$

By triangle inequality, $\|\tilde{x}\|_{\psi_2} \leq \|\tilde{c}\|_{\psi_2} + \|\tilde{z}\|_{\psi_2}$. Note that $\|\tilde{c}\|_{\psi_2}$ is finite for all x since C_{Ai} is a bounded random vector, and $\|\tilde{z}\|_{\psi_2}$ is finite for all x since Z_{Ai} is a zero-mean gaussian random vector. It follows that $\|\tilde{x}\|_{\psi_2}$ is finite for all x, hence X_{Ai} is unconditionally sub-gaussian.

By definition, $\Sigma_{AA} + \Delta_A \Delta_A^\top$ is unconditional population covariance matrix of X_A and $W_{AA} + D_A D_A^\top$ is unconditional sample covariance matrix of X_A. Using Theorem 5.39 in [62], with probability at least $1 - \log^{-1}(n)$

$$\|\!|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1/2}(W_{AA} + D_A D_A^\top)(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1/2} - I|\!\|_2 \leq C\sqrt{\frac{s\log(\log(n))}{n}}.$$

By submultiplicity of operator norm,

$$\begin{aligned}
&\|\!|(W_{AA} + D_A D_A^\top)^{-1} - (\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1}|\!\|_2 \\
&\quad \leq \|\!|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1}|\!\|_2 \|\!|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{1/2}(W_{AA} + D_A D_A^\top)^{-1}(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{1/2} - I|\!\|_2.
\end{aligned}$$

Therefore, with probability at least $1 - \log^{-1}(n)$

$$\|(W_{AA} + D_A D_A^\top)^{-1} - (\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1}\|_2 \leq C\|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1}\|_2 \sqrt{\frac{s \log(\log(n))}{n}}.$$

By triangle inequality,

$$\begin{aligned}
&\|(W_{AA} + D_A D_A^\top)^{-1}\|_\infty \\
&\leq \|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1}\|_\infty + \|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1} - (W_{AA} + D_A D_A^\top)^{-1}\|_\infty \\
&\leq \sqrt{s}\|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1}\|_2 + \sqrt{s}\|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1} - (W_{AA} + D_A D_A^\top)^{-1}\|_2 \\
&\leq \sqrt{s}\|(\Sigma_{AA} + \Delta_A \Delta_A^\top)^{-1}\|_2 \left(1 + \mathcal{O}\left(\sqrt{\frac{s \log(\log(n))}{n}}\right)\right).
\end{aligned}$$

□

Lemma 11. *With probability at least* $1 - \log^{-1}(n)$

$$\|D_A^\top \Sigma_{AA}^{-1} D_A - D_A^\top W_{AA}^{-1} D_A\|_2 \leq C\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2 \sqrt{\frac{(G-1)\log(\log(n))}{n}}.$$

Proof of Lemma 11. By submultiplicity of operator norm,

$$\begin{aligned}
&\|D_A^\top \Sigma_{AA}^{-1} D_A - D_A^\top W_{AA}^{-1} D_A\|_2 \\
&\leq \|D_A^\top \Sigma_{AA}^{-1} D_A\|_2 \|I - (D_A^\top \Sigma_{AA}^{-1} D_A)^{-1/2} D_A^\top W_{AA}^{-1} D_A (D_A^\top \Sigma_{AA}^{-1} D_A)^{-1/2}\|_2.
\end{aligned}$$

By Theorem 3.2.5 and Theorem 3.2.11 in [41],

$$(n-G)(D_A^\top \Sigma_{AA}^{-1} D_A)^{1/2} (D_A^\top W_{AA}^{-1} D_A)^{-1} (D_A^\top \Sigma_{AA}^{-1} D_A)^{1/2} \sim W_{G-1}(n-s-1, I).$$

By Lemma 9 in [65], with probability at most $2\exp\left(-(n-s-1)t^2/2\right)$,

$$\|\frac{n-s-1}{n-G}(D_A^\top \Sigma_{AA}^{-1} D_A)^{-1/2} D_A^\top W_{AA}^{-1} D_A (D_A^\top \Sigma_{AA}^{-1} D_A)^{-1/2} - I\|_2 \geq \delta(n-s-1, G-1, t),$$

where

$$\delta(n-s-1, G-1, t) = 2\left(\sqrt{\frac{G-1}{n-s-1}} + t\right) + \left(\sqrt{\frac{G-1}{n-s-1}} + t\right)^2.$$

Let

$$t = \sqrt{\frac{2\log(2\log n))}{n-s-1}}.$$

Then with probability at least $1 - \log^{-1}(n)$

$$\|\frac{n-s-1}{n-G}(D_A^\top \Sigma_{AA}^{-1} D_A)^{-1/2} D_A^\top W_{AA}^{-1} D_A (D_A^\top \Sigma_{AA}^{-1} D_A)^{-1/2} - I\|_2 \leq 8\sqrt{\frac{2(G-1)\log(2\log(n))}{n-s-1}}.$$

Hence, with probability at least $1 - \log^{-1}(n)$

$$\|(D_A^\top \Sigma_{AA}^{-1} D_A)^{-1/2} D_A^\top W_{AA}^{-1} D_A (D_A^\top \Sigma_{AA}^{-1} D_A)^{-1/2} - I\|_2 \leq C\sqrt{\frac{(G-1)\log(\log(n))}{n}}.$$

□

Lemma 12. *With probability at least* $1 - \log^{-1}(n)$

$$\begin{aligned}\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2 \leq & (G-1)\|\Delta_A^\top \Sigma_{AA}^{-1} \Delta_A\|_2 \\ & + \mathcal{O}\left(\frac{(G-1)s\log(\log(n))}{n} \vee \sqrt{\|\Delta_A^\top \Sigma_{AA}^{-1} \Delta_A\|_2 \frac{(G-1)\log(\log(n))}{n}}\right).\end{aligned}$$

Proof of Lemma 12. Since $D_A^\top \Sigma_{AA}^{-1} D_A$ is a positive semi-definite matrix,

$$\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2 \leq \mathrm{Tr}(D_A^\top \Sigma_{AA}^{-1} D_A).$$

Recall that $D_A \sim \mathcal{N}(\Delta_A, \Sigma_{AA}/n \otimes I)$. Therefore for all $i \in \{1, .., (G-1)\}$

$$n e_i^\top D_A^\top \Sigma_{AA}^{-1} D_A e_i \sim \chi_s^2\left(n e_i^\top \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A e_i\right).$$

From [32, Lemma 11], for all $i \in \{1, .., (G-1)\}$ with probability at least $1 - \log^{-1}(n)$

$$\begin{aligned}e_i^\top D_A^\top \Sigma_{AA}^{-1} D_A e_i \leq & e_i^\top \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A e_i \\ & + \mathcal{O}\left(\frac{s\log((G-1)\log(n))}{n} \vee \sqrt{e_i^\top \Delta_A^\top \Sigma_{AA}^{-1} \Delta_A e_i \frac{\log((G-1)\log(n))}{n}}\right),\end{aligned}$$

or equivalently

$$\begin{aligned}\mathrm{Tr}(D_A^\top \Sigma_{AA}^{-1} D_A) \leq & \mathrm{Tr}(\Delta_A^\top \Sigma_{AA}^{-1} \Delta_A) \\ & + \mathcal{O}\left(\frac{(G-1)s\log((G-1)\log(n))}{n} \vee \sqrt{\mathrm{Tr}(\Delta_A^\top \Sigma_{AA}^{-1} \Delta_A) \frac{\log((G-1)\log(n))}{n}}\right).\end{aligned}$$

Since $\mathrm{Tr}(\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A) \leq (G-1)\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2$ and $G = \mathcal{O}(1)$, it follows that with probability at least $1 - \log^{-1}(n)$

$$\begin{aligned}\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2 \leq & (G-1)\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2 \\ & + \mathcal{O}\left(\frac{(G-1)s\log(\log(n))}{n} \vee \sqrt{\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2 \frac{(G-1)\log(\log(n))}{n}}\right).\end{aligned}$$

□

Lemma 13. *With probability at least* $1 - \mathcal{O}(\log^{-1}(n))$

$$\begin{aligned}\|D_A^\top W_{AA}^{-1} D_A\|_2 \leq & C\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2 \\ & + \mathcal{O}\left(\frac{(G-1)s\log(\log(n))}{n} \vee \sqrt{\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2 \frac{(G-1)\log(\log(n))}{n}}\right).\end{aligned}$$

Proof of Lemma 13. By triangle inequality and Lemma 12,

$$\begin{aligned}\|D_A^\top W_{AA}^{-1} D_A\|_2 = & \frac{\|D_A^\top W_{AA}^{-1} D_A\|_2}{\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2}\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2 \\ \leq & \frac{\|D_A^\top W_{AA}^{-1} D_A\|_2}{\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2}\big((G-1)\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2 \\ & + \mathcal{O}\left(\frac{(G-1)s\log(\log(n))}{n} \vee \sqrt{\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2 \frac{(G-1)\log(\log(n))}{n}}\right).\end{aligned}$$

From Lemma 11, with probability at least $1 - \log^{-1}(n)$

$$\begin{aligned}\frac{\|D_A^\top W_{AA}^{-1} D_A\|_2}{\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2} \leq & \frac{\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2 + \|D_A^\top W_{AA}^{-1} D_A - D_A^\top \Sigma_{AA}^{-1} D_A\|_2}{\|D_A^\top \Sigma_{AA}^{-1} D_A\|_2} \\ \leq & 1 + C\sqrt{\frac{(G-1)\log(\log(n))}{n}} \leq C'.\end{aligned}$$

Combining with the previous display, we obtain with probability at least $1 - \mathcal{O}(\log^{-1}(n))$ that

$$\begin{aligned}\|D_A^\top W_{AA}^{-1} D_A\|_2 \leq & C\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2 \\ & + \mathcal{O}\left(\frac{(G-1)s\log(\log(n))}{n} \vee \sqrt{\|\Delta_A^\top \Sigma_{AA}^{-1}\Delta_A\|_2 \frac{(G-1)\log(\log(n))}{n}}\right).\end{aligned}$$

□

CHAPTER 5

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

5.1 Summary

The answers to important scientific questions depend on the analysis of complex modern datasets, which present a myriad of challenges for the existing statistical techniques. Two of these challenges are covariance estimation and over-selection of variables. In this dissertation we investigated how to address these challenges in generalized eigenvalue problems in multivariate analysis. Our main findings are summarized below.

1. Careful consideration should be given to the choice of variable selection method in nonconvex problems. In particular, a popular ℓ_1 penalization technique can fail to achieve a sparse solution, and an ℓ_1 constraint is preferred if strong variable selection is desired (Chapter 2).
2. Joint consideration of variable selection and covariance estimation problems allows one to achieve optimal theoretical guarantees and fast computations in sparse high-dimensional discriminant analysis (Chapters 3 and 4).

These insights provide a starting point to new research directions.

5.2 Variable selection and consistency in nonconvex problems

Many statistical problems naturally correspond to nonconvex optimization problems. In Chapter 2 we considered a special example of nonconvex problem, a generalized eigenvalue problem, and showed that ℓ_1 penalization is not an effective variable selection tool.

There exist numerous other examples of nonconvex optimization problems which use ℓ_1 penalty for variable selection. [2] utilizes it for generalized matrix factorization, [3] for transposable covariance model, [9] for covariance estimation, [53, 76] for principal components analysis. All of these problems can be written in the form

$$x_\lambda = \arg\min_{x \in \mathbb{A}} \left\{ f(x) + \lambda \|x\|_1 \right\}, \tag{5.1}$$

where $\mathbb{A}$ is a convex set and $f(x)$ is nonconvex. Problem (5.1) is a Lagrangian of

$$x_\tau = \arg\min_{x \in \mathbb{A}} \{ f(x) \} \quad \text{subject to } \|x\|_1 \leq \tau. \tag{5.2}$$

Proposition 3. *For every $\lambda \geq 0$ there exists $\tau \geq 0$ such that $x_\lambda = x_\tau$.*

The reverse, however, is not generally true for nonconvex $f(x)$. In Chapter 2 we showed that this is not true for the generalized eigenvalue problem due to the duality gap between (5.1) and (5.2). Poor variable selection performance of ℓ_1-penalized generalized eigenvalue problem is a consequence of this duality gap.

Therefore, it is of interest to investigate the presence of the duality gap in other nonconvex problems with ℓ_1 penalty, and its effect on the variable selection performance. To achieve this goal, we plan to pursue the following analysis:

1. Conduct an empirical study to assess the relationship between the tuning parameter λ and sparsity level of the corresponding solution x_λ for [2, 3, 9, 53, 76]
2. Characterize the class of nonconvex functions $f(x)$ such that (5.1) fails to achieve sparse solution
3. Provide guidance on assessing the presence of duality gap for a given nonconvex problem (5.1) and offer alternative formulation that achieves sparse solution (i.e. ℓ_1-constrained version (5.2)).

While the use of ℓ_1-constraint in (5.2) overcomes the challenge of finding sparse solution, the theoretical properties of the resulting estimator remain unknown. The difficulty lies in nonconvexity as the convergence to global solution is not guaranteed. Empirically, however, the local solutions demonstrate excellent performance, often outperforming competing convex methods. There has been a lot of interest in deriving theoretical properties of local solutions to nonconvex problems. [35] study the first order conditions of M-estimators with nonconvex penalties, and prove that any stationary point will lie within statistical precision of the underlying parameter vector. [66] study an approximate regularization path following algorithm for M-estimators with nonconvex penalties, and provide statistical guarantees for the local solution to the algorithm.

Unfortunately, these results are not suitable for either (5.1) or (5.2). Problem (5.1) is nonconvex due to the function $f(x)$, whereas the results of [35, 66] are generally restricted to nonconvex problems with convex $f(x)$ and nonconvex penalties.

Given this gap in the literature, it is of interest to develop a unified framework for the theoretical analysis of (5.1) that will support excellent empirical

performance of local solutions. To achieve this goal, we plan to try two strategies:

1. Similar to [35], we plan to restrict the analysis to the first-order optimality conditions for (5.1). This technique has potential for providing theoretical results for all stationary points, and as such, is independent of the choice of optimization algorithm.
2. Given the biconvex nature of the problems of interest [2, 3, 9, 53, 68, 69, 76], we plan to restrict the analysis to the local solution of biconvex optimization algorithm. This strategy is similar to the one employed by [66] for nonconvex penalties.

5.3 Nonlinear discriminant analysis in high dimensions

Linear projections are motivated by multivariate normal data where correlations measure the dependence between variables. In practice, the normality assumption is often violated, leading to complex relationships between the features and, as a result, unsatisfactory performance of linear methods. Given the success of direct estimation approach in linear discriminant analysis (Chapters 3 and 4), it is of interest to investigate whether the same direct estimation strategy can be successfully applied in nonlinear settings.

5.3.1 Sparse quadratic discriminant analysis

Let (X_i, Y_i), $i = 1, \cdots, n$, be independent pairs with $Y_i \in \{1, \cdots, G\}$ and $X_i | Y_i = g \sim \mathcal{N}(\mu_g, \Sigma_g)$. In linear discriminant analysis, it is assumed that $\Sigma_1 = \cdots = \Sigma_G$, leading to linear classification boundaries. We relax this assumption in this section. For a new observation $x \in \mathbb{R}^p$, the optimal population classification rule $h(x)$ is

$$h(x) = \arg\min_g \left\{ (x - \mu_g)^\top \Sigma_g^{-1} (x - \mu_g) - \log\det \Sigma_g^{-1} - 2 \log \pi_g \right\},$$

where $\pi_g = P(Y_i = g)$. Let S_g be the sample covariance matrix for group g. The sample classification rule $\hat{h}(x)$ is

$$\hat{h}(x) = \arg\min_g \left\{ (x - \bar{X}_g)^\top S_g^{-1} (x - \bar{X}_g) - \log\det S_g^{-1} - 2 \log \frac{n_g}{n} \right\}.$$

Since $\hat{h}(x)$ induces quadratic classification boundaries, the resulting classification procedure is called quadratic discriminant analysis (QDA).

Unfortunately, QDA performs poorly on high-dimensional datasets. This unsatisfactory performance is largely due to the estimation of G precision matrices Σ_g^{-1}, a task that is extremely challenging when $p >> n$. In fact, even when $p = 0.5n$ and the assumption of equal covariance matrices is violated, the misclassification error rate of QDA is worse than the misclassification error rate of sparse LDA methods (see simulations of Section 3.7.4 in Chapter 3).

Several extensions of traditional QDA haven been proposed in the literature. A common strategy is joint estimation of G precision matrices Σ_g^{-1}. [24] propose to regularize group sample covariance matrices S_g by shrinkage. [18, 27, 45, 55] use penalized likelihood technique, where the penalty enforces similarity either between the covariance matrices Σ_g or the precision matrices Σ_g^{-1}.

While these methods perform better than traditional QDA in high-dimentional settings, their main focus is on estimating G precision matrices Σ_g^{-1}. The precision matrices play a crucial role in determining the optimal population classification rule $h(x)$, however, they are not the primary objects of interest. Moreover, in many applications it is implicitly assumed that only a few out of p variables are relevant for the classification. This assumption has not been explored by the current high-dimensional QDA proposals.

Our goal is to develop a new methodology for sparse quadratic discriminant analysis that overcomes the drawbacks of existing QDA methods. To achieve this goal, we plan to pursue the following analysis

1. Determine the key object of interest in quadratic discriminant analysis that governs the projection and the resulting classification rule (in linear discriminant analysis this object is $\Sigma_g^{-1}\Delta$, see Chapter 3)
2. Understand the structural consequence of the sparsity assumption on that object (i.e. row/column sparsity, sparsity of individual elements)
3. Formulate an optimization problem that combines the loss function with the sparsity-inducing penalty
4. Perform simulation study to empirically assess the performance of the new procedure
5. Develop theoretical guarantees for variable selection and classification consistency

5.3.2 Sparse kernel discriminant analysis

Let (X_i, Y_i), $i = 1, \cdots, n$, be independent pairs with $Y_i \in \{1, \cdots, G\}$ and $X_i|Y_i = g \sim \mathcal{N}(\mu_g, \Sigma)$. Recall that Fisher's Linear Discriminant Analysis solves

$$\text{maximize}_v \left\{ \frac{v^\top B v}{v^\top W v} \right\}, \tag{5.3}$$

where

$$B = \sum_{g=1}^{G} n_g (\bar{X}_g - \bar{X})(\bar{X}_g - \bar{X})^\top,$$
$$W = \sum_{g=1}^{G} (n_g - 1) S_g = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (X_{ig} - \bar{X}_g)(X_{ig} - \bar{X}_g)^\top.$$

Consider a possibly nonlinear transformation $\phi(\cdot) : \mathbb{R}^p \to H$, where H is a new feature space of higher dimension. Consider Fisher's LDA in the new feature space H,

$$\text{maximize}_v \left\{ \frac{v^\top B^H v}{v^\top W^H v} \right\}, \tag{5.4}$$

where

$$B^H = \sum_{g=1}^{G} n_g (\bar{X}_g^H - \bar{X}^H)(\bar{X}_g^H - \bar{X}^H)^\top,$$
$$W^H = \sum_{g=1}^{G} (n_g - 1) S_g^H = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\phi(X_{ig}) - \bar{X}_g^H)(\phi(X_{ig}) - \bar{X}_g^H)^\top,$$
$$\bar{X}_g^H = \frac{1}{n_g} \sum_{i=1}^{n_g} \phi(X_{ig}),$$
$$\bar{X} = \frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \phi(X_{ig}).$$

The main advantage of (5.4) over (5.3) is its flexibility with respect to nonlinear patterns. While the assumption $X_i|Y_i = g \sim \mathcal{N}(\mu_g, \Sigma)$ is likely to be violated in practice, it is always possible to project to sufficiently higher-dimensional

space H so that $\phi(X_i)|Y_i = g \sim \mathcal{N}(\mu_g^H, \Sigma^H)$ holds. By transforming the original features, the nonlinear classification boundary in $\mathbb{R}^p$ becomes linear in H, hence (5.4) can be used to achieve optimal classification performance.

Direct implementation of (5.4) requires the knowledge of $\phi(\cdot)$ and computation of B^H and W^H, which is prohibitive in high dimensions. To overcome this drawback, note that B^H and W^H depend on $\phi(\cdot)$ only through the inner products. These inner products can be efficiently computed without explicit mapping to H by using Mercer kernels. For example, gaussian kernel $k(x, y) = \exp(-\|x - y\|_2^2/c)$ has a corresponding $\phi(\cdot)$ such that $k(x, y) =< \phi(x), \phi(y) >$.

Given that (5.4) corresponds to higher-dimensional feature space, its application for large p is likely to produce unfavorable results due to overfitting and singularity of W^H. Kernel regularization is commonly used to solve this problem [40, 74]. While it prevents overfitting and singularity of W^H, the resulting classification rule relies on all p original features. As such, the interpretation of results becomes very challenging.

[1] introduces feature weights within the kernels, and estimates some of these weights as zero through ℓ_1 penalization. The resulting problem is nonconvex, however the local solution can be effectively found using biconvex optimization scheme. The nonconvexity, however, presents significant challenges for theoretical analysis, and as a result, the method comes with no variable selection or classification guarantees.

Inspired by the direct estimation approach to linear discriminant analysis (Chapter 3), our goal is to develop a direct estimation procedure for kernel discriminant analysis borrowing strength from the ideas of [1]. In short, we want

to construct kernels that adaptively choose a subset of variables to produce interpretable, nonlinear, low-dimensional structures for classification. To achieve this goal, we plan to pursue the following analysis

1. Gain a better understanding of the ideas of [1] and explore the possibility of convex formulation of sparse kernel discriminant analysis, possibly by modifying the convex formulation of sparse LDA (Chapter 3)
2. Perform simulation studies to assess empirical tradeoffs between sparse LDA, sparse QDA and sparse kernel LDA
3. Perform simulation studies to assess the sensitivity with respect to the choice of kernel
4. Develop theory for variable selection and classification consistency of sparse kernel LDA

5.4 Sparse canonical correlation analysis

Canonical Correlation Analysis (CCA), described in Chapter 1.2.3, is a standard multivariate analysis tool that is used to find linear combinations of two sets of features with the maximum correlation. Unfortunately, it is not suitable for the analysis of modern high-dimensional datasets. First, when the number of features is large, it's impossible to interpret linear combinations that are based on all the features. Secondly, CCA relies on the inverse of sample covariance matrix, which is singular when $p >> n$.

There has been a lot of interest in sparse canonical correlation analysis with linear combinations based only on the subset of features. [15, 68] use ℓ_1 norm to

perform variable selection and assume independent covariance structure. [14] split the data into two parts, one part is used for precision matrix estimation and the other part for variable selection through thresholded singular value decomposition. Other methods for sparse CCA are [43] (iterative soft-thresholded singular value decomposition), [63, 64] (penalized regression framework) and [67] (greedy algorithm).

The existing proposals for sparse CCA suffer from two main drawbacks. First, the majority of them separate the problems of covariance estimation and variable selection [14, 15, 68]. Given the success of direct estimation approach in discriminant analysis (Chapters 3 and 4), I conjecture that better performance can be achieved by combining these problems. Secondly, except for [14], these methods have no theoretical guarantees on either the variable selection or estimation consistency.

Our goal is to develop a computationally efficient and theoretically sound procedure for sparse canonical correlation analysis by jointly solving the problems of covariance estimation and variable selection. Let $X \in \mathbb{R}^{n\times p_1}$ and $Y \in \mathbb{R}^{n\times p_2}$ be two sets of features on the same sample of size n. We propose to estimate canonical directions (w_1, w_2) as

$$\begin{aligned}(\hat{w}_1, \hat{w}_2) = \arg\min_{w_1,w_2} \left\{ \frac{1}{2n}\|Xw_1 - Yw_2\|_2^2 + \lambda_1\|w_1\|_1 + \lambda_2\|w_2\|_1 \right\} \\ \text{s.t.} \quad \frac{1}{n}w_1^\top X^\top X w_1 = 1, \frac{1}{n}w_2^\top Y^\top Y w_2 = 1.\end{aligned} \tag{5.5}$$

We call this method penalized CCA.

Proposition 4. *Penalized CCA is a special case of Generalized Least-Square Matrix Decomposion [2].*

Proof. Generalized penalized matrix factorization problem (GPMF) is defined

as

$$\arg\max_{v,u} \left\{ u^\top QZRv - \lambda_1 \|v\|_1 - \lambda_2 \|u\|_1 \right\}$$
$$\text{s.t.} \quad u^\top Qu \leq 1, v^\top Rv \leq 1.$$

Penalized CCA problem (5.5) can be rewritten as

$$\arg\max_{w_1,w_2} \left\{ w_1^\top \frac{X^\top Y}{n} w_2 - \lambda_1 \|w_1\|_1 - \lambda_2 \|w_2\|_1 \right\}$$
$$\text{s.t.} \quad \frac{1}{n} w_1^\top X^\top X w_1 = 1, \frac{1}{n} w_2^\top Y^\top Y w_2 = 1.$$

Moreover, the equality constraints can be substituted by inequality constraints

$$\arg\max_{w_1,w_2} \left\{ w_1^\top \frac{X^\top Y}{n} w_2 - \lambda_1 \|w_1\|_1 - \lambda_2 \|w_2\|_1 \right\}$$
$$\text{s.t.} \quad \frac{1}{n} w_1^\top X^\top X w_1 \leq 1, \frac{1}{n} w_2^\top Y^\top Y w_2 \leq 1.$$

This is equivalent to GPMF with $Q = \frac{X^\top X}{n}$, $R = \frac{Y^\top Y}{n}$ and $QZR = \frac{X^\top Y}{n}$. □

While (5.5) is nonconvex, a biconvex iterative scheme can be used to find local solution:

$$\begin{aligned}
&\hat{w}_1 = \arg\min_{w_1} \frac{1}{2n} \|Xw_1 - Y\hat{w}_2^{(t)}\|_2 + \lambda_1 \|w_1\|_1; \\
&\hat{w}_1^{(t+1)} = \hat{w}_1 / \sqrt{\frac{1}{n} \hat{w}_1^\top X^\top X \hat{w}_1}; \\
&\hat{w}_2 = \arg\min_{w_2} \frac{1}{2n} \|Xw_1^{(t+1)} - Yw_2\|_2 + \lambda_2 \|w_2\|_1; \\
&\hat{w}_2^{(t+1)} = \hat{w}_2 / \sqrt{\frac{1}{n} \hat{w}_2^\top Y^\top Y \hat{w}_2}.
\end{aligned} \tag{5.6}$$

Algorithm (5.6) iterates between solving two LASSO regression problems. Due to the form of the LASSO updates, (5.6) can be viewed as a modified power method for generalized singular value decomposition. A similar idea has been explored by [63] and [68].

Preliminary results demonstrate the convergence of biconvex optimization algorithm, and the existence of λ_1 and λ_2 with small estimation error for corresponding solutions. However, the convergence to global solution is not guaranteed, which makes the theoretical analysis very challenging. As such, in future work we plan to pursue the following directions:

1. Perform simulation studies to compare the empirical performance of penalized CCA with other sparse CCA methods in the literature.
2. Study variable selection consistency and estimation consistency of the local solutions to algorithm (5.6). Our first strategy is to analyze the first-order conditions of optimization problem (5.5), this approach has been previously used by [35]. Our second strategy is to use the connection between (5.6) and power method for generalized singular value decomposition, and adapt the existing proof of convergence for the power method.

BIBLIOGRAPHY

[1] Genevera I Allen. Automatic feature selection via weighted kernels and regularization. *Journal of Computational and Graphical Statistics*, 22(2):284–299, 2013.

[2] Genevera I Allen, Logan Grosenick, and Jonathan Taylor. A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, 109(505):145–159, 2014.

[3] Genevera I Allen and Robert Tibshirani. Transposable regularized covariance models with an application to missing data imputation. *Annals of Applied Statistics*, 4(2):764–790, 2010.

[4] U Alon, N Barkai, D A Notterman, K Gish, S Ybarra, D Mack, and A 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 of the United States of America*, 96(12):6745–6750, 1999.

[5] Francis R Bach, R Jenatton, and J Mairal. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, 2011.

[6] Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, 1999.

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

[8] Peter J. J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *Annals of Statistics*, 36(1):199–227, 2008.

[9] Jacob Bien and Robert Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820, 2011.

[10] T Bodnar and Y Okhrin. Properties of the singular, inverse and generalized inverse partitioned Wishart distributions. *Journal of Multivariate Analysis*, 99(10):2389–2405, 2008.

[11] Stephen P Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge Univ Press, Cambridge, 2004.

[12] Tony Cai and Weidong Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106(496):1566–1577, 2011.

[13] Sumit Chakraborty, Todd Gruber, Clifton E Barry, Helena I Boshoff, and Kyu Y Rhee. Para-aminosalicylic acid acts as an alternative substrate of folate metabolism in Mycobacterium tuberculosis. *Science*, 339(6115):88–91, 2013.

[14] Mengjie Chen, Chao Gao, Zhao Ren, and Harrison H Zhou. Sparse CCA via Precision Adjusted Iterative Thresholding. *arXiv.org*, 2013.

[15] Eric C Chi, Genevera I Allen, Hua Zhou, Omid Kohannim, Kenneth Lange, and Paul M Thompson. Imaging genetics via sparse canonical correlation analysis. In *Proceedings of IEEE International Symposium on Biomedical Imaging: from nano to macro*, pages 740–743, 2013.

[16] Line Clemmensen, Trevor Hastie, Daniela M Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

[17] Hengjian Cui, Runze Li, and Wei Zhong. Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *Journal of the American Statistical Association*, pages 00–00, 2014.

[18] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society Ser. B*, 76(2):373–397, 2014.

[19] Luiz Pedro S de Carvalho, Crystal M Darby, Kyu Y Rhee, and Carl Nathan. Nitazoxanide Disrupts Membrane Potential and Intrabacterial pH Homeostasis of Mycobacterium tuberculosis. *ACS medicinal chemistry letters*, 2(11):849–854, 2011.

[20] Luiz Pedro S de Carvalho, Steven M Fischer, Joeli Marrero, Carl Nathan, Sabine Ehrt, and Kyu Y Rhee. Metabolomics of Mycobacterium tuberculosis reveals compartmentalized co-catabolism of carbon substrates. *Chemistry & Biology*, 17(10):1122–1131, 2010.

[21] Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.

[22] Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of Statistics*, 36(6):2605, 2008.

[23] Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society Ser. B*, 74(4):745–771, 2012.

[24] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

[25] G H Golub and C F Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 4 edition, 2012.

[26] J. Gorski, F. Pfeuffer, and K. Klamroth. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007.

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

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

[29] Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The Elements of Statistical Learning*. Springer, New York, second edition, 2009.

[30] Daniel Hsu, Sham M Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:52–6, 2012.

[31] Jian Huang, Patrick Breheny, and Shuangge Ma. A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science*, 27(4):481–499, 2012.

[32] Mladen Kolar and Han Liu. Optimal feature selection in high-dimensional discriminant analysis. *arXiv.org*, 2013.

[33] W J Krzanowski, P Jonathan, W V McCarthy, and M R Thomas. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Journal of the Royal Statistical Society Ser. C*, 44(1):101–115, 1995.

[34] B Laurent and P Massart. Adaptive Estimation of a Quadratic Functional by Model Selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

[35] Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *arXiv.org*, 2013.

[36] Q Mai and H Zou. A note on the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics*, 55(2):243–246, 2013.

[37] Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.

[38] K V Mardia, J T Kent, and J M Bibby. *Multivariate Analysis*. Academic Press, New York, 1979.

[39] Geoffrey J McLachlan. *Discriminant analysis and statistical pattern recognition*. John Wiley and Sons, Inc.

[40] S Mika, G Ratsch, J Weston, B Scholkopf, and K Muller. Fisher discriminant analysis with kernels. In *Proceedings of the 1999 IEEE Signal Processing Society Workshop*, pages 41–48, 1999.

[41] Robb J Muirhead. *Aspects of multivariate statistical theory*. John Wiley and Sons, Inc., New York, 1982.

[42] Guillaume Obozinski, Martin J Wainwright, and Michael I Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39(1):1–47, 2011.

[43] Elena Parkhomenko, David Tritchler, and Joseph Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009.

[44] Kevin Pethe, Patricia C Sequeira, Sanjay Agarwalla, Kyu Rhee, Kelli Kuhen, Wai Yee Phong, Viral Patel, David Beer, John R Walker, Jeyaraj Duraiswamy, Jan Jiricek, Thomas H Keller, Arnab Chatterjee, Mai Ping Tan, Manjunatha Ujjini, Srinivasa P S Rao, Luis Camacho, Pablo Bifani, Puiying A Mak, Ida Ma, S Whitney Barnes, Zhong Chen, David Plouffe, Pamela Thayalan, Seow Hwee Ng, Melvin Au, Boon Heng Lee, Bee Huat Tan, Sindhu Ravindran, Mahesh Nanjundappa, Xiuhua Lin, Anne Goh,

Suresh B Lakshminarayana, Carolyn Shoen, Michael Cynamon, Barry Kreiswirth, Veronique Dartois, Eric C Peters, Richard Glynne, Sydney Brenner, and Thomas Dick. A chemical genetic screen in Mycobacterium tuberculosis identifies carbon-source-dependent growth inhibitors devoid of in vivo efficacy. *Nature Communications*, 1:57, 2010.

[45] Bradley S Price, Charles J Geyer, and Adam J Rothman. Ridge Fusion in Statistical Learning. *Journal of Computational and Graphical Statistics*, (just-accepted):00–00, 2014.

[46] Zhiwei Qin, Katya Scheinberg, and Donald Goldfarb. Efficient block-coordinate descent algorithms for the Group Lasso. *Mathematical Programming Computation*, 5(2):143–169, 2013.

[47] S Ramaswamy, P Tamayo, R Rifkin, S Mukherjee, C H Yeang, M Angelo, C Ladd, M Reich, E Latulippe, J P Mesirov, T Poggio, W Gerald, M Loda, E S Lander, and T R Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26):15149–15154, 2001.

[48] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *arXiv.org*, math.PR, 2013.

[49] Andrew L Rukhin. Generalized Bayes estimators of a normal discriminant function. *Journal of Multivariate Analysis*, 41(1):154–162, 1992.

[50] Shayle R Searle. *Linear Models for Unbalanced Data*. Wiley-Interscience, 2006.

[51] G A F Seber. *Multivariate observations*. John Wiley and Sons, Inc., New York, 1984.

[52] Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics*, 39(2):1241–1265, 2011.

[53] Haipeng Shen and Jianhua Z Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99(6):1015–1034, 2008.

[54] N Simon, Jerome H Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[55] Noah Simon and Rob Tibshirani. Discriminant Analysis with Adaptively Pooled Covariance. *arXiv.org*, 2011.

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

[57] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6567–6572, 2002.

[58] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18(1):104–117, 2003.

[59] Nickolay T. Trendafilov and Ian T. Jolliffe. Projected gradient approach to the numerical solution of the SCoTLASS. *Computational Statistics & Data Analysis*, 50(1):242–253, 2006.

[60] A W van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

[61] Charles F Van Loan. Generalizing the Singular Value Decomposition. *SIAM Journal on Numerical Analysis*, 13(1):76–83, 1976.

[62] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv.org*, math.PR, 2010.

[63] Sandra Waaijenborg, Philip C Verselewel de Witt Hamer, and Aeilko H Zwinderman. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology*, 7:1–29, 2008.

[64] Sandra Waaijenborg and Aeilko H Zwinderman. Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC bioinformatics*, 10(1):315, 2009.

[65] Martin J Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using -Constrained Quadratic Programming (Lasso). *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 55(5):2183–2202, 2009.

[66] Zhaoran Wang, Han Liu, and Tong Zhang. Optimal Computational and

Statistical Rates of Convergence for Sparse Nonconvex Learning Problems. *arXiv.org*, stat.ML, 2013.

[67] Ami Wiesel, Mark Kliger, and Alfred O Hero, III. A greedy approach to sparse canonical correlation analysis. *arXiv.org*, 2008.

[68] Daniela M Witten and Robert Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):28, 2009.

[69] Daniela M Witten and Robert Tibshirani. Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society Ser. B*, 73(5):753–772, 2011.

[70] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[71] Michael C Wu, Lingsong Zhang, Zhaoxi Wang, David C Christiani, and Xihong Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.

[72] Ping Xu, Guy N Brock, and Rudolph S Parrish. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics & Data Analysis*, 53(5):1674–1687, 2009.

[73] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Ser. B*, 68(1):49–67, 2006.

[74] S Zafeiriou, G Tzimiropoulos, M Petrou, and T Stathaki. Regularized Kernel Discriminant Analysis With a Robust Kernel for Face Recognition and Verification. *IEEE Transactions on Neural Networks and Learning Systems*, 23(3):526–534.

[75] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[76] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.