ON HIGH DIMENSIONAL SPARSE REGRESSION AND ITS INFERENCE

Qiang Sun

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2014

Approved by:

Hongtu Zhu

Joseph Ibrahim

Donglin Zeng

Michael Kosorok

Yufeng Liu

Hongyu An

ABSTRACT

QIANG SUN: On High Dimensional Sparse Regression and Its Inference
(Under the direction of Hongtu Zhu and Joseph Ibrahim)

In the first part of this work, we aim to develop a sparse projection regression modeling (SPReM) framework to perform multivariate regression modeling with a large number of responses and a multivariate covariate of interest. We propose two novel heritability ratios to simultaneously perform dimension reduction, response selection, estimation, and testing, while explicitly accounting for correlations among multivariate responses. Our SPReM is devised to specifically address the low statistical power issue of many standard statistical approaches, such as the Hotelling's T^2 test statistic or a mass univariate analysis, for high-dimensional data. We formulate the estimation problem of SPREM as a novel sparse unit rank projection (SURP) problem and propose a fast optimization algorithm for SURP. Furthermore, we extend SURP to the sparse multi-rank projection (SMURP) by adopting a sequential SURP approximation. Theoretically, we have systematically investigated the convergence properties of SURP and the convergence rate of SURP estimates. Our simulation results and real data analysis have shown that SPReM outperforms other state-of-the-art methods.

In the second part of this work, we propose a Hard Thresholded Regression (HTR) framework for simultaneous variable selection and unbiased estimation in high dimensional linear regression. This new framework is motivated by its close connection with the L_0 regularization and best subset selection under orthogonal design, while enjoying several key computational and theoretical advantages over many existing penalization methods (e.g., SCAD or MCP). Computationally, HTR is a fast two-stage estimation procedure consisting of the first step for calculating a coarse initial estimator and the second step for solving a

linear program. Theoretically, under some mild conditions, the HTR estimator is shown to enjoy the strong oracle property and thresholded property even when the number of covariates may grow at an exponential rate. We also propose to incorporate the regularized covariance estimator into the estimation procedure in order to better trade off between noise accumulation and correlation modeling. Under this scenario with regularized covariance matrix, HTR includes Sure Independence Screening as a special case. Both simulation and real data results show that HTR outperforms other state-of-the-art methods.

In the third part of this work, we focus on multicategory classification and propose the sparse multicategory discriminant analysis. Many supervised machine learning tasks can be cast as multicategory classification problems. Linear discriminant analysis has been well studied in two class classification problems and can be easily extended to multicatigory cases. For high dimensional classification, traditional linear discriminant analysis fails due to diverging spectra and accumulation of noise. Therefore, researchers have proposed penalized LDA (Fan et al. 2012, Witten and Tibshirani 2011). However, most available methods for high dimensional multi-class LDA are based on an iterative algorithm, which is computationally expensive and not theoretically justified. In this paper, we present a new framework for sparse multicategory discriminant analysis (SMDA) for high dimensional multi-class classification by simultaneous extracting the discriminant directions. Our SMDA can be cast as an convex programming which distinguishes itself from other state-of-the-art method. We evaluate the performances of the resulting methods on the extensive simulation study and a real data analysis.

Dedicated to my parents

Xiaozhe Sun and Caini Yong,

for their boundless love;

and to my sister,

Xin Sun,

for her support.

ACKNOWLEDGMENTS

First of all, I would like to take this opportunity to express my great appreciation to my advisors as well as my friends, Professor Hongtu Zhu and Professor Joseph G. Ibrahim. I am deeply indebted to them for their supervision, sweet encouragement, patience, and their deep insight into statistics, which have directly contributed to my general understanding of statistics, influenced my way of thinking about and performing scientific research and most importantly, the right attitude to life. It was Dr. Zhu and Dr. Ibrahim who taught me in person to build up my background and rigorous thinking in statistics through countless hours of face-to-face instructions as well as emails. Numerous scenarios pumped out of my head are that he was communicating with me editting papers I wrote till midnight; helping me go through every statement I attempted to make. But what they did is way beyond what I can mention here. They cares about students. They encouraged me to communicate with other scientists, provided me opportunities to know them in person, and helped me initiate collaborations with them. It can never be overstated that he is a leading and standard example of how to combine research of highest standard with a kind-hearted attitude.

I am deeply grateful to Professor Donglin Zeng and Professor Yufeng Liu for his invaluable advice, helpful comments, insightful discussion and writing important recommendation letters to support my job application. My sincere thanks also go to Professors Hongyu An and Michael Kosorok for joining my oral committee and providing many suggestions on my research. I wish to thank other faculty members in the biostatistics department and statistics department for their excellent lectures.

I would like to express my sincere appreciations to my friends and classmates: Baiguo An, Mihye Ahn, Guanhua Chen, Chao Huang, Jill Johnston, Yutao Ke, Dehan Kong, Linglong Kong, Joanne Lin, Lan Liu, Shangbang Rao, Yuying Xie, Liang Yin, Chong Zhang, Yingqi

Zhao, Ruoqing Zhu and many others for their helpful discussion on my research. Also thanks goes out to my buddies, Yutao Ke and Ming Gao, for their support and help.

Finally, I want to thank my parents, who made it possible for me to pursue the dream of becoming a statistician; my sister, who has been supporting me over the years.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER1: INTRODUCTION

In this paper document, we give some perspectives on high dimensional sparse regression and inference. We aim to build a framework of high dimensional inference, by considering three sub-topics. The first project is about hypothesis testing in multivariate linear regression with ultra high dimensional responses. The second project touches the fundamental framework of simultaneous variable selection and unbiased estimation in ultra-high dimensional space. The third part of this work concerns high dimensional multicategory classification problem where many traditional method fails due to the diverging spectra of sample covariance matrix and noise accumulation issue in high dimenional regime. We first start from literature review.

CHAPTER2: LITERATURE REVIEW

Traditionally, statistical inference considers a probability model for a population and considers data that arose as a sample from the population. For many problems, the estimates of the population characteristics, or parameters, can be substantially refined as the sample size n towards infinity with fixed number of unknown parameters p. Recently, researchers are interested in high dimensional statistical inference, when the number of unknown parameters p is much larger than the sample size n, that is $p \gg n$. This encompasses supervise regression and classification models where the number of covariates is of much larger order than n, unsupervised settings such as clustering or graphical modeling with more variables than observations or multiple testing where the number of considered testing hypotheses is larger than sample size. Such framework has become increasingly frequent and important in diverse fields of sciences, engineering, and humanities, ranging from genomics and health sciences to economics, finance and machine learning and characterizes many contemporary problems in statistics. For example, in imaging genetic studies between genotypes and phenotypes, hundreds of thousands of as single-nucleotide polymorphisms (SNPs) are considered as potential covariates for high dimensional imaging measures; in disease classification using microarray or proteomics data, tens of thousands of expression s of molecules are potential predictors. When interactions are considered, the dimensionality grows exponentially and result in ultra-high dimensionality, where ultra-high dimensionality refers to the case where the dimensionality grows at a non-polynomial rate as the sample size increases. Donoho et al. (2000) convincingly demonstrates the need for developments in high dimensional data analysis and presents the curses and blessings of dimensionality.

The high dimensionality poses challenges to statistical accuracy, model interpretability

and computational complexity, while in conventional studies, when the sample size n is much larger than the number of variables or parameters p, none of the three aspects needs to be sacrificed for the efficiency of others. However, traditional method fails due challenges posed by high dimensionality. We introduce the difficulties introduced by high dimensionality in the following.

A notorious difficulty of high dimensional model selection comes from the collinearity among the predictors, as pointed out by Fan and Lv (2008). The collinearity can easily be spurious in high dimensional geometry, which can make us select a wrong model and thus lead to completely wrong scientific conclusions. Statistically, this is due to the model identifiability issues in high dimensional framework.

Another well recognized issue for high dimensional statistical analysis goes to the noise accumulation problem both in statistics and computer science. The quantification of the impact of high dimensionality has been fully characterized both in regression and classification problems (Bühlmann and Geer 2011). The prediction error can be unbounded while the classification error can be as bad as random guessing due to noise accumulation in estimating the coefficient parameters and the population centroids respectively.

The philosophy that will generally rescue us, is to "believe" that in fact only a few, say s_0 of the unknown parameters are non-zero, namely the parameters are assumed to be sparse. With sparsity, variable selection can improve estimation accuracy and model interpretability by effectively identifying the subset of important predictors and thus achieving parsimonious representation.

Sparsity arises in many scientific endeavors. In genomic studies, it is generally believed that only a fraction of molecules are related to biological outcomes. For example, in disease classification, it is commonly believed that only one specific gene or tens of genes are responsible for the disease development. Selection tens of genes helps not only statisticians in constructing a more reliable classification rule, but also biologists to understand molecular mechanisms.

To study the theoretical property of high dimensional sparse regression and classification,

as pointed out in Fan and Li (2006), it is helpful to differentiate two types of statistical endeavors in high dimensional statistical learning: accuracy of estimated model parameters by controlling the risk bound and accuracy of the expected loss of the estimated model. The former is called consistency and appears in many contexts where we want to identify the significant predictors and characterize the precise contribution of each to the response variable. The latter property is called persistence in Greenshtein et al. (2004) and arises frequently in machine learning problems such as classification. More recently, Fan and Li (2001) has proposed the oracle property for high dimensional sparse regression by requiring the estimator identifying the true subset model and achieving the optimal estimation rate simultaneously.

Another important issue involves the estimation of a covariance matrix or its inverse (the precision matrix). Examples include portfolio management and risk assessment (Fan et al. 2008), high dimensional classification such as the Fisher discriminant (Hastie et al. 2009), graphic models (Meinshausen and Bühlmann 2006), statistical inference such as controlling false discoveries in multiple testing (Leek and Storey 2008, Efron 2010), finding quantitative trait loci based on longitudinal data (Yap et al. 2009, Xiong et al. 2011) and testing the capital asset pricing model (Sentana 2009), among others. Yet, the dimensionality is often either comparable with the sample size or even larger. In such cases, the sample covariance is known to have poor performance (Johnstone 2001), and some regularization is needed.

Realizing the importance of estimating large covariance matrices and the challenges that are brought by the high dimensionality, in recent years researchers have proposed various regularization techniques to consistently estimate Σ. One of the key assumptions is that the covariance matrix is sparse, namely many entries are 0 or nearly so (Bickel and Levina 2008b, Cai et al. 2010, Lam and Fan 2009, Rothman et al. 2009, Cai and Liu 2011a). Fan et al. (2013) further extends such framework to conditional sparsity by allowing the presence of common factors. This is useful in financial returns which depend on the equity market risks, housing prices which depend on the economic health and gene expressions, among many others.

The major contribution of this dissertation involves building a framework of high dimensional hypothesis test, a framework of simultaneous variable selection and estimation and a unified framework for sparse multicategory discriminant analysis. All of the projects involve incorporating covariance estimation into the regression framework to better trade off between noise accumulation and correlation modeling for possibility of relaxing conditions for consistent variable selection. We will separately introduce the the back ground in the following sections respectively.

2.0.1 Multivariate Regression and High Dimensional Test

Multivariate regression modeling with a multivariate response $\mathbf{y} \in \mathbb{R}^q$ and a multivariate covariate $\mathbf{x} \in \mathbb{R}^p$ is a standard statistical tool in modern high-dimensional inference, with wide applications in various large-scale applications, such as genome-wide association studies (GWAS) and neuroimaging studies. For instance, in GWAS, our primary problem of interest is to identify genetic variants ($\mathbf{x}$) that cause phenotypic variation ($\mathbf{y}$). Specifically, in imaging genetics, multivariate imaging measures ($\mathbf{y}$), such as volumes of regions of interest (ROIs), are phenotypic variables, whereas covariates ($\mathbf{x}$) include single nucleotide polymorphisms (SNPs), age, and gender, among others. The joint analysis of imaging and genetic data may ultimately lead to discoveries of genes for neuropsychiatric and neurological disorders such as autism and schizophrenia (Scharinger et al. 2010, Paus 2010, Peper et al. 2007, Chiang et al. 2011). Moreover, in many neuroimaging studies, there is a great interest in the use of imaging measures ($\mathbf{x}$), such as functional imaging data and cortical and subcortical structures, to predict multiple clinical and/or behavioral variables ($\mathbf{y}$) (Knickmeyer et al. 2008, Lenroot and Giedd 2006). This motivates us to systematically investigate a multivariate linear model with a multivariate response $\mathbf{y}$ and a multivariate covariate $\mathbf{x}$.

Throughout this paper, we consider n independent observations $(\mathbf{y}_i, \mathbf{x}_i)$ and a Multivariate Linear Model (MLM) given by

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \text{ or } \mathbf{y}_i = \mathbf{B}^T\mathbf{x}_i + \mathbf{e}_i, \tag{2.1}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, and $\mathbf{B} = (\beta_{jl})$ is a $p \times q$ coefficient matrix with $\text{rank}(\mathbf{B}) = r^\star \leq \min(p, q)$. Moreover, the error term $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^T$ has $E(\mathbf{e}_i) = 0$ and $\text{Cov}(\mathbf{e}_i) = \Sigma_R$ for all i, where Σ_R is a $q \times q$ matrix. Many hypothesis testing problems of interest, such as comparison across groups, can often be formulated as

$$H_0 : \mathbf{CB} = \mathbf{B}_0 \text{ v.s. } H_1 : \mathbf{CB} \neq \mathbf{B}_0, \tag{2.2}$$

where $\mathbf{C}$ is an $r \times p$ matrix and $\mathbf{B}_0$ is an $r \times q$ matrix. Without loss of generality, we center the covariates, standardize the responses, and assume $\text{rank}(\mathbf{C}) = r$.

We focus on a specific setting that q is relatively large, but p is relatively small. Such a setting is general enough to cover two-sample (or multi-sample) hypothesis testing for high-dimensional data (Chen and Qin 2010, Lopes et al. 2011). There are at least three major challenges including (i) a large number of regression parameters, (ii) a large covariance matrix, and (iii) correlations among multivariate responses. When the number of responses and the number of covariates are even moderately high, fitting the conventional MLM usually requires estimating a $p \times q$ matrix of regression coefficients, whose number pq can be much larger than n. Although accounting for complicated correlations among multiple responses is important for improving the overall prediction accuracy of multivariate analysis (Breiman and Friedman 1997, Cook et al. 2010), it requires estimating $q(q+1)/2$ unknown parameters in an unstructured covariance matrix.

There is a great interest in the development of efficient methods for handling MLMs with large q. Four popular traditional methods include the mass univariate analysis, the Hotelling's T^2 test, partial least squares regression, and dimension reduction methods. As pointed by Klei et al. (2008) and many others, testing each response variable individually in the mass univariate analysis requires a substantial penalty of controlling for multiplicity. The Hotelling's T^2 test is not well-defined, when $q > n$. Even when $q \leq n$, the power of the Hotelling's T^2 can be very low if q is nearly as large as n. Partial least squares regression (PLSR) aims to find a linear regression model by projecting $\mathbf{y}$ and $\mathbf{x}$ to a smaller latent space

(Chun and Keles 2010, Krishnan et al. 2011), but it focuses on prediction and classification. Although dimension reduction techniques, such as principal component analysis (PCA), are considered to reduce the dimensions of both the response and covariates (Formisano et al. 2008, Kherif et al. 2002, ROWE and Hoffmann 2006, Teipel et al. 2007), most of the methods ignore the variation of covariates and their associations with responses. Thus, such methods can be sub-optimal for our problem.

Some recent developments primarily include regularization methods and envelope models (Peng et al. 2010, Tibshirani 1996, Breiman and Friedman 1997, Cook et al. 2010, Cook, R. D., Helland, I. S. and Su 2013, Lin et al. 2012). Cook, Li and Chiaromonte (2010) developed a powerful envelope modeling framework for MLMs. Such envelope methods use dimension reduction techniques to remove the immaterial information, while achieving efficient estimation of the regression coefficients by accounting for correlations among the response variables. However, the existing envelope methods are limited to the $n > \max(p, q)$ scenario. Recently, much attention has been given to regularization methods for enforcing sparsity in $\mathbf{B}$ (Peng et al. 2010, Tibshirani 1996). These regularization methods, however, do not provide a standard inference tool (e.g., standard deviation) on the regression coefficient matrix $\mathbf{B}$. Lin et al. (2012) developed a projection regression model (PRM) and its associated estimation procedure to assess the relationship between a multivariate phenotype and a set of covariates without providing any theoretical justification.

In this dissertation, we present a new general framework, called sparse projection regression model (SPReM), for simultaneously performing dimension reduction, response selection, estimation, and testing in a general high dimensional MLM setting. We introduce two novel heritability ratios, which extend the idea of principal components of heritability from familial studies (Klei et al. 2008, Ott and Rabinowitz 1999), for MLM and overcome over-fitting and noise accumulation in high dimensional data by enforcing the sparsity constraint. We develop a fast algorithm for both **s**parse **u**nit **r**ank **p**rojection (SURP) and **s**parse **mu**lti-**r**ank **p**rojection (SMURP). Furthermore, a test procedure based on the wild-bootstrap method is proposed, which leads to a single $p-$value for the test of an association between all response

variables and covariates of interest, such as genetic markers. Simulations show that our method can control the overall Type I error well, while achieving high statistical power.

2.0.2 High Dimensional Sparse Regression

Consider the linear model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.3}$$

wher y_i is a univariate response, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^T$ is a $p-$dimensional covariate vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ regression coefficient vector, and $\{\epsilon_i : i = 1, \ldots, n\}$ are independent and identically distributed (i.i.d) errors. The theory of linear models is well established for traditional applications, where the dimension p is fixed and the sample size n is much larger than p. With the development of many modern technologies, however, in many biological, medical, social, and economical studies, p is comparable with, or much larger than, n, making valid statistical inferences a great challenge. Let $\mathcal{A}$ be subset of indices such that $\mathcal{A} = \{j | \beta_j^o \neq 0\}$ and p_A be the cardinality of $\mathcal{A}$, where $\boldsymbol{\beta}^o = (\beta_1^o, \ldots, \beta_p^o)^T$ is the true parameter $\boldsymbol{\beta}$. For prediction accuracy and variable selection consistency, it is common to assume a sparsity assumption, that is, $p_A << p$.

For model (2.3), many regularization methods for variable selection minimize

$$Q(\boldsymbol{\beta}) = \frac{1}{2n} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \sum_{j=1}^{p} p_\lambda(\beta_j), \tag{2.4}$$

where $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\mathbf{X}$ is an $n \times p$ non-stochastic matrix with the ith row $\mathbf{x}_i^T$, $|| \cdot ||_2$ represents the L_2 norm, and $p_\lambda(\cdot)$ is a penalty function (e.g., SCAD or Lasso), which depends on a tuning parameter $\lambda > 0$. The most well-known best subset selection corresponding to the L_0 penalty function can achieve simultaneous parameter estimation and variable selection (Akaike 1973, Schwarz 1978). The subset selection methods coupled with different selection criteria-including the C_p statistics, the Akaike information criterion (AIC), the Bayesian information criterion (BIC), minimum description length (MDL), and the risk

inflation criterion (RIC) are special cases of the L_0 penalized regression, resulting from the assignment of different values to λ. However, solving the L_0 regularization with a fixed λ is an NP-hard problem and its computational methods based on exhaustive search rapidly become impractical when the number of covariates increases (Huo and Ni 2007, Fan and Peng 2004, Fan and Li 2001, Zhang 2010). To address such computational issue, different convex/nonconvex penalty functions have been used in $Q(\boldsymbol{\beta})$ and been extensively investigated in order to mimic the L_0 regularization (Tibshirani 1996, Fan and Li 2001, Fan and Peng 2004, Zhang 2010, Meinshausen and Bühlmann 2006, Leng et al. 2004, Zou 2006).

Instead of developing another penalty function, we develop a new hard thresholded regression (HTR) modeling framework for performing simultaneous variable selection and unbiased estimation in model (2.3) in this dissertation. The key idea of HTR is to minimize

$$H(\boldsymbol{\beta}) = ||\mathbf{W} \times \mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})||_1 + \lambda||\boldsymbol{\beta}||_1, \tag{2.5}$$

where $\mathbf{W}$ is a $p_0 \times p$ weighted matrix based on some initial estimates of $\boldsymbol{\beta}$, which will be introduced in Section 2. As shown in Sections 2 and 3, HTR simultaneously enjoys two key computational and theoretical properties as follows.

- (i) Since $H(\boldsymbol{\beta})$ is convex and HTR can be casted as a linear program, minimizing $H(\boldsymbol{\beta})$ is computationally efficient even in high-dimensional settings.
- (ii) Under some mild conditions, the HTR estimate, which minimizes $H(\boldsymbol{\beta})$, is an oracle estimator and achieves unbiased estimation.

Due to its nice properties (i) and (ii), our HTR estimate may be a good addition to the extensive regularization literature.

Our HTR shares some important similarities with the regularization methods in (2.5). The penalty function of $H(\boldsymbol{\beta})$ is the same as that of the popular Lasso (Tibshirani 1996), when $p_\lambda(\beta_j) = \lambda|\beta_j|$. As shown in Section 2, HTR has a close connection with the L_0 and hard-thresholding regularizations (Akaike 1973, Schwarz 1978, Zheng et al. 2013), since all of them reduce to the best subset selection under the orthonormal design, that is, $n^{-1}\mathbf{X}^T\mathbf{X} =$

$\mathbf{I}_p$, where $\mathbf{I}_p$ is a $p \times p$ identity matrix. A comparison of the regularization path between the L_0 regularization regression and HTR is shown in Figure 2.1.

Our HTR differs significantly from the regularization methods (2.5) in several major ways. A major advantage of HTR over nonconvex regularizations is its computational efficiency (i), even though they may enjoy nice theoretical properties, such as oracle property (Barron et al. 1999, Lin et al. 2008). Although there are many impressive works on nonconvex regularization methods (Wang et al. 2013a, Kim and Kwon 2012, Zhang and Zhang 2012, Fan and Lv 2011, Kim et al. 2008, Wang et al. 2013b), several important questions still remain. Specifically, due to the non-convexity of the penalty function, multiple local minima always exist, while it is difficulty to identify the oracle estimator among multiple minima, even if the oracle estimator may be known to exist along the solution path.

A major advantage of HTR over convex regularization methods is its nice theoretical property (ii). Due to the convexity of the penalty function, convex regularization methods, such as Lasso, suffer from the bias issue and thus they can be suboptimal in terms of risk estimation. See Fan and Li (2001) for detailed discussions. Moreover, the shrinkage bias introduced by convex regularization methods poses major challenges to statistical inferences, such as constructing confidence intervals or testing, in high dimensional settings (Zhang and Zhang 2011, van de Geer et al. 2013, Chatterjee and Lahiri 2011). There is a major conflict of optimal prediction and consistent variable selection in the lasso method (Meinshausen and Bühlmann 2006, Leng et al. 2004, Zou 2006).

We make three major contributions in this part as follows.

- We systematically investigate a fast two-step estimation procedure for HTR. The first step is to calculate a ridge estimator and the second step is to solve a linear programming.

- We provide a comprehensive theoretical analysis of HTR. We show that the HTR estimator enjoys the strong oracle property even when the number of covariates may grow at an exponential rate.

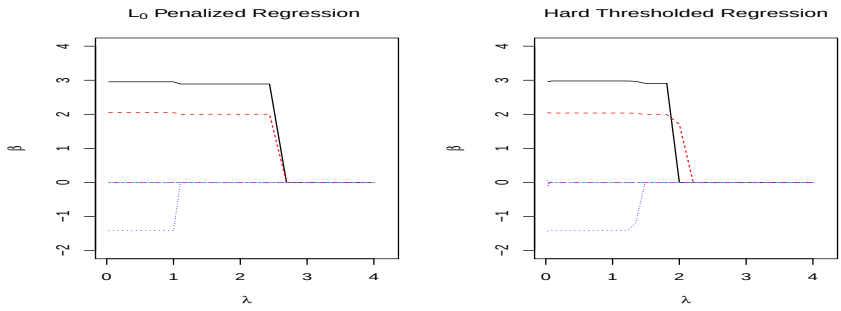

Figure 2.1: Solution paths of L_0 regularization regression and HTR: We consider a simple example that $y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$, where $\boldsymbol{\beta} = (3, 2, -1.5, 0, 0, 0)^T$ and ε_i's are independently and identically distributed as $N(0, 1)$. We plot the estimates of regression coefficients $\hat{\beta}_j, j = 1, 2, \ldots, 6$ for this example. *Left Panel* L_0 Penalized Regression estimates, as a function of λ; *Right Panel* Hard Thresholded Regression estimates, as a function of λ.

- We propose to incorporate the regularized covariance estimator into the estimation procedure in order to better trade off between noise accumulation and correlation modeling.

2.0.3 Sparse Multicategory Discriminant Analysis

Discriminant analysis is widely used in classification problems. Fisher's linear discriminant analysis is proposed by R.A. Fisher, and has been successfully used in the machine learning literature. Nowadays, high throughput data from microarray and proteomics technologies has been frequently used in many contemporary statistical studies. In the case of microarray data, the dimensionality is frequently in thousands, whereas the sample size is typically only order of tens. The large p small n case poses challenges for classification problems.

When the feature space dimension p is far more larger than the sample size n, the Fisher's linear discriminant rule fails owing to diverging spectra as demonstrated by Bickel and Levina (2008b), who showed that the independence rule in which the correlation structure is ignored performs better than the naive bayes rule. However, in my data analysis, for example, the microarray studies, correlation structure can is an essential characteristic of the data and is usually not negligible. To circumvent this issue, Fan et al. (2012) proposes the regularized optimal affine discriminant (ROAD) method. Their method focuses on the binary classification problem, where, on the other hand, many real world problems have more than two classes to deal with. Typical examples include text catergorization and microarray data analysis, etc. Witten and Tibshirani (2011) proposes the penalized LDA and extend the framework to multicategoty problem by using a sequential approach. However, their problem is not convex and is extremely computational expensive.

In this paper we propose a unified approach, called sparse multicategory discriminant analysis (SMDA), which enjoys following attractive properties.

- It reduces to penalized version of the ROAD estimator when there are only two classes.
- It results a fast convex programming algorithm comparing to the penalized LDA framework proposed by Witten and Tibshirani (2011).
- Interpretable discriminant directions are produced owing to the penalized penalty.

This dissertation is organized as follows. We present Sparse Projection Regression Model in chapter 2 . Chapter 3 is contributed to the Hard Thresholded Regression which can be cast as linear programming. Sparse Multicategory Discriminant Analysis is discussed in Chapter 4. Conclusions and discussions are touched in Chapter 5.

CHAPTER3: SPARSE PROJECTION REGRESSION MODEL

We develop a Sparse Projection Regression Model (SPReM) framework to perform multivariate regression modeling with a large number of responses and a multivariate covariate of interest. We propose two novel heritability ratios to simultaneously perform dimension reduction, response selection, estimation, and testing, while explicitly accounting for correlations among multivariate responses. Our SPReM is devised to specifically address the low statistical power issue of many standard statistical approaches, such as the Hotelling's T^2 test statistic or a mass univariate analysis, for high-dimensional data. We formulate the estimation problem of SPReM as a novel sparse unit rank projection (SURP) problem and propose a fast optimization algorithm for SURP. Furthermore, we extend SURP to the sparse multi-rank projection (SMURP) by adopting a sequential SURP approximation. Theoretically, we have systematically investigated the convergence properties of SURP and the convergence rate of SURP estimates. Our simulation results and real data analysis have shown that SPReM outperforms other state-of-the-art methods.

3.1 Model Setup and Heritability Ratios

We introduce SPReM as follows. The key idea of our SPReM is to appropriately project $\mathbf{y}_i$ in a high-dimensional space onto a low-dimensional space, while accounting for the correlation structure Σ_R among the response variables and the hypothesis test in (2.2). Let $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_k]$ be a $q \times k$ nonrandom and unknown direction matrix, where $\mathbf{w}_j$ are $q \times 1$ vectors. A projection regression model (PRM) is given by

$$\mathbf{W}^T\mathbf{y}_i = (\mathbf{BW})^T\mathbf{x}_i + \mathbf{W}^T\mathbf{e}_i = \boldsymbol{\beta}_{\mathbf{w}}^T\mathbf{x}_i + \boldsymbol{\varepsilon}_i, \tag{3.1}$$

where $\boldsymbol{\beta}_{\mathbf{w}}$ is a $p \times k$ regression coefficient matrix and the random vector $\boldsymbol{\varepsilon}_i$ has $E(\boldsymbol{\varepsilon}_i) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}_i) = \mathbf{W}^T \Sigma_R \mathbf{W}$. When $k = 1$, PRM reduces to the pseduo-trait model considered in (Amos et al. 1990, Amos and Laing 1993, Klei et al. 2008, Ott and Rabinowitz 1999). If $k << \min(n, q)$ and $\mathbf{W}$ were known, then one could use likelihood (or estimating equation) based methods to efficiently estimate $\boldsymbol{\beta}_{\mathbf{w}}$, and (2.2) would reduce approximately to

$$H_{0W} : \mathbf{C}\boldsymbol{\beta}_{\mathbf{w}} = \mathbf{b}_0 \text{ v.s. } H_{1W} : \mathbf{C}\boldsymbol{\beta}_{\mathbf{w}} \neq \mathbf{b}_0, \tag{3.2}$$

where $\mathbf{C}\boldsymbol{\beta}_{\mathbf{w}} = \mathbf{CBW}$ and $\mathbf{b}_0 = \mathbf{B}_0\mathbf{W}$. In this case, the number of null hypotheses in (3.2) is much smaller than that of (2.2). It is also expected that different $\mathbf{W}$'s strongly influence the statistical power of testing the hypotheses in (2.2).

A fundamental question arises

"how do we determine an 'optimal' $\mathbf{W}$ to achieve good statistical power of testing (2.2)?"

To determine $\mathbf{W}$, we develop a novel deflation approach to sequentially determine each column of $\mathbf{W}$ at a time starting from $\mathbf{w}_1$ to $\mathbf{w}_k$. We focus on how to determine $\mathbf{w}_1$ below and then discuss how to extend it to the scenario with $k > 1$.

To determine an optimal $\mathbf{w}_1$, we consider two principles. The first principle is to maximize the mean value of the square of the signal-to-noise ratio, called the heritability ratio, for model (3.1). For each i, the signal-to-noise ratio in model (3.1) is defined as the ratio of mean to standard deviation of a signal or measurement $\mathbf{w}^T\mathbf{y}_i$, denoted by $\mathrm{SNR}_i = \mathbf{w}^T\mathbf{B}^T\mathbf{x}_i/(\mathbf{w}^T\Sigma_R\mathbf{w})^{0.5}$. Thus, the heritability ratio (HR) is given by

$$\mathrm{HR}(\mathbf{w}) = n^{-1}\sum_{i=1}^{n} \mathrm{SNR}_i^2 = \frac{\mathbf{w}^T\mathbf{B}^T S_X \mathbf{B}\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}, \tag{3.3}$$

where $S_X = n^{-1}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T$. The HR has several important interpretations. If the $\mathbf{x}_i$ are independently and identically distributed (i.i.d) with $E(\mathbf{x}_i) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{x}_i) = \Sigma_X$, then as

$n \to \infty$, we have

$$\mathrm{HR}(\mathbf{w}) \to^p \frac{\mathbf{w}^T\mathbf{B}^T\Sigma_X\mathbf{B}\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}} = \frac{\mathrm{Var}(\mathbf{w}^T\mathbf{B}^T\mathbf{x}_i)}{\mathrm{Var}(\boldsymbol{\varepsilon}_i)},$$

where $\to^p$ denotes convergence in probability. Thus, $\mathrm{HR}(\mathbf{w})$ is close to the ratio of the variance of signal $\mathbf{w}^T\mathbf{B}^T\mathbf{x}_i$ to that of noise $\boldsymbol{\varepsilon}_i$. Moreover, $\mathrm{HR}(\mathbf{w})$ is close to the heritability ratio considered in (Amos et al. 1990, Amos and Laing 1993, Klei et al. 2008, Ott and Rabinowitz 1999) for familial studies, but we define HR from a totally different perspective. With such new perspective, one can easily define HR for more general designs, such as cross-sectional or longitudinal design. One might directly maximize $\mathrm{HR}(\mathbf{w})$ to calculate an 'optimal' $\mathbf{w}_1$, but such a $\mathbf{w}_1$ can be sub-optimal for testing the hypotheses in (2.2) as discussed below.

The second principle is to explicitly account for the hypotheses in (2.2) under model (2.1) and the reduced ones in (3.2) under model (3.1). We define four spaces associated with the null and alternative hypotheses of (2.2) and (3.2) as follows:

$$S_{H_0} = \{\mathbf{B} : \mathbf{CB} = \mathbf{B}_0\},\ S_{H_W} = \{\mathbf{B} : \mathbf{CBW} = \mathbf{B}_0\mathbf{W}\},$$
$$S_{H_1} = \{\mathbf{B} : \mathbf{CB} \neq \mathbf{B}_0\},\ S_{H_{1W}} = \{\mathbf{B} : \mathbf{CBW} \neq \mathbf{B}_0\mathbf{W}\}.$$

It can be shown that they satisfy the following relationship:

$$S_{H_0} \subset S_{H_W} \text{ and } S_{H_{1W}} \subset S_{H_1} \text{ for any } \mathbf{W} \neq \mathbf{0}.$$

Due to potential information loss during dimension reduction, both $S_{H_W} - S_{H_0}$ and $S_{H_1} - S_{H_{1W}}$ may not be the empty set, but we need to choose $\mathbf{W}$ such that $S_{H_1} - S_{H_{1W}} \approx \emptyset$. The next question is how to achieve this.

We consider a data transformation procedure. Let $\mathbf{C}_1$ be a $(p-r) \times p$ matrix such that

$$\mathrm{rank}[\mathbf{C}^T\ \mathbf{C}_1^T] = p \text{ and } \mathbf{C}\mathbf{C}_1^T = \mathbf{0}. \tag{3.4}$$

Let $\mathbf{D} = [\mathbf{C}^T \ \mathbf{C}_1^T]^T$ be a $p \times p$ matrix and $\tilde{\mathbf{x}}_i = (\tilde{\mathbf{x}}_{i1}^T, \tilde{\mathbf{x}}_{i2}^T) = \mathbf{D}^{-T}\mathbf{x}_i$ be a $p \times 1$ vector, where $\tilde{\mathbf{x}}_{i1}$ and $\tilde{\mathbf{x}}_{i2}$ are, respectively, the $r \times 1$ and $(p-r) \times 1$ subvectors of $\tilde{\mathbf{x}}_i$. We define $\tilde{\mathbf{B}} = [\tilde{\mathbf{B}}_1^T \ \tilde{\mathbf{B}}_2^T]^T = \mathbf{DB}$ or $\mathbf{B} = \mathbf{D}^{-1}\tilde{\mathbf{B}}$, where $\tilde{\mathbf{B}}_1$ and $\tilde{\mathbf{B}}_2$ are, respectively, the first r rows and the last $p-r$ rows of $\tilde{\mathbf{B}}$. Therefore, model (3.1) can be rewritten as

$$\begin{aligned} \mathbf{W}^T\mathbf{y}_i &= (\mathbf{D}^{-1}\tilde{\mathbf{B}}\mathbf{W})^T\mathbf{x}_i + \mathbf{W}^T\mathbf{e}_i \\ &= \mathbf{W}^T(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T\tilde{\mathbf{x}}_{i1} + \mathbf{W}^T\mathbf{B}_0^T\tilde{\mathbf{x}}_{i1} + \mathbf{W}^T\tilde{\mathbf{B}}_2^T\tilde{\mathbf{x}}_{i2} + \mathbf{W}^T\mathbf{e}_i. \end{aligned} \tag{3.5}$$

In (3.5), due to (3.4), we only need to consider the transformed covariate vector $\tilde{\mathbf{x}}_{i1}$, which contains useful information associated with $\tilde{\mathbf{B}}_1 - \mathbf{B}_0 = \mathbf{CB} - \mathbf{B}_0$.

We define a generalized heritability ratio based on model (3.5). Specifically, for each i, we define a new signal-to-noise ratio as the ratio of mean to standard deviation of signal $\mathbf{w}^T(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T\tilde{\mathbf{x}}_{i1} + \mathbf{w}^T\mathbf{e}_i$, denoted by $\mathrm{SNR}_{i,\mathbf{C}} = \mathbf{w}^T(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T\tilde{\mathbf{x}}_{i1}/(\mathbf{w}^T\Sigma_R\mathbf{w})^{0.5}$. The generalized heritability ratio is then defined as

$$\mathrm{GHR}(\mathbf{w};\mathbf{C}) = n^{-1}\sum_{i=1}^{n}\mathrm{SNR}_{i,\mathbf{C}}^2 = \frac{\mathbf{w}^T(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T S_{\tilde{X}_1}(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}, \tag{3.6}$$

where $S_{\tilde{X}_1} = n^{-1}\sum_{i=1}^{n}\tilde{\mathbf{x}}_{i1}\tilde{\mathbf{x}}_{i1}^T$. If the $\mathbf{x}_i$s are random, then we have

$$\mathrm{GHR}(\mathbf{w};\mathbf{C}) \to^p \frac{\mathbf{w}^T(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T\mathrm{Cov}(\tilde{\mathbf{x}}_{i1})(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}} = \frac{\mathbf{w}^T\Sigma_C\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}, \tag{3.7}$$

where $\Sigma_C = (\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T(\mathbf{D}^{-T}\Sigma_X\mathbf{D}^{-1})_{(r,r)}(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)$, and $(\mathbf{D}^{-T}\Sigma_X\mathbf{D}^{-1})_{(r,r)}$ is the upper $r \times r$ submatrix of $\mathbf{D}^{-T}\Sigma_X\mathbf{D}^{-1}$. Particularly, if $\mathbf{C} = [\mathbf{I}_r \ \mathbf{0}]$, then Σ_C reduces to $\mathbf{w}^T(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T(\Sigma_X)_{(1,1)}(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)\mathbf{w}$, in which $(\Sigma_X)_{(1,1)}$ is the upper $r \times r$ submatrix of Σ_X. Thus, $\mathrm{GHR}(\mathbf{w};\mathbf{C})$ can be interpreted as the ratio of the variance of $\mathbf{w}^T(\tilde{\mathbf{B}}_1 - \mathbf{B}_0)^T\tilde{\mathbf{x}}_{i1}$ relative to that of $\mathbf{w}^T\mathbf{e}_i$. We propose to calculate an optimal $\mathbf{w}_*$ as follows:

$$\mathbf{w}_* = \underset{\mathbf{w}}{\mathrm{argmax}}\,\mathrm{GHR}(\mathbf{w};\mathbf{C}). \tag{3.8}$$

We expect that such an optimal $\mathbf{w}_*$ can substantially reduce the size of both $S_{H_1} - S_{H_{1W}}$ and $S_{H_W} - S_{H_0}$ and thus the use of such an optimal $\mathbf{w}_*$ can enhance the power of testing the hypotheses in (2.2). Without loss of generality, we assume $\mathbf{B}_0 = \mathbf{0}$ from now on.

We consider a simple example to illustrate the appealing properties of $\mathrm{GHR}(\mathbf{w}; \mathbf{C})$.

Example We consider model (2.1) with $p = q = 5$ and want to test the nonzero effect of the first covariate on all five responses. In this case, $r = 1$, $\mathbf{C} = (1, 0, 0, 0, 0)$, $\mathbf{B}_0 = (0, 0, 0, 0, 0)$, and $\mathbf{D} = I_5$, which is a 5×5 identity matrix. Without loss of generality, it is assumed that $(\Sigma_X)_{(1,1)} = 1$.

We consider three different cases of Σ_R and $\mathbf{B}$. In the first case, we set $\Sigma_R = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_5^2)$ and the first column of $\mathbf{B}$ to be $(1, 0, 0, 0, 0)$. It follows from (3.6) that

$$\mathrm{GHR}(\mathbf{w}; \mathbf{C}) = \frac{w_1^2}{\sigma_1^2 w_1^2 + \sigma_2^2 w_2^2 + \cdots + \sigma_5^2 w_5^2} \text{ and } \mathbf{w}_*^T = (c_0, 0, 0, 0, 0),$$

where c_0 is any nonzero scalar. Therefore, $\mathbf{w}_*$ picks out the first response, which is the sole one that is associated with the first covariate.

In the second case, we set $\Sigma_R = \mathrm{diag}(\sigma_1^2, \cdots, \sigma_5^2)$ with $\sigma_1^2 \geq \cdots \geq \sigma_5^2$ and the first row of $\mathbf{B}$ to be $(1, 1, 0, 0, 0)$. It follows from (3.6) that

$$\mathrm{GHR}(\mathbf{w}; \mathbf{C}) = \frac{(w_1 + w_2)^2}{\sigma_1^2 w_1^2 + \sigma_2^2 w_2^2 + \cdots + \sigma_5^2 w_5^2} \text{ and } \mathbf{w}_*^T = (\frac{\sigma_2^2}{\sigma_1^2} c_0, c_0, 0, 0, 0),$$

where c_0 is any nonzero scalar. Therefore, $\mathbf{w}_*$ picks out both the first and second response with larger weight on the second component. This is desirable since β_{11} and β_{21} are equal in terms of strength of effect and the noise level for the second response is smaller than that of the first one.

In the third case, we set the first row of $\mathbf{B}$ to be $(1, 1, 0, 0, 0)$ and the first and second columns of Σ_R are set as $\sigma^2(1, \rho, 0, 0, 0)$ and $\sigma^2(\rho, 1, 0, 0, 0)$, respectively. It follows from

(3.6) that

$$\mathrm{GHR}(\mathbf{w};\mathbf{C}) = \frac{(w_1+w_2)^2}{\sigma^2 w_1^2 + 2\sigma_2^2\rho w_1 w_2 + \sigma_2^2 w_2^2 + Q(w_3,w_4,w_5)} \text{ and } \mathbf{w}_*^T = (c_0, c_0, 0, 0, 0),$$

where $Q(w_3, w_4, w_5)$ is a non-negative quadratic form of (w_3, w_4, w_5). Thus, the optimal $\mathbf{w}_*$ chooses the first two responses with equal weight, since they are correlated with each other with same variance and $\beta_{11} = \beta_{21} = 1$.

For high dimensional data, it is difficult to accurately estimate $\mathbf{w}_*$, since the sample covariance matrix estimator $\hat{\Sigma}_R$ can be either ill-conditioned or not invertible for large $q > n$. One possible solution is to focus only on a small number of important features for testing. However, a naive search for the best subset is NP-hard. We develop a penalized procedure to address these two problems, while obtaining a relatively accurate estimate of $\mathbf{w}$. Let $\tilde{\Sigma}_R$ and $\hat{\Sigma}_C$ be, respectively, estimators of Σ_R and Σ_C. Here we use $\tilde{\Sigma}_R$ to denote the covariance estimator other than sample covariance matrix $\hat{\Sigma}_R$. To obtain $\hat{\Sigma}_C$, we need to plug $\widehat{\mathbf{B}}$, an estimator of $\mathbf{B}$, into Σ_C. Without loss of generality, we consider the ordinary least squares estimate of $\mathbf{B}$. By imposing a sparse structure on $\mathbf{w}_1$, we recast the optimization problem as

$$\max\{\frac{\mathbf{w}^T\hat{\Sigma}_C\mathbf{w}}{\mathbf{w}^T\tilde{\Sigma}_R\mathbf{w}}\} \quad \text{s.t. } ||\mathbf{w}||_1 \le t, \tag{3.9}$$

where $||\cdot||_1$ is the L_1 norm and $t > 0$.

3.1.1 Sparse Unit Rank Projection

When $r = 1$, we call the problem in (3.8) as the unit rank projection problem and its corresponding sparse version in (3.9) as the sparse unit rank projection (SURP) problem. Actually, many statistical problems, such as two-sample test and marginal effect test problems, can be formulated as the unit rank projection problem (Lopes et al. 2011). We consider two cases including $\boldsymbol{\ell} = (\mathbf{CB})^T = \mathbf{0}$ and $\boldsymbol{\ell} = (\mathbf{CB})^T \neq \mathbf{0}$. When $\boldsymbol{\ell} = (\mathbf{CB})^T = \mathbf{0}$, the solution set of (3.6) is trivial, since any $\mathbf{w} \neq \mathbf{0}$ is a solution of (3.6). As discussed later, this property is extremely important for controlling the type I error rate.

When $\boldsymbol{\ell} = (\mathbf{CB})^T \neq \mathbf{0}$, (3.6) reduces to the following optimization problem:

$$\mathbf{w}_* = \underset{\mathbf{w}^T\Sigma_R\mathbf{w}=1}{\operatorname{argmax}} \mathbf{w}^T\Sigma_C\mathbf{w} = \underset{\mathbf{w}^T\Sigma_R\mathbf{w}\leq 1}{\operatorname{argmax}} \mathbf{w}^T\Sigma_C\mathbf{w} = \underset{\mathbf{w}^T\Sigma_R\mathbf{w}\leq 1}{\operatorname{argmax}} \mathbf{w}^T\boldsymbol{\ell}, \tag{3.10}$$

where $\boldsymbol{\ell}$ is the sole eigenvector of Σ_C, since Σ_C is a unit-rank matrix. To impose an L_1 sparsity on $\mathbf{w}$, we propose to solve the penalized version of (3.10) given by

$$\mathbf{w}_\lambda = \underset{\mathbf{w}^T\Sigma_R\mathbf{w}\leq 1}{\operatorname{argmax}} \mathbf{w}^T\boldsymbol{\ell} - \lambda||\mathbf{w}||_1. \tag{3.11}$$

Although (3.11) can be solved by using some standard convex programming methods, such methods are too slow for most large-scale applications, such as imaging genetics. We therefore reformulate our problem below. Without special saying, we focus on $\boldsymbol{\ell} = (\mathbf{CB})^T \neq \mathbf{0}$. By omitting a scaling factor $||\Sigma_R^{-1/2}\boldsymbol{\ell}||_2$, which will not affect the generalized heritability ratio, we note that (3.10) is equivalent to the following

$$\mathbf{w}_0 = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} - \mathbf{w}^T\boldsymbol{\ell}. \tag{3.12}$$

We consider a penalized version of (5.12) as

$$\mathbf{w}_{0,\lambda} = \underset{\mathbf{w}}{\operatorname{argmin}} f(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} - \mathbf{w}^T\boldsymbol{\ell} + \lambda||\mathbf{w}||_1. \tag{3.13}$$

A nice property of (5.13) is that it does not explicitly involve the inequality constraint, which leads to a fast computation. We define (5.12) as the oracle, since $\mathbf{w}_\lambda$ converges to $\mathbf{w}_0$ as $\lambda \to 0$. It can be shown that

$$\mathbf{w}_0 = \Sigma_R^{-1}\boldsymbol{\ell}. \tag{3.14}$$

We obtain an equivalence between (5.13) and (3.11) as follows.

Theorem 3.1.1 *Problem (5.13) is equivalent to problem (3.11) and* $\mathbf{w}_\lambda \propto \mathbf{w}_{0,\lambda}$.

We discuss some connections between our SURP problem and the optimization problem considered in Fan et al. (2012) for performing classification in high dimensional space. However, rather than recasting the problem as in (3.10) and then (5.13), they formulate it as

$$\mathbf{w}_c = \underset{||\mathbf{w}||_1 \leq c, \mathbf{w}^T\boldsymbol{\ell}=1}{\operatorname{argmin}} \mathbf{w}^T\Sigma_R\mathbf{w},$$

which can further be reformulated as

$$\mathbf{w}_\lambda = \underset{\mathbf{w}^T\boldsymbol{\ell}=1}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} + \lambda||\mathbf{w}||_1. \tag{3.15}$$

Since (5.14) involves a linear equality constraint, they replace it by a quadratic penalty as

$$\mathbf{w}_{\lambda,\gamma} = \operatorname{argmin} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} + \lambda||\mathbf{w}||_1 + \frac{1}{2}\gamma(\mathbf{w}^T\boldsymbol{\ell} - 1)^2. \tag{3.16}$$

This new formulation requires the simultaneously tuning of λ and γ, which can be computationally intensive. However, in Fan et al. (2012), they stated that the solution to (5.15) is not sensitive to γ, since solution is always in the direction of $\Sigma_R^{-1}\boldsymbol{\ell}$ when $\lambda = 0$, as validated by simulations. Their formulation (5.14) is close to the formulation (5.13). This result sheds some light on why $\mathbf{w}_{\lambda,\gamma}$ is not sensitive to γ. Finally, we can show that the solution path to (5.13) has a piecewise linear property.

Proposition 3.1.2 *Let $\boldsymbol{\ell} \in \mathbb{R}^q$ be a constant vector and Σ_R be positive definite. Then, $\mathbf{w}_{0,\lambda}$ is a continuous piecewise linear function in λ.*

We derive a coordinate descent algorithm to solve (5.13). Without loss of generality, suppose that $\mathbf{w} = (\tilde{w}_1, \tilde{\mathbf{w}}_2^T)^T = (\tilde{w}_1, \cdots, \tilde{w}_q)^T$, $\tilde{w}_j$ for all $j \geq 2$ are given, and we need to optimize (5.13) with respect to $\tilde{w}_1$. In this case, the objective function (5.13) becomes

$$f_1(\tilde{w}_1, \tilde{\mathbf{w}}_2) = \frac{1}{2}(\tilde{w}_1, \tilde{\mathbf{w}}_2^T)\begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} \tilde{w}_1 \\ \tilde{\mathbf{w}}_2 \end{pmatrix} - (\tilde{\ell}_1\tilde{w}_1 + \tilde{\boldsymbol{\ell}}_2^T\tilde{\mathbf{w}}_2) + \lambda|\tilde{w}_1| + \lambda||\tilde{\mathbf{w}}_2||_1,$$

where $\boldsymbol{\ell} = (\tilde{\ell}_1, \tilde{\boldsymbol{\ell}}_2^T)$ and σ_{11}, Σ_{12}, and Σ_{22} are subcomponents of Σ_R. Then, by taking the sub-gradient with respect to $\tilde{w}_1$, we have

$$f_1'(\tilde{w}_1, \tilde{\mathbf{w}}_2) = \tilde{w}_1\sigma_{11} + \Sigma_{12}\tilde{\mathbf{w}}_2 + \lambda\Gamma_1 - \tilde{\ell}_1$$

where $\Gamma_1 = \text{sign}(\tilde{w}_1)$ for $\tilde{w}_1 \neq 0$ and is between -1 and 1 if $\tilde{w}_1 = 0$. Let $S_\lambda(t) = \text{sign}(t)(|t| - \lambda)^+$ be the soft-thresholding operator. By setting $f_1'(\tilde{w}_1, \tilde{\mathbf{w}}_2) = 0$, we have $\tilde{w}_1 = S_\lambda(\tilde{\ell}_1 - \Sigma_{12}\tilde{\mathbf{w}}_2)/\sigma_{11}$. Based on this result, we can obtain a coordinate descent algorithm as follows.

Algorithm

(a) Initialize $\mathbf{w}$ at a starting point $\mathbf{w}^{(0)}$ and set $m = 0$.

(b) Repeat:

- (b.1) Increase m by 1: $m \leftarrow m + 1$
- (b.2) for $j \in 1, \cdots, p$, if $\tilde{w}_j^{(m-1)} = 0$, then set $\tilde{w}_j^{(m)}$=0;
  otherwise: $\tilde{w}_j^{(m)} = \text{argmin}\, f(\tilde{w}_1^{(m)}, \cdots, \tilde{w}_{j-1}^{(m)}, \tilde{w}_j, \tilde{w}_{j+1}^{(m-1)}, \cdots, \tilde{w}_q^{(m-1)})$

(c) Until numerical convergence: we require $|f(\mathbf{w}^{(m)}) - f(\mathbf{w}^{(m-1)})|$ to be sufficiently small.

3.1.2 Extension to Multi-rank Cases

In this subsection, we extend the sparse unit rank projection procedure to handle multiple rank test problems when $r > 1$. We propose the $k-$th projection direction as the solution to the following problem:

$$\text{argmax}\, \frac{\mathbf{w}_k^T \Sigma_C \mathbf{w}_k}{\mathbf{w}_k^T \Sigma_R \mathbf{w}_k} \quad \text{s.t. } \mathbf{w}_k^T \Sigma_R \mathbf{w}_j = 0, \ \forall j < k. \tag{3.17}$$

It can be shown that (3.17) is equivalent to

$$\text{argmax}\, \mathbf{w}_k^T \Sigma_C \mathbf{w}_k \quad \text{s.t. } \mathbf{w}_k^T \Sigma_R \mathbf{w}_k \leq 1, \mathbf{w}_k^T \Sigma_R \mathbf{w}_j = 0, \ \forall j < k. \tag{3.18}$$

Following the reasoning in Witten and Tibshirani (2011), we recast (3.18) into an equivalent problem.

Proposition 3.1.3 *Problem* (3.18) *is equivalent to the following problem:*

$$\underset{\mathbf{w}}{\operatorname{argmax}} \frac{\mathbf{w}^T\mathbf{B}^T\mathbf{C}^T\Sigma_{11}^{1/2}P_{\perp}^{k-1}\Sigma_{11}^{1/2}\mathbf{C}\mathbf{B}\mathbf{w}}{\mathbf{w}^T\Sigma_R\mathbf{w}}, \tag{3.19}$$

where $P_{\perp}^{k-1}$ *is the projection matrix onto the orthogonal space spanned by* $\{\Sigma_{11}^{1/2}\mathbf{C}\mathbf{B}\mathbf{w}_j, 1 \leq j \leq k-1\}$*, in which* $\Sigma_{11} = (D^{-T}\Sigma_X D^{-1})_{(r,r)}$.

Based on Proposition 3.1.3, we consider several strategies of imposing the sparsity structure on $\mathbf{w}_k$. A simple strategy is to consider the following problem given by

$$\underset{\mathbf{w}_k}{\operatorname{argmax}}\, \mathbf{w}_k^T\Sigma_C^k\mathbf{w}_k - \lambda||\mathbf{w}_k||_1 \quad \text{s.t.} \quad \mathbf{w}_k^T\Sigma_R\mathbf{w}_k \leq 1, \tag{3.20}$$

where $\Sigma_C^k = \mathbf{B}^T\mathbf{C}^T\Sigma_{11}^{1/2}P_{\perp}^{k-1}\Sigma_{11}^{1/2}\mathbf{C}\mathbf{B}$. When the rank of $\mathbf{C}$ is greater than 1, the problem in (3.20) is no longer convex, since it involves maximizing an objective function that is not concave. A potential solution is to use the minorization-maximization (MM) algorithm (Lange et al. 2000). Specifically, for any fixed $\mathbf{w}^{(m)}$, we take a Taylor series expansion of $\mathbf{w}_k^T\Sigma_C^k\mathbf{w}_k$ at $\mathbf{w}^{(m)}$ and get

$$\mathbf{w}_k^T\Sigma_C^k\mathbf{w}_k - \lambda||\mathbf{w}_k||_1 \geq 2\mathbf{w}_k^T\Sigma_C^k\mathbf{w}_k^{(m)} - \mathbf{w}_k^{(m)T}\Sigma_C^k\mathbf{w}_k^{(m)} - \lambda||\mathbf{w}_k||_1. \tag{3.21}$$

Thus, the right hand side of (3.21) minorizes the objective function (3.20) at $\mathbf{w}_k^{(m)}$ and is a convex function, which can be solved by using some convex optimization methods. However, based on our extensive experience, the MM algorithm is too slow for most large-scale problems, such as imaging genetics.

To further improve computational efficiency, we consider a surrogate of (3.20). Recall the discussion in the second principle, we are only interested in extracting informative directions for testing hypotheses of interest. We consider a spectral decomposition

of $(D^{-T}\Sigma_X D^{-1})_{(r\times r)}$ as $(D^{-T}\Sigma_X D^{-1})_{(r\times r)} = \sum_{j=1}^{r}\gamma_j \boldsymbol{\ell}_j\boldsymbol{\ell}_j^T$, where $(\gamma_j, \boldsymbol{\ell}_j)$ are eigenvalue-eigenvector pairs with $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_r$. Then, instead of solving (3.20), we propose to solve r SURP problems as

$$\mathbf{w}_\lambda^k = \text{argmin}\,\frac{1}{2}\mathbf{w}_k^T\Sigma_R\mathbf{w}_k - \sqrt{\gamma_k}\boldsymbol{\ell}_k^T\mathbf{C}\mathbf{B}\mathbf{w}_k + \lambda_k||\mathbf{w}_k||_1 \;\text{ for } 1 \leq k \leq r. \tag{3.22}$$

Solving (3.22) leads to r sparse projection directions. In (3.22), since we sequentially extract the direction vector according to the input signal Σ_C, it may produce a less informative direction vector compared with those from (3.20). However, such formulation leads to a fast computational algorithm and our simulation results demonstrate its reasonable performance. Thus, (3.22) is preferred in practice.

3.1.3 Test Procedure

We consider three statistics for testing H_{0W} against H_{1W} in (3.2). Based on model (3.1), we calculate the ordinary least squares estimate of $\boldsymbol{\beta}_\mathbf{w}$, given by $\hat{\boldsymbol{\beta}}_\mathbf{w} = (\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T)^{-1}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{y}_i^T\mathbf{W}$. Subsequently, we calculate a $k \times k$ matrix, denoted by T_n, as follows:

$$T_n = (\mathbf{C}\hat{\boldsymbol{\beta}}_\mathbf{w} - \mathbf{b}_0)^T\Sigma_{\tilde{\Omega}}^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_\mathbf{w} - \mathbf{b}_0), \tag{3.23}$$

where $\Sigma_{\tilde{\Omega}}$ is a consistent estimate of the covariance matrix of $\mathbf{C}\hat{\boldsymbol{\beta}}_\mathbf{w} - \mathbf{b}_0$. Specifically, let $\tilde{\beta}_\mathbf{w}$ be the restricted least squares (RLS) estimate of $\boldsymbol{\beta}$ under H_0, which is given by

$$\tilde{\boldsymbol{\beta}}_\mathbf{w} = \hat{\boldsymbol{\beta}}_\mathbf{w} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_\mathbf{w} - \mathbf{b}_0).$$

Then, we can set $\Sigma_{\tilde{\Omega}} = \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\sum_{i=1}^{N}a_i^2\mathbf{x}_i\tilde{\boldsymbol{\epsilon}}_i^T\tilde{\boldsymbol{\epsilon}}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$, where $a_i = 1/\{1 - \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\}$ and $\tilde{\boldsymbol{\epsilon}}_i = \mathbf{W}^T\mathbf{y}_i - \tilde{\boldsymbol{\beta}}_\mathbf{w}^T\mathbf{x}_i$. When $k > 1$, we use the determinant, trace and eigenvalues of T_n as test statistics, which are given by

$$W_n = \det(T_n),\;\; \text{Tr}_n = \text{trace}(T_n),\;\text{ and }\text{Roy}_n = \max(\text{eig}(T_n)), \tag{3.24}$$

where det, trace, and eig, respectively, denote the determinant, trace and eigenvalues of a symmetric matrix. When $k = 1$, all three statistics in (3.24) reduce to the Wald-type (or Hotelling's T^2) test statistic. For simplicity, we focus on Tr_n throughout the paper.

We propose a wild bootstrap method to improve the finite sample performance of the test statistic Tr_n. First, we fit model (2.1) under the null hypothesis (2.2) to calculate the estimated regression coefficient matrix, denoted by $\widehat{\mathbf{B}}_{\mathbf{0}}$, with corresponding residuals $\hat{\mathbf{e}}_i = \mathbf{y}_i - \widehat{\mathbf{B}}_0^T \mathbf{x}_i$ for $i = 1, \ldots, n$. Then we generate G bootstrap samples $\mathbf{z}_i^{(g)} = (\widehat{\mathbf{B}}_0)^T \mathbf{x}_i + \eta_i^{(g)} \hat{\mathbf{e}}_i$ for $i = 1, \ldots, n$, where $\eta_i^{(g)}$ are independently and identically distributed as a distribution F, which is chosen to be ± 1 with equal probability. For each generated wild-bootstrap sample, we repeat the estimation procedure for estimating the optimal weights and the calculation of the test statistic $\mathrm{Tr}_n^{(g)}$. Subsequently, the p-value of Tr_n is computed as $\frac{1}{G} \sum_{g=1}^{G} \mathbf{1}(\mathrm{Tr}_n^{(g)} \geq \mathrm{Tr}_n)$, where $\mathbf{1}(\cdot)$ is an indicator function.

3.1.4 Tuning Parameter Selection

We consider several methods to select the tuning parameter λ. The first one is cross validation (CV), which is primarily a way of measuring the predictive performance of a statistical model. However, the CV technique can be computationally expensive for large-scale problems. The second one is the information criterion, which has been widely to measure the relative goodness of fit of a statistical model. However, neither of these two methods are applicable for SURP, since our primary interest is to find informative directions for appropriately testing the null and alternative hypotheses of (2.2). If the null hypothesis is true, it is expected that $\mathbf{C}\widehat{\mathbf{B}}$ only contains noisy components and the estimated direction vectors should be random. In this case, the test statistics Tr_n, W_n, and Roy_n should not be sensitive to the value of λ. This motivates us to use the rejection rate to select the tuning parameter as follows:

$$\hat{\lambda} = \underset{0 \leq \lambda \leq \lambda_{\max}}{\operatorname{argmax}} \{(\text{Rejection Rate})_\lambda\}, \tag{3.25}$$

where $\lambda_{\max}$ is the largest λ to make $\mathbf{w}$ nonzero.

3.2 Asymptotic Theory

We investigate several theoretical properties of SURP and its associated estimator. By substituting $\tilde{\Sigma}_R$ and $\widehat{\boldsymbol{\ell}} = \mathbf{C}\widehat{\mathbf{B}}$ into (5.13), we can calculate an estimate of $\mathbf{w}_0$ as

$$\hat{\mathbf{w}}_\lambda = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\tilde{\Sigma}_R\mathbf{w} - \mathbf{w}^T\widehat{\boldsymbol{\ell}} + \lambda||\mathbf{w}||_1. \tag{3.26}$$

The following question arises naturally:

how close is $\hat{\mathbf{w}}_\lambda$ to $\mathbf{w}_0$?

We address this question in Theorems 3.2.1 and 3.2.2.

We consider the scenario that there are a few nonzero components in $\mathbf{w}_0$, that is, a few response variables are associated with the covariates of interest. Such a scenario is common in many large-scale problems. We make a note here that the sparsity of $\mathbf{w}_0 = \Sigma_R^{-1}\boldsymbol{\ell}$ does not require neither Σ_R^{-1} nor $\boldsymbol{\ell}$ to be sparse, and hence are more quite flexible. Let $S_0 = \{j : w_{0,j} \neq 0\}$ be the active set of $\mathbf{w}_0 = (w_{0,1}, \cdots, w_{0,q})^T$ and s_0 is the number of elements in S_0. We use the banded covariance estimator of Σ_R (Bickel and Levina 2008b) such that $||\tilde{\Sigma}_R - \Sigma_R||_2 = O_p((\frac{\log q}{n})^{\frac{\alpha}{2(\alpha+1)}})$ for some well behaved covariance class $\mathcal{U}(\varepsilon_0, \alpha, C_1)$, which is defined as

$$\begin{aligned}\mathcal{U}(\varepsilon_0, \alpha, C_1) \quad = \quad & \{\Sigma = (\sigma_{jj'}) : \max_j \sum_{j'} \{|\sigma_{j'j}| : |j' - j| > k\} \leq C_1 k^{-\alpha} \text{ for all } k > 0 \\ & \text{and } 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0\}.\end{aligned}$$

We have the following results.

Theorem 3.2.1 *Assume that* $\Sigma_R \in \mathcal{U}(\varepsilon_0, \alpha, C_1)$ *and*

$$\lambda = \max\{(k_n t_1^0 + C_1 k_n^{-\alpha})||\mathbf{w}_0||_2, t_2^0\} \asymp (\frac{\log(q \vee n)}{n})^{\frac{\alpha}{2(\alpha+1)}}||\mathbf{w}_0||_2, \tag{3.27}$$

where $k_n \asymp (\frac{\log(q\vee n)}{n})^{-\frac{1}{2(\alpha+1)}}$, $t_1^0 := \sqrt{2(\eta_1+1)\frac{1}{\gamma(\epsilon_0,\delta)}}\sqrt{\frac{\log(q\vee n)}{n}}$, *and* $t_2^0 := \frac{C_0}{\varepsilon_0}\sqrt{2(\eta_2+1)}\sqrt{\frac{\log(q\vee n)}{n}}$, *in which* $\gamma(\varepsilon_0, \delta)$ *and* $\delta = \delta(\varepsilon_0)$ *only depends on* ε_0. *Then, with probability at least* $1 - (q \vee$

$n)^{-\eta_1} - (q \vee n)^{-\eta_2}$, *we have*

$$||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_2 \leq C\lambda\sqrt{s_0}, \tag{3.28}$$

where C is a constant not depending on q and n. Furthermore, for $||\boldsymbol{\ell}||_2 > \delta_0$, we have

$$||\frac{\hat{\mathbf{w}}_\lambda}{||\hat{\mathbf{w}}_\lambda||_2} - \frac{\mathbf{w}_0}{||\mathbf{w}_0||_2}||_2 \leq \frac{2C\lambda\sqrt{s_0}}{||\mathbf{w}_0||_2}. \tag{3.29}$$

Theorem 3.2.1 gives an oracle inequality and the L_2 convergence rate of $\hat{\mathbf{w}}_\lambda$ in the sparse case, which indicates direction consistency and is important to ensure the good performance of test statistics. This result has several important implications. If $\sqrt{s_0}(\frac{\log q}{n})^{\frac{\alpha}{2(\alpha+1)}} = o(1)$, then $||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_2$ converges to zero in probability. Therefore, our SURP should perform well for the extremely sparse cases with $s_0 << n$. This is extremely important in practice, since the extremely sparse cases are common for many large-scale problems. Although we consider the banded covariance estimator of Σ_R in Theorem 3.2.1 (Bickel and Levina 2008b), the convergence rate of $\hat{\mathbf{w}}_\lambda$ can be established for other estimators of Σ_R and $\boldsymbol{\ell}$ as follows.

Theorem 3.2.2 *Suppose that we have $||\tilde{\Sigma}_R - \Sigma_R||_2 = O_p(a_n) = o_p(1)$ and $||\hat{\boldsymbol{\ell}} - \boldsymbol{\ell}||_\infty = O_p(b_n) = o_p(1)$, then*

$$||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_2 = O_p((a_n \vee b_n)\sqrt{s_0}). \tag{3.30}$$

Furthermore, for $||\boldsymbol{\ell}||_2 > \delta_0$, we have

$$||\frac{\hat{\mathbf{w}}_\lambda}{||\hat{\mathbf{w}}_\lambda||_2} - \frac{\mathbf{w}_0}{||\mathbf{w}_0||_2}||_2 = O_p(\frac{(a_n \vee b_n)\sqrt{s_0}}{||\mathbf{w}_0||_2}). \tag{3.31}$$

Theorem 3.2.2 gives the L_2 convergence rate of $\hat{\mathbf{w}}_\lambda$ for any possible estimators of Σ_R and $\boldsymbol{\ell}$. A direct implication is that we can consider other estimators of Σ_R in order to achieve better estimation of Σ_R under different assumptions of Σ_R. For instance, if Σ_R has an approximate factor structure with sparsity, then we may consider the principal orthogonal complement thresholding (POET) method in Fan et al. (2013) to estimate Σ_R. Moreover, if we can achieve good estimation of $\boldsymbol{\ell}$ for large p, then we can extend model (2.1) to the

scenario with large p. We will systematically investigate these generalizations in our future work.

Remark The SPReM estimator $\hat{\mathbf{w}}_\lambda$ is closely connected with those estimators in Witten and Tibshirani (2011) and Fan et al. (2012) in the framework of penalized linear discriminant analysis. However, little is known about the theoretical properties of such estimators. To the best of our knowledge, Theorems 3.2.1 and 3.2.2 are the first results on the convergence rate of such estimators under the restricted eigen-vectors of problem (3.9).

Remark The SPReM estimator $\hat{\mathbf{w}}_\lambda$ does not have the oracle property due to the asymptotic bias introduced by the L_1 penalty. See detailed discussions in (Fan and Li 2001, Zou 2006). However, our estimation procedure may be modified to achieve the oracle property by using some non-concave penalties or adaptive weights. We will investigate this issue in more depth in our future work.

3.3 Numerical Examples

3.3.1 Simulation 1: Two Sample Test in High Dimensions

In this subsection, we consider high-dimensional two-sample test problems and compare SPReM with the High-dimensional Two-Sample test (HTS) method in Chen and Qin (2010) and the Random Projection (RP) method proposed by Lopes et al. (2011). Both HTS and RP are the state-of-the-art methods for detecting a shift between the means of two high-dimensional normal distributions. It has been shown in Lopes et al. (2011) that the random projection method outperforms several competing methods when q/n converges to a constant or ∞.

We simulated two sets of samples $\{\mathbf{y}_1, ..., \mathbf{y}_{n_1}\}$ and $\{\mathbf{y}_{n_1+1}, \ldots, \mathbf{y}_n\}$ from $N(\boldsymbol{\beta}_1, \Sigma_R)$ and $N(\boldsymbol{\beta}_2, \Sigma_R)$, respectively, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $q \times 1$ mean vectors and $\Sigma_R = \sigma^2(\rho_{jj'})$, in which $(\rho_{jj'})$ is a $q \times q$ correlation matrix. We set $n = 2n_1 = 100$ and the dimension of the multivariate response q is 50, 100, 200, 400, and 800, respectively. We are interested in testing the null hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ against $H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$. This two-sample test

problem can be formulated as a special case of model (2.1) with $n = n_1 + n_2$. Moreover, we have $\mathbf{B}^T = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$ and $\mathbf{C} = (1, -1)$. Without loss of generality, we set $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \mathbf{0}$ to assess type I error rate and then introduce a shift in the first ten components of $\boldsymbol{\beta}_2$ to be 1 to assess power. We set σ^2 to be 1 and 3 and consider three different correlation matrices as follows.

- Case 1 is an independent covariance matrix with $(\rho_{jj'}) = \text{diag}(1, \cdots, 1)$.
- Case 2 is a weak correlation matrix with $\rho_{jj'} = \mathbf{1}(j' = j) + 0.3 \times \mathbf{1}(j' \neq j)$.
- Case 3 is a strong correlation covariance matrix with $\rho_{jj'} = 0.8^{|j'-j|}$.

Simulation results are summarized in Tables 3.1 and 3.2. As expected, both HTS and RP perform worse as q gets larger, whereas our SPReM works very well even for relatively large q. This is consistent with our theoretical results in Theorems 3.2.1 and 3.2.2. Moreover, HTS and RP cannot control the type I error rate well in all scenarios, whereas our SPReM based on the wild bootstrap method works reasonably well. According to the best of our knowledge, none of the existing methods for the two sample test in high dimensions work well in this sparse setting. For cases (ii) and (iii), $\Sigma_R^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)$ is not sparse, but SPReM performs reasonably well under the correlated scenarios. This may indicate the potential of extending SPReM and its associated theory to non-sparse cases. As expected, increasing σ^2 decreases statistical power in rejecting the null hypothesis. Since both SPReM and RP significantly outperform HTS, we increased q to 2,000 and presented some additional comparisons between SPReM and RP based on 100 simulated data sets in Figure 1.

3.3.2 Simulation 2: Multiple Rank Cases

In this subsection, we evaluate the finite sample performance of SMURP. The simulation studies were designed to establish the association between a relatively high-dimensional imaging phenotype with a genetic marker (e.g., SNP or haplotype), which is common in imaging genetics studies, while adjusting for age and other environmental factors. We set the sample size $n = 100$ and the dimension of the multivariate phenotype q to be 50, 100,

200, 400 and 800, respectively, and then simulated the multivariate phenotype according to model (1). The random errors were simulated from a multivariate normal distribution with mean 0 and covariance matrix with diagonal elements 1. For the off-diagonal elements in the covariance matrix, which characterize the correlations among the multivariate phenotypes, we categorized each component of the multivariate phenotype into three categories: high correlation, medium correlation and very low correlation with the corresponding number of components $(1, 1, q-2)$ in each category, and then we set the three degrees of correlation among the different components of the multivariate phenotype according to Table 3. The final covariance matrix is set to be $\Sigma_R = \sigma^2(\rho_{jj'})$, where $(\rho_{jj'})$ is the correlation matrix. We considered $\sigma^2 = 1$ and 3.

For the covariates, we included two SNPs with an additive effect and 3 additional continuous covariates. We varied the minor allele frequency (MAF) of the first SNP, whereas we fixed the MAF of the second SNP to be 0.5. For the first SNP, we considered 6 scenarios assuming the MAFs are 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5, respectively. We simulated the three additional continuous covariates from a multivariate normal distribution with mean 0, standard deviation 1, and equal correlation 0.3. We first set $\mathbf{B} = 0$ to assess type I error rate. To assess power, we set the first response to be the only components of the multivariate phenotype associated with the first SNP and the second response to be the component related to the second SNP effect. Specifically, we set the coefficients of the two SNPs to be 1 for the selected responses and all other regression coefficients to be 0. We are interested in testing the joint effects of the two SNPs on phenotypic variance.

We applied SPReM to 100 simulated data sets. Note that to the best of our knowledge, no other methods can be used to test the multi-rank test problem and thus we only focus on SPReM here. Table 4 presents the estimated rejection rates corresponding to different MAFs, q, and σ^2. Our SPReM works very well even for relatively large q under both $\sigma^2 = 1$ and 3. Specifically, the wild bootstrap method can control the type I error rate well in all scenarios. For the power, SPReM performs reasonably well under the small MAFs and $q = 800$. It may indicate that our method can perform well for much larger q if the sample

size gets larger. As expected, increasing σ^2 decreases statistical power in rejecting the null hypothesis.

3.3.3 Alzheimer's Disease Neuroimaging Initiative (ADNI) Data Analysis

The development of SPReM is motivated by the joint analysis of imaging, genetic, and clinical variables in the ADNI study. "Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year publicprivate partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org. "

The Huamn 610-Quad BeadChip (Illumina, Inc. San Diego, CA) was used to genotype

818 subjects in the ADNI-1 database, which resulted in a set of 620,901 SNPs and copy number variation (CNV) markers. Since the Apolipoprotein E (ApoE) SNPs, rs429358 and rs7412, are not on the Human 610-Quad Bead-Chip, they were genotyped separately and added to the data set manually. For simplicity, we only considered the $10,479$ SNPs collected on the chromosome 19, which houses the famous ApoE gene commonly suspected of having association with Alzheimer's disease. A complete GWAS of ADNI will be reported elsewhere. The SNP data were preprocessed by standard quality control steps including dropping any SNP that has more than 5% missing data, imputing the missing values in each SNP with its mode, dropping SNPs with minor allele frequency < 0.05, and screening out SNPs violating the Hardy-Weinberg equilibrium. Finally, we obtained $8,983$ SNPs on chromosome 19, including the ApoE allele as the last SNP in our dataset.

Our problem of interest is to perform a genome-wide search for establishing the association between the $10,479$ SNPs collected on the chromosome 19 and the brain volume of 93 regions of interest (ROIs). We fitted model (1) with all 93 ROIs as responses and a covariate vector including an intercept, a specific SNP, age, gender, whole brain volume, and the top 5 principal components to account for population stratification. To reduce population stratification effects, we only used 761 Caucasians from all 818 subjects. Subjects with missing values were removed, which leads to 747 subjects. We set $\lambda = \lambda_{\max}$ in our SPReM for computational efficiency. To test the SNP effect on all 93 ROIs, we calculated the test statistic and its $p-$value for each SNP. We further performed a standard massive univariate analysis. Specifically, we fitted a linear model with the same set of covariates and calculated a $p-$value for every pair of ROIs and SNPs.

We developed a computationally efficient strategy to approximate the $p-$value of each SNP with different MAFs. In the real data analysis, we considered a pool of SNPs consisting of 6 MAF groups including MAF$\in (0.05, 0.075]$, MAF$\in (0.075, 0.15]$, MAF$\in (0.15, 0.25]$, MAF$\in (0.25, 0.35]$, MAF$\in (0.35, 0.45]$, and MAF$\in (0.45, 0.50]$. Each MAF group contains 40 SNPs. For each SNP, we generated 10,000 wild bootstrap samples under the null hypothesis to obtain 10,000 bootstrapped test statistics. Then, based on $40 \times 10,000$ bootstrapped

samples for each MAF group, we use the Satterthwaite method to approximate the null distribution of the test statistic by a Gamma distribution with parameters (a_T, b_T). Specifically, we set $a_T = \mathcal{E}^2/\mathcal{V}$ and $b_T = \mathcal{V}/\mathcal{E}$ by matching the mean $(\mathcal{E})$ and the variance $(\mathcal{V})$ of the test statistics and those of the Gamma distribution. The histograms and the fitted gamma distributions along with the QQ-plots are, respectively, presented in Figures 2-3. Figures 2 and 3 reveal that our gamma approximations work reasonably well for a wide range of MAFs when $\lambda = \lambda_{\max}$. Since we only use Gamma(a_T, b_T) to approximate the $p-$value of large test statistic, we only need a good approximation at the tail of the Gamma distribution. See Figure 3 for details. For each SNP, we matched its MAF with the closest MAF group in the pool and then calculated the $p-$value of the test statistic based on the approximated gamma distribution. We present the manhattan plot in Figure 4 and the top 10 SNPs with their $p-$values for SPReM and the mass univariate analysis in Table 5 for $\lambda = \lambda_{\max}$.

We have several important findings. The ApoE allele was identified as the top one significant covariate with $-\log_{10}(p) \sim 15$ and 9 respectively, indicating a strong association between the ApoE allele and imaging phenotype, a biomarker of Alzheimer's disease diagnosis. This finding agrees with the previous result in Vounou et al. (2012). We also found some interesting results regarding rs207650 on the TOMM40 gene, which is one of the top 10 significant SNPs with $-\log_{10}(p) \sim 5$ and 4 respectively. The TOMM40 gene is located in close proximity to the ApoE gene and has also been linked to AD in some recent studies (Vounou et al. 2012). We are also able to detect some additional SNPs, such as rs11667587 on the NOVA2 gene, among others, on the chromosome 19, which are not identified in existing genome-wide association studies. The new findings may shed more light on further Alzheimer's research. The $p-$values for those top 10 SNPs calculated from SPReM are much smaller than those calculated from the mass univariate analysis. In other words, to achieve comparable $p-$values, the mass univariate analysis requires many more samples. This strongly demonstrates the effectiveness of our proposed method.

3.4 Discussion

In this paper, we have developed a general SPReM framework based on the two heritability ratios. Our SPReM methodology has a wide range of applications, including sparse linear discriminant analysis, two sample tests, and general hypothesis tests in MLMs, among many others. We have systematically investigated the L_2 convergence rate of $\hat{\mathbf{w}}_\lambda$ in the ultra-high dimensional framework. We further extend the SURP problem to the SMURP and offered a sequential SURP approximation algorithm. We carried out simulation studies and examined a real data set to demonstrate the excellent performance of our SPReM framework compared to other state-of-the-art methods.

3.5 Assumptions and Proofs

Throughout the paper, the following assumptions are needed to facilitate the technical details, although they may not be the weakest conditions.

Assumption A1. $\mathbf{C}(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T \asymp 1$, that is, there exists constant c_0 and C_0 such that $c_0 \leq \mathbf{C}(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T \leq C_0$.

Assumption A2. $0 \leq \varepsilon_0 \leq \lambda_{\min}(\Sigma_R) \leq \lambda_{\max}(\Sigma_R) \leq 1/\varepsilon_0$.

Assumption A3. The covariance estimator $\tilde{\Sigma}_R$ satisfies: $||\tilde{\Sigma}_R - \Sigma_R||_2 = O_p(a_n) \leq o_p(1)$.

Remark : Assumption A1 is a very weak and standard assumption for regression models. Assumption A2 has been widely used in the literature. Assumption A3 requires a relatively accurate covariance estimator in terms of spectral norm convergence. We may use some good penalized estimators of Σ_R under different assumptions of Σ_R (Bickel and Levina 2008b, Cai et al. 2010, Lam and Fan 2009, Rothman et al. 2009, Fan et al. 2013).

Proof of Theorem 3.1.1 The Karush Kuhn Tucker (KKT) conditions for problem (3.11) are given by:

$$\boldsymbol{\ell} - \lambda\Gamma - \gamma\Sigma_R\mathbf{w} = 0, \gamma \geq 0, \gamma(\frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} - \frac{1}{2}) = 0, \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} \leq \frac{1}{2},$$

where Γ is a $q \times 1$ vector and equals the subgradient of $||\mathbf{w}||_1$ with respect to $\mathbf{w}$. We consider two scenarios. First, suppose that $|\ell_j| > \lambda$ for some j. We must have $\gamma \Sigma_R \mathbf{w} \neq 0$, which leads to $\gamma > 0$ and $\mathbf{w}^T \Sigma_R \mathbf{w} = 1$. Thus, the KKT conditions reduce to

$$\boldsymbol{\ell} - \lambda\Gamma - \gamma\Sigma_R\mathbf{w} = 0,\ \gamma \geq 0,\ \mathbf{w}^T\Sigma_R\mathbf{w} = 1.$$

If we write $\tilde{\mathbf{w}} = \gamma\mathbf{w}$, this is equivalent to solving problem (5.13) with $\tilde{\mathbf{w}}$ and then take normalization. Second, if $|\ell_j| \leq \lambda$ for any j, then $\mathbf{w} = 0$ and $\gamma = 0$, which is the solution of (5.13) as well. This finishes the proof.

Proof of Proposition 3.1.2 It follows from Theorem 2 of Rosset and Zhu (2007).

Proof of Proposition 3.1.3 The proof is similar to that of Proposition 1 of Witten and Tibshirani (2011). Letting $\tilde{\mathbf{w}}_k = \Sigma_R^{1/2}\mathbf{w}_k$, then problem (3.18) can be rewritten as

$$\text{argmax}\, \tilde{\mathbf{w}}_k^T \Sigma_R^{-1/2}\mathbf{B}^T\mathbf{C}^T\Sigma_{11}^{1/2}\mathbf{C}\mathbf{B}\Sigma_R^{-1/2}\tilde{\mathbf{w}}_k \ \text{ s.t. } \ ||\tilde{\mathbf{w}}_k||^2 \leq 1,$$

which is equivalent to

$$\text{argmax}\, \tilde{\mathbf{w}}_k\mathbf{A}P_\perp^{k-1}\mathbf{u}_k \text{ s.t } \ ||\tilde{\mathbf{w}}_k||^2 \leq 1, ||\mathbf{u}_k||^2 \leq 1, \tag{3.32}$$

where $\mathbf{A} = \mathbf{B}^T\mathbf{C}^T\Sigma_{11}^{1/2}$. Thus, $\tilde{\mathbf{w}}_k$ and $\mathbf{u}_k$ that solve problem (3.32) are the k-th left and right singular vectors of $\mathbf{A}$ (Witten and Tibshirani 2011). Therefore, we have $P_\perp^{k-1} = \mathbf{I} - \sum_{j=1}^{k-1}\mathbf{u}_j\mathbf{u}_j^T$ and $\mathbf{u}_k$ is the k-th eigenvector of $\mathbf{A}^T\mathbf{A}$, or equivalently the k-th right singular vector of $\mathbf{A}$. For problem (3.32), $\tilde{\mathbf{w}}_k$ is the k-th left singular vector of $\mathbf{A}$. Therefore, the solution of (3.19) is the k-th discriminant vector of (3.18).

Proof of Theorem 3.2.1 In this theorem, we specifically use the banded covariance estimator $\tilde{\Sigma}_R = B_{k_n}(\hat{\Sigma}_R)$, where $B_k(\Sigma) = [\sigma_{jj'}I(|j' - j| \leq k)]$ and $\hat{\Sigma}_R$ is the sample covariance matrix of $\mathbf{y}_i - \hat{\mathbf{B}}^T\mathbf{x}_i$.

First, we define $\mathcal{J} = \{||\tilde{\Sigma}_R - B_{k_n}(\Sigma_R)||_\infty \leq t_1\} \bigcap \{||\hat{\boldsymbol{\ell}} - \boldsymbol{\ell}||_\infty \leq t_2\}$, where t_1 and t_2 are

specified as in Lemma 3.5.2. Then, it follows from Lemma 3.5.2 that $P(\mathcal{J}) \geq 1 - 3(q \vee n)^{-\eta_1} - 2(q \vee n)^{-\eta_2}$.

On the set $\mathcal{J}$, by taking $\lambda = \max\{k_n t_1 + C_1 k_n^{-\alpha}, t_2\}$ and using Lemma 3.5.1, we have

$$\begin{aligned}
&\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T \tilde{\Sigma}_R (\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda ||\hat{\mathbf{w}}_\lambda||_1 \leq (\mathbf{w}_0^T (\Sigma_R - \tilde{\Sigma}_R) + \boldsymbol{\varepsilon}^T)(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda ||\mathbf{w}_0||_1 \\
&\leq ||\tilde{\Sigma}_R - \Sigma_R||_2 ||\mathbf{w}_0||_2 ||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_1 + ||\boldsymbol{\varepsilon}||_\infty ||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_1 + \lambda ||\mathbf{w}_0||_1 \\
&\leq (k_n t_1 + C_1 k_n^{-\alpha}) ||\mathbf{w}_0||_2 ||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_1 + t_2 ||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_1 + \lambda ||\mathbf{w}_0||_1 \\
&\leq \lambda ||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_1 + \lambda ||\mathbf{w}_0||_1.
\end{aligned}$$

Let $\mathbf{w}_{0,S_0} = [w_{0,j} I(j \in S_0)]$, where $w_{0,j}$ is the $j-$th component of $\mathbf{w}_0$. The above equation can be rewritten as

$$\begin{aligned}
&(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T (\tilde{\Sigma}_R - \Sigma_R + \Sigma_R)(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda ||\hat{\mathbf{w}}_{\lambda,S_0}||_1 + \lambda ||\hat{\mathbf{w}}_{\lambda,S_0^c}||_1 \\
&\leq \lambda ||\hat{\mathbf{w}}_{\lambda,S_0} - \mathbf{w}_{0,S_0}||_1 + \lambda ||\mathbf{w}_{0,S_0}||_1 + \lambda ||\hat{\mathbf{w}}_{\lambda,S_0^c}||_1,
\end{aligned}$$

which yields

$$\{\lambda_{\min} - O(1)(\frac{\log(q)}{n})^{\frac{\alpha}{2(\alpha+1)}}\} ||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_2^2 \leq 2\lambda \sqrt{s_0} ||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_2.$$

Finally, we obtain the following inequality

$$||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_2 \leq \frac{2\lambda\sqrt{s_0}}{\lambda_{min} - O(1)(\frac{\log(q)}{n})^{\frac{\alpha}{2(\alpha+1)}}} \leq C\lambda\sqrt{s_0},$$

which finishes the proof.

Proof of Theorem 3.2.2 It follows from Lemma (3.5.1) that

$$\begin{aligned}
&\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T \tilde{\Sigma}_R (\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda ||\hat{\mathbf{w}}_\lambda||_1 \leq (\mathbf{w}_0^T (\Sigma_R - \tilde{\Sigma}_R) + (\hat{\boldsymbol{\ell}} - \hat{\boldsymbol{\ell}}))(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda ||\mathbf{w}_0||_1 \\
&\leq (||\tilde{\Sigma}_R - \Sigma_R||_2 ||\mathbf{w}_0||_2 + ||\hat{\boldsymbol{\ell}} - \boldsymbol{\ell}||_\infty) ||\hat{\mathbf{w}}_{\lambda,S} - \mathbf{w}_{0,S}||_1 + \lambda ||\mathbf{w}_{0,S}||_1
\end{aligned}$$

Note that $||\hat{\mathbf{w}}_\lambda||_1 \leq ||\mathbf{w}_{0,S_0}||_1 - ||\mathbf{w}_{0,S_0} - \hat{\mathbf{w}}_{\lambda,S_0}||_1 + ||\hat{\mathbf{w}}_{\lambda,S_0^c}||_1$. Then, by taking

$$\lambda = ||\tilde{\Sigma}_R - \Sigma_R||_2||\mathbf{w}_0||_2 + ||\hat{\boldsymbol{\ell}} - \boldsymbol{\ell}||_\infty \asymp O_p(a_n||\mathbf{w}_0||_2 \vee b_n),$$

we have

$$\begin{aligned}
&\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) \leq (||\tilde{\Sigma}_R - \Sigma_R||_2||\mathbf{w}_0||_2 + ||\hat{\boldsymbol{\ell}} - \boldsymbol{\ell}||_\infty)(||\hat{\mathbf{w}}_{\lambda,S_0} - \mathbf{w}_{0,S_0}||_1 + ||\hat{\mathbf{w}}_{\lambda,S_0^c}||_1) \\
&\quad - \lambda(||\mathbf{w}_{0,S_0}||_1 - ||\mathbf{w}_{0,S} - \hat{\mathbf{w}}_{\lambda,S_0}||_1 + ||\hat{\mathbf{w}}_{\lambda,S_0^c}|) + \lambda|\mathbf{w}_{0,S_0}||_1 \\
&= O_p(a_n \vee b_n)||\hat{\mathbf{w}}_{\lambda,S_0} - \mathbf{w}_{0,S_0}||_1 \leq O_p(a_n \vee b_n)\sqrt{s_0}||\hat{\mathbf{w}}_{\lambda,S} - \mathbf{w}_{0,S_0}||_2.
\end{aligned}$$

By using Weyl's inequality, we have

$$(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) \geq (\lambda_{\min}(\Sigma_R) - ||\tilde{\Sigma}_R - \Sigma_R||_2)||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_2^2$$

where $||\tilde{\Sigma}_R - \Sigma_R||_2 = O_p(a_n) = o_p(1)$. Finally, we have

$$||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||_2^2 \leq \frac{O_p(a_n \vee b_n)\sqrt{s_0}||\hat{\mathbf{w}}_{\lambda,S} - \mathbf{w}_{0,S}||_2}{\lambda_{min}(\Sigma) - O_p(a_n)}, \tag{3.33}$$

which finishes the proof.

Lemma 3.5.1 *We have the following basic inequality*

$$\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0)^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda||\hat{\mathbf{w}}_\lambda||_1 \leq \{\mathbf{w}_0^T(\Sigma_R - \tilde{\Sigma}_R) + (\hat{\boldsymbol{\ell}} - \boldsymbol{\ell})^T\}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda||\mathbf{w}_0||_1. \tag{3.34}$$

Proof We rewrite the optimization problem (3.26) as

$$\hat{\mathbf{w}}_\lambda = \text{argmin}\,\frac{1}{2}(\mathbf{w} - \tilde{\Sigma}_R^{-1}\hat{\boldsymbol{\ell}})^T\tilde{\Sigma}_R(\mathbf{w} - \tilde{\Sigma}_R^{-1}\hat{\boldsymbol{\ell}}) + \lambda||\mathbf{w}||_1.$$

Thus, we have

$$\frac{1}{2}(\hat{\mathbf{w}}_\lambda - \tilde{\Sigma}_R^{-1}\hat{\boldsymbol{\ell}})^T\tilde{\Sigma}_R(\hat{\mathbf{w}}_\lambda - \tilde{\Sigma}_R^{-1}\hat{\boldsymbol{\ell}}) + \lambda||\hat{\mathbf{w}}_\lambda||_1 \leq \frac{1}{2}(\mathbf{w}_0 - \tilde{\Sigma}_R^{-1}\hat{\boldsymbol{\ell}})^T\tilde{\Sigma}_R(\mathbf{w}_0 - \tilde{\Sigma}_R^{-1}\hat{\boldsymbol{\ell}}) + \lambda||\mathbf{w}_0||_1,$$

which yields

$$\begin{aligned}\frac{1}{2}||\hat{\mathbf{w}}_\lambda - \mathbf{w}_0||^2_{\tilde{\Sigma}_R} + \lambda||\hat{\mathbf{w}}_\lambda||_1 &\leq (\hat{\boldsymbol{\ell}} - \tilde{\Sigma}_R\mathbf{w}_0)^T(\tilde{\mathbf{w}}_\lambda - \mathbf{w}_0) \\ &= \{\mathbf{w}_0^T(\Sigma_R - \tilde{\Sigma}_R) + (\hat{\boldsymbol{\ell}} - \boldsymbol{\ell})^T\}(\hat{\mathbf{w}}_\lambda - \mathbf{w}_0) + \lambda||\mathbf{w}_0||_1,\end{aligned}$$

in which we have used $\hat{\boldsymbol{\ell}} = \Sigma_R\mathbf{w}_0 + \hat{\boldsymbol{\ell}} - \boldsymbol{\ell}$ in the last equality.

Lemma 3.5.2 *For all $t_1 \geq t_1^0$ and $t_2 \geq t_2^0$, we have*

$$P(\mathcal{J}) \geq 1 - 3(q \vee n)^{-\eta_1} - 2(q \vee n)^{-\eta_2}. \tag{3.35}$$

Proof First, it follows from Lemma A.3 of Bickel and Levina (2008b) that

$$\begin{aligned}P(||\tilde{\Sigma}_R - B_{k_n}(\Sigma_R)||_\infty \geq t_1) &\leq 2(k+1)q\exp\{-n(t_1^0)^2\gamma(\varepsilon_0,\delta)\} \\ &\leq 2(k_n+1)(q\vee n)\exp\{-2n(\eta_1+1)\frac{1}{\gamma(\varepsilon_0,\delta)}\frac{\log(q\vee n)}{n}\gamma(\varepsilon_0,\delta))\} \\ &\leq 3((q\vee n)k_n)\exp\{-(\eta_1+1)\log((q\vee n)k_n))\} \\ &\leq 3((q\vee n)k_n)^{-(\eta_1+1)+1} \leq 3(q\vee n)^{-\eta_1},\end{aligned}$$

where $t_1^0 = \sqrt{2(\eta_1+1)\frac{1}{\gamma(\epsilon_0,\delta)}}\sqrt{\frac{\log(q\vee n)}{n}}$.

Second, we know that $\frac{\sqrt{n}(\hat{\ell}_j - \ell_j)}{\sigma_j C_X}$ is $Sub(1)$-distributed, where $C_X = \mathbf{C}(\frac{1}{n}\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T$. Then by the union sum inequality, we have

$$\begin{aligned}P(\max_j|\frac{\sqrt{n}(\hat{\ell}_j - \ell_j)}{C_0/\varepsilon_0}| \geq t_2) &\leq P(\max_j|\frac{\sqrt{n}(\hat{\ell}_j - \ell_j)}{\sigma_j C_X}| \geq t_2) \\ &\leq 2(q\vee n)\exp\{-\frac{(t_2^0)^2}{2}\}.\end{aligned} \tag{3.36}$$

By taking $(t_2^0)^2 = 2\eta_2\log(q\vee n)$, we can rewrite the above inequality as

$$P(||\hat{\boldsymbol{\ell}} - \boldsymbol{\ell}||_\infty \geq \frac{C_0}{\varepsilon_0}\sqrt{\frac{(2\eta_2+2)\log(q\vee n)}{n}}) \leq 2(q\vee n)^{-\eta_2}$$

Finally, we get

$$P(\mathcal{J}) \geq 1 - P(||\tilde{\Sigma}_R - B_k(\Sigma_R)||_\infty \geq t_1^0) - P(||\hat{\boldsymbol{\ell}} - \boldsymbol{\ell}||_\infty \geq \frac{C_0}{\varepsilon_0}\sqrt{\frac{(2\eta_2+2)\log(q \vee n)}{n}})$$
$$\geq 1 - 3(q \vee n)^{-\eta_1} - 2(q \vee n)^{-\eta_2},$$

which finishes the proof.

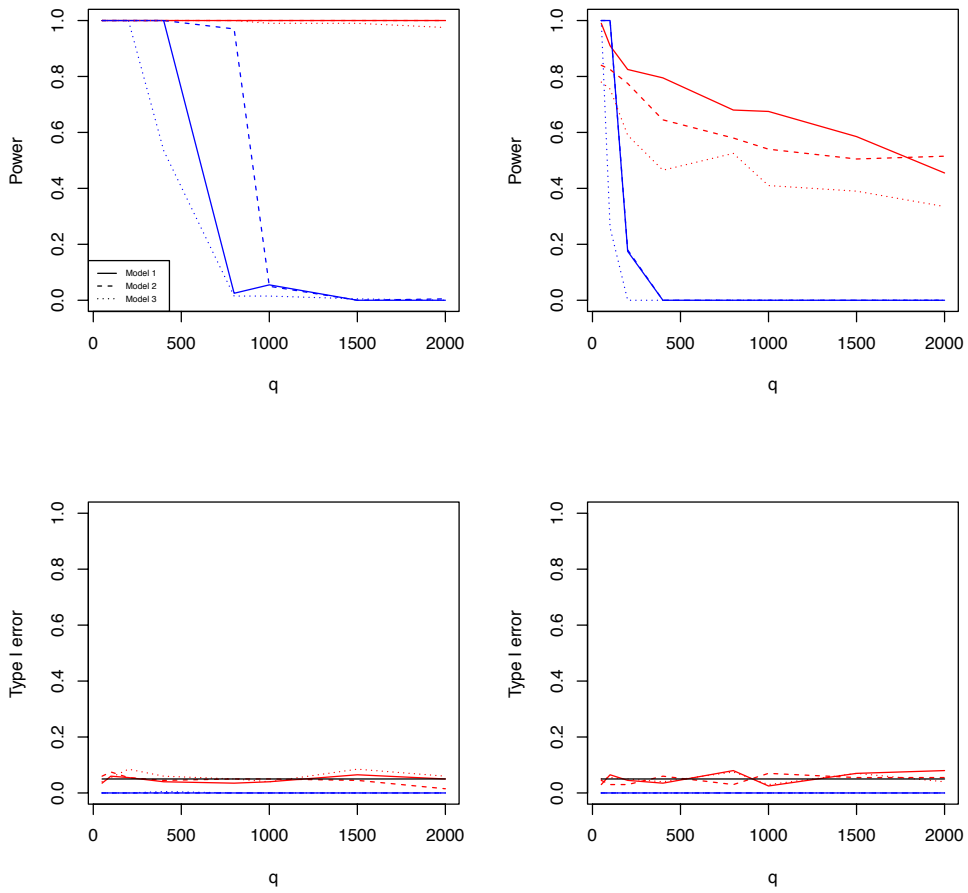

Figure 3.1: Simulation 1 results: the estimated rejection rates as functions of q for two different σ^2 values. The upper and lower rows are, respectively, for powers and for type I error rates, whereas the left and right columns correspond to $\sigma^2 = 1$ and $\sigma^2 = 3$, respectively. In all panels, the lines obtained from SPReM and RP are, respectively, presented in red and in blue, and the results for independence, weak, and strong correlation structures are, respectively, presented as thick, dashed, and dotted lines.

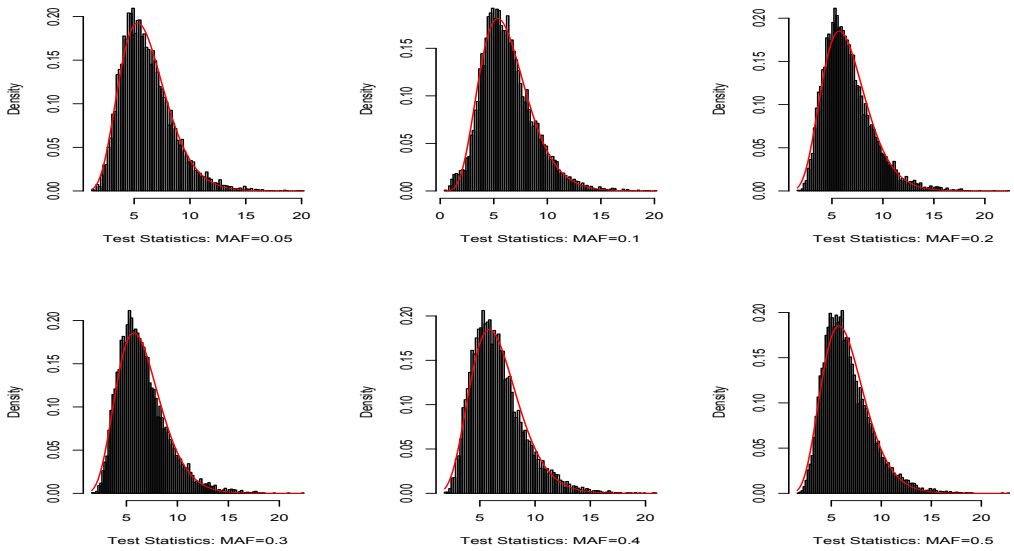

Figure 3.2: Histograms and their gamma approximations based on the wild bootstrap samples under the null hypothesis for different MAFs for $\lambda = \lambda_{\max}$.

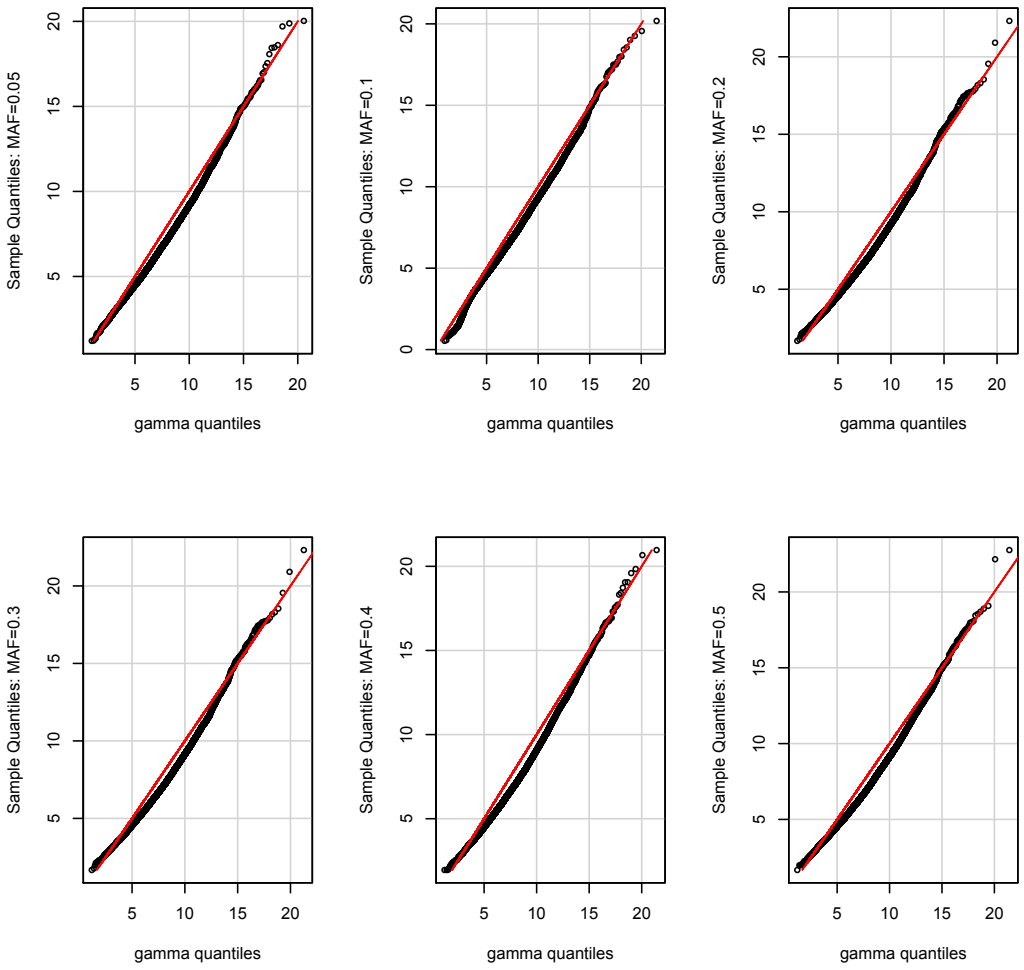

Figure 3.3: QQ-plot of the gamma approximations based on the wild bootstrap samples under the null hypothesis for different MAFs for $\lambda = \lambda_{\max}$.

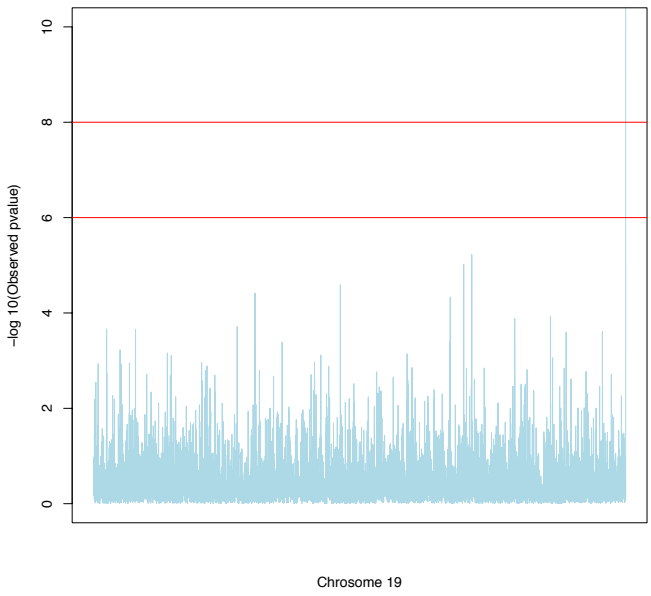

Figure 3.4: ADNI GWAS results: Manhattan plot of $-\log_{10}(p)$-values on chromosome 19 by SPReM for $\lambda = \lambda_{\max}$.

Table 3.1: Simulation 1: power and type I error are reported for two sample test at 5 different qs at significance level $\alpha = 5\%$ when $\sigma^2 = 1$.

	Power					Type I error				
q	50	100	200	400	800	50	100	200	400	800
	case 1									
SPReM	1.000	1.000	1.000	1.000	1.000	0.035	0.060	0.055	0.040	0.035
RP	1.000	1.000	1.000	1.000	0.025	0.000	0.000	0.000	0.000	0.000
HTS	0.965	0.320	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	case 2									
SPReM	1.000	1.000	1.000	1.000	1.000	0.060	0.075	0.055	0.045	0.050
RP	1.000	1.000	1.000	1.000	0.970	0.000	0.000	0.000	0.000	0.000
HTS	1.000	0.245	0.030	0.005	0.000	0.000	0.000	0.000	0.000	0.000
	case 3									
SPReM	1.000	1.000	1.000	1.000	1.000	0.040	0.055	0.085	0.060	0.050
RP	1.000	1.000	1.000	0.535	0.015	0.000	0.000	0.000	0.005	0.000
HTS	1.000	0.140	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 3.2: Simulation 1: power and type I error are reported for two sample test at 5 different qs at significance level $\alpha = 5\%$ when $\sigma^2 = 3$.

	Power					Type I error				
q	50	100	200	400	800	50	100	200	400	800
	case 1									
SPReM	0.990	0.910	0.825	0.795	0.680	0.030	0.065	0.045	0.035	0.080
RP	1.000	1.000	0.175	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HTS	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	case 2									
SPReM	0.840	0.825	0.775	0.645	0.580	0.045	0.030	0.030	0.060	0.030
RP	1.000	1.000	0.180	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HTS	0.105	0.015	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	case 3									
SPReM	0.780	0.755	0.590	0.465	0.525	0.050	0.055	0.050	0.040	0.075
RP	1.000	0.260	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
HTS	0.095	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 3.3: Correlation matrix of responses used in the simulation

	High	Med	Low
High	0.9	0.6	0.3
Med	0.6	0.9	0.1
Low	0.3	0.1	0.1

Table 3.4: Simulation 2: the estimates of rejection rates were reported at 6 different MAFs, 5 different qs, and 2 different σ^2 values at significance level $\alpha = 5\%$. For each case, 100 simulated data sets were used.

	Power					Type I error				
	$\sigma^2 = 1$									
MAF$\backslash q$	50	100	200	400	800	50	100	200	400	800
0.050	0.950	0.955	0.930	0.940	0.930	0.045	0.060	0.030	0.070	0.080
0.100	0.995	0.990	0.990	0.980	0.975	0.045	0.055	0.040	0.045	0.045
0.200	1.000	1.000	1.000	1.000	1.000	0.045	0.045	0.080	0.030	0.060
0.300	1.000	1.000	1.000	1.000	1.000	0.065	0.040	0.020	0.065	0.060
0.400	1.000	1.000	1.000	1.000	1.000	0.050	0.070	0.035	0.060	0.070
0.500	1.000	1.000	1.000	1.000	1.000	0.060	0.050	0.030	0.020	0.035
	$\sigma^2 = 3$									
0.050	0.915	0.875	0.765	0.795	0.735	0.050	0.040	0.030	0.050	0.065
0.100	0.970	0.960	0.940	0.875	0.865	0.040	0.055	0.070	0.080	0.050
0.200	0.995	0.985	0.975	0.975	0.970	0.015	0.050	0.060	0.010	0.065
0.300	1.000	1.000	0.990	0.970	0.955	0.045	0.055	0.055	0.080	0.040
0.400	0.995	1.000	1.000	0.990	0.985	0.055	0.035	0.045	0.050	0.070
0.500	0.995	1.000	1.000	0.985	0.980	0.085	0.055	0.055	0.065	0.030

Table 3.5: Comparison between SPReM and the massive univariate analysis (MUA) for ADNI data analysis: the top 10 SNPs and their $-\log_{10}(p)$ values for $\lambda = \lambda_{\max}$.

SNP	apoe_allele	rs11667587	rs2075650	rs7248284	rs3745341
SPReM	5.04E-16	5.95E-06	9.58E-06	2.56E-05	3.83E-05
MUA	3.43E-11	4.42E-04	1.12E-04	8.75E-04	1.00E-03
SNP	rs4803646	rs8106200	rs2445830	rs8102864	rs740436
SPReM	4.65E-05	1.16E-04	1.32E-04	1.93E-04	2.17E-04
MUA	7.56E-04	3.70E-03	1.33E-02	9.34E-04	1.63E-03

CHAPTER4: HARD THRESHOLDED REGRESSION

In this chapter, we propose a Hard Thresholded Regression (HTR) framework for simultaneous variable selection and unbiased estimation in high dimensional linear regression. This new framework is motivated by its close connection with the L_0 regularization and best subset selection under orthogonal design, while enjoying several key computational and theoretical advantages over many existing penalization methods (e.g., SCAD or MCP). Computationally, HTR is a fast two-stage estimation procedure consisting of the first step for calculating a coarse initial estimator and the second step for solving a linear program. Theoretically, under some mild conditions, the HTR estimator is shown to enjoy the strong oracle property and thresholded property even when the number of covariates may grow at an exponential rate. We also propose to incorporate the regularized covariance estimator into the estimation procedure in order to better trade off between noise accumulation and correlation modeling. Under this scenario with regularized covariance matrix, HTR includes Sure Independence Screening as a special case. Both simulation and real data results show that HTR outperforms other state-of-the-art methods.

4.1 Methods

4.1.1 Hard Thresholded Regression (HTR)

Consider n independent observations $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ from model (2.3) with the true parameter vector $\boldsymbol{\beta}^o$. Without loss of generality, we standardize each column of $\mathbf{X} = (\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_p)$ so that $||\tilde{\mathbf{x}}_k||_2 = \sqrt{n}$ for $k = 1, \ldots, p$. The target of HTR in (2.5) is to estimate $\boldsymbol{\beta}^o$ from the data. Our HTR algorithm is a two-stage approach.

1. Compute an initial estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{init}$, with a reasonably small risk error

bound. For instance, let $\hat{\boldsymbol{\beta}}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda_{init}\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$ be a ridge estimator of $\boldsymbol{\beta}$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix and $\lambda_{init} \geq 0$ is a tuning parameter. When $\lambda_{init} = 0$, $\hat{\boldsymbol{\beta}}^{ridge}$ reduces to the ordinary least squares estimator of $\boldsymbol{\beta}$. We will use $\hat{\boldsymbol{\beta}}^{ridge}$ as a candidate of $\hat{\boldsymbol{\beta}}_{init}$ and examine its risk error bound in Section 2.5.

2. Construct the weight matrix $\mathbf{W}$ based on $\hat{\boldsymbol{\beta}}_{init}$, denoted by $\widehat{\mathbf{W}}$, and then write the HTR estimator as

$$\hat{\boldsymbol{\beta}}_{HTR} = \text{argmin}\, \frac{1}{n}||\widehat{\mathbf{W}} \times \mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})||_1 + \lambda||\boldsymbol{\beta}||_1. \tag{4.1}$$

Throughout the paper, we set $\widehat{\mathbf{W}}$ as

$$\widehat{\mathbf{W}} = \text{diag}(\hat{w}_j) \text{ and } \hat{w}_j = |\hat{\beta}_{init,j}|^\gamma \text{ for } j = 1, 2, \ldots, p, \tag{4.2}$$

where γ is a positive constant and $\hat{\beta}_{init,j}$ is the $j-$th component of $\hat{\boldsymbol{\beta}}$.

Numerically, computation of $\hat{\boldsymbol{\beta}}_{HTR}$ is very straightforward, since the objective function in (4.1) is convex and can be recast into a linear programming problem. Specifically, we introduce a $p \times 1$ slack vector $\boldsymbol{\eta} = \{\eta_j = \frac{1}{n}|[\widehat{\mathbf{W}}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]_j|, j = 1, \ldots, p\}$, $\boldsymbol{\beta}^+ = \{\beta_j^+\}_{j\geq 1}$, and $\boldsymbol{\beta}^- = \{\beta_j^-\}_{j\geq 1}$. Then, the minimization in (4.1) can be rewritten as

$$\min \sum_{j=1}^{p} \{\eta_j + \lambda(\beta_j^+ + \beta_j^-)\} \text{ subject to } \boldsymbol{\eta} \geq \mathbf{0}, \boldsymbol{\beta}^+ \geq \mathbf{0}, \boldsymbol{\beta}^- \geq \mathbf{0}, \text{ and}$$
$$-\boldsymbol{\eta} \leq \frac{1}{n}\widehat{\mathbf{W}}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \leq \boldsymbol{\eta},$$

where the optimization variables are $\boldsymbol{\eta}$, $\boldsymbol{\beta}^+$, and $\boldsymbol{\beta}^-$ in R^p. Finally, $\boldsymbol{\beta}$ can be recovered by $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$.

There are at least two major motivations for HTR. The first one comes from the score equation of the maximum likelihood estimator. Let $\ell_n(\boldsymbol{\beta})$ and $U_n(\boldsymbol{\beta})$ be, respectively, the likelihood (or quasi-likelihood) and score functions of $\boldsymbol{\beta}$. The score equation and its weighted

version are given by

$$U_n(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) = \mathbf{0} \text{ and } \widehat{\mathbf{W}} \times U_n(\boldsymbol{\beta}) = \mathbf{0}, \tag{4.3}$$

which are equivalent to $||U_n(\boldsymbol{\beta})||_1 = 0$ and $||\widehat{\mathbf{W}} \times U_n(\boldsymbol{\beta})||_1 = 0$, respectively. For model (2.3), $U_n(\boldsymbol{\beta})$ reduces to $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and thus $\hat{\boldsymbol{\beta}}_{HTR}$ can be regarded as the penalized weighted score estimator with the L_1 norm $||\boldsymbol{\beta}||_1$. Moreover, $\mathbf{R}(\boldsymbol{\beta}) = (R_1(\boldsymbol{\beta}), R_2(\boldsymbol{\beta}), \ldots, R_p(\boldsymbol{\beta}))^T = U_n(\boldsymbol{\beta})$ can be regarded as the risk function of $\boldsymbol{\beta}$ and $\widehat{\mathbf{W}}$ is the risk calibration weight matrix for imposing additional information learned from the first stage. Therefore, based on (4.3), it is possible to extend HTR to more general scenarios, such as generalized linear model.

The second motivation comes from the Dantzig selector (Candes and Tao 2007) and the least absolute gradient selector (LAGS) (Yang 2012). These two selectors are equivalent to solving the objective function as

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{n} ||\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})||_a + \lambda ||\mathbf{V}\boldsymbol{\beta}||_1, \tag{4.4}$$

where $\mathbf{V}$ is a $p \times p$ weight matrix. The Dantzig selector and LAGS correspond to $(||\cdot||_a, \mathbf{V}) = (||\cdot||_\infty, \mathbf{I}_p)$ and $(||\cdot||_a, \mathbf{V}) = (||\cdot||_\infty, \text{diag}(1/|\hat{\beta}_{init,1}|, \cdots, 1/|\hat{\beta}_{init,p}|))$, respectively. As pointed by Candes and Tao (2007), one would want to constrain the size of the correlated residual vector $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ rather than the size of the residual vector $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, since such an estimation procedure is invariance under orthogonal transformations of $\mathbf{X}$. Moreover, since the correlated residual vector measures the correlation between the predictors and the response, one would obviously want to include the explanatory variables that are highly correlated with the response $\mathbf{y}$ in the model.

A major drawback of the Dantzig selector is shrinkage bias, leading to suboptimal risk estimation, even though a double Dantzig selector can reduce the bias (James and Radchenko 2009). Moreover, to address the same bias issue, similar to the adaptive Lasso (Zou 2006), LAGS uses adaptive weights calculated from $\hat{\boldsymbol{\beta}}_{init}$ to directly penalize different regression coefficients. An advantage of HTR is that it directly reduces the effects of those risk functions

$R_j(\boldsymbol{\beta})$ associated with 'insignificant' β_js' in both estimation and variable selection. When $s << \min(p, n)$ and p is comparable with n, we expect that HTR outperforms LAGS in terms of bias and mean squared error. See Section 4 for details.

4.1.2 Orthonormal Design Case

We examine the orthonormal design case in order to delineate some connections between HTR and other existing regularization methods. In this case, we have $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_p$ and $\hat{\boldsymbol{\beta}}^{ols} = (\hat{\beta}_1^{ols}, \cdots, \hat{\beta}_p^{ols})^T = n^{-1}\mathbf{X}^T\mathbf{y}$.

Best subset selection of size k reduces to choosing the k largest coefficients in absolute value and setting the rest to 0. Specifically, for some value of λ, this is equivalent to

$$\hat{\beta}_j = \hat{\beta}_j^{ols} 1_{|\hat{\beta}_j^{ols}| > \lambda} \text{ for } j = 1, \cdots, p, \tag{4.5}$$

which have a strong connection with hard shrinkage. For the Lasso (Tibshirani 1996), its solutions have the following form

$$\hat{\beta}_{lasso,j} = \text{sgn}(\hat{\beta}_j^{ols})(|\hat{\beta}_j^{ols}| - \lambda)_+ \text{ for } j = 1, \cdots, p, \tag{4.6}$$

which has a strong connection with the soft shrinkage proposals of Donoho and Johnstone (1994), Donoho et al. (1995). However, there is a major shrinkage bias in (4.6).

Many convex/nonconvex penalty functions in (2.4) have been proposed to reduce the effect of the shrinkage bias in Lasso for statistical inferences (Candes and Tao 2007, Fan and Li 2001, Zou 2006, Zhang 2010). For instance, with the hard-thresholding penalty $p_\lambda(t) = 0.5[\lambda^2 - (\lambda - t)_+^2]\mathbf{1}(t \geq 0)$, we can obtain the hard thresholding estimator in (4.5). In the case of orthonormal design, the hard thresholding penalty is also equivalent to the L_0-penalty $p_\lambda(t) = 0.5\lambda^2\mathbf{1}(t \neq 0)$. However, for nonorthonormal designs, although nonconvex regularization can be beneficial in selecting important covariates in model (2.3), additional computational and theoretical questions arise due to the nonconvexity of the penalty function.

Both HTR and LAGS try to mimic best subset selection, while avoiding various issues associated with convex/nonconvex penalty functions used in $Q(\boldsymbol{\beta})$. Specifically, we keep the L_1-penalty function $p_\lambda(t) = \lambda|t|$, whereas we replace the loss function in $Q(\boldsymbol{\beta})$ by the score equation (or risk function) of $\boldsymbol{\beta}$. In the case of orthonormal design, HTR reduces to

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{j=1}^{p} \hat{w}_j |\beta_j - \hat{\beta}_j^{ols}| + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{4.7}$$

whose solutions are given by

$$\hat{\beta}_{HTR,j} = \hat{\beta}_j^{ols} \mathbf{1}(\lambda \leq \hat{w}_j) \text{ for } j = 1, \cdots, p. \tag{4.8}$$

By taking the ridge estimator, we obtain $\hat{w}_j = \hat{\beta}_j^{ols}/(1+\lambda_{init})$ and thus $\hat{\boldsymbol{\beta}}_{HTR}$ reduces to the hard thresholding estimator in (4.5) for some value of λ. We can also use the 'biased' lasso estimate $\hat{\beta}_{lasso,j}$ to construct $\hat{w}_j$ in the first stage and then calculate an unbiased estimator $\hat{\boldsymbol{\beta}}_{HTR}$ by calibrating the bias in $\hat{\beta}_{lasso,j}$. Thus, for HTR, we only need a coarse initial estimator in the first stage, which could then help us in identifying the activation set S of the true $\boldsymbol{\beta}^\circ$.

We make a note that HTR is different from the hard-thresholding procedure. Given $\hat{\boldsymbol{\beta}}_{init}$ and $\lambda_n > 0$, the hard thresholding (HT) estimator $\hat{\boldsymbol{\beta}}^{HT}$ is defined as

$$\hat{\boldsymbol{\beta}}^{HT} = \begin{cases} \hat{\boldsymbol{\beta}}_{init}, \text{ if } |\hat{\boldsymbol{\beta}}_{init}| \geq \lambda_n, \\ 0, \text{ if } |\hat{\boldsymbol{\beta}}_{init}| \leq \lambda_n. \end{cases} \tag{4.9}$$

The hard-thresholding rule aims to remove the false positives at the second stage, while largely preserving the parameter estimator calculated in the first stage. In contrast, our HTR always re-estimates $\boldsymbol{\beta}$ in order to calibrate the estimation bias introduced in the first stage. Therefore, a coarse initial estimator of $\boldsymbol{\beta}$ is sufficient in the first stage of HTR.

4.1.3 Theoretical Results

We formally establish the strong oracle property of $\hat{\boldsymbol{\beta}}_{HTR}$, when the number of parameters is large and grows with the sample size n. We start with the following regularity conditions. Throughout the paper, the following conditions are needed to facilitate the technical details, although they may not be the weakest conditions.

Regularity Conditions (RCs)

(RC1) $0 < b \leq \lambda_{\min}(n^{-1}\mathbf{X}^T\mathbf{X}) \leq \lambda_{\max}(n^{-1}\mathbf{X}^T\mathbf{X}) \leq B < \infty$.

(RC2) $\lim_{n\to\infty} \log(p)/\log(n) \leq v$ for some $0 \leq v < 1$.

(RC3) $\lambda n^{-1/2} \to 0$ and $\lambda n^{1/2(\gamma - v(\gamma+1))} \to \infty$.

(RC4) The initial estimates $\hat{\boldsymbol{\beta}}_{init}$ satisfy $E[||\hat{\boldsymbol{\beta}}_{init} - \boldsymbol{\beta}||_2^2|\mathbf{X}] = O(pn^{-1})$.

Remarks. Condition (RC1) assumes that the predictor matrix has reasonably good behavior, which is also considered in Fan and Peng (2004). Condition (RC2) specifies that the growth rate of p is at most a polynomial, that is, $p = O(n^v)$, $v < 1$. It is worth pointing out that Condition (RC2) is weaker than that used in Fan and Peng (2004), for which they assume that p satisfies $p^3 = o(n)$. Condition (RC3) specifies the relationship between λ and n. To construct the risk calibration weight matrix $\widehat{\mathbf{W}}$, we take a fixed γ such that $\gamma > 2v/(1-v)$. Condition (RC4) requires that the initial estimator used in the first stage has a reasonably good behavior in terms of the risk error bound. Such an error bound is generally available for many standard estimators of $\boldsymbol{\beta}$.

As an illustration, we show below that the ridge estimator used in the first stage satisfies (RC4) as given in the following proposition.

Proposition 4.1.1 *(Risk Error For Ridge Estimates) Under (RC1),* $\hat{\boldsymbol{\beta}}^{ridge}$ *satisfies*

$$E[||\hat{\boldsymbol{\beta}}^{ridge} - \boldsymbol{\beta}||_2^2|\mathbf{X}] \leq 2\frac{\lambda_{init}^2||\boldsymbol{\beta}^\circ||_2^2 + \sigma^2 npB}{n^2 b^2}. \tag{4.10}$$

Furthermore, if $\lambda_{init}^2||\boldsymbol{\beta}^\circ||_2^2 = O(np)$*, then we have*

$$E[||\hat{\boldsymbol{\beta}}^{ridge} - \boldsymbol{\beta}||_2^2|\mathbf{X}] = O(\frac{p}{n}). \tag{4.11}$$

We next study the strong oracle properties of $\hat{\boldsymbol{\beta}}_{HTR}$. Before we state the main theorem, we introduce the oracle estimator, denoted as $\hat{\boldsymbol{\beta}}^\circ$, as

$$\hat{\boldsymbol{\beta}}^\circ = \underset{\boldsymbol{\beta},\beta_j=0,\forall j\notin S}{\operatorname{argmin}} \frac{1}{2n}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 = \begin{pmatrix} (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y} \\ \mathbf{0} \end{pmatrix}, \tag{4.12}$$

in which without loss of generality, it is assumed that the first s regression coefficients are nonzero and the remaining $p-s$ regression coefficients are zero. Moreover, $\mathbf{X}_1$ is the corresponding design matrix for the first s regression coefficients. Theorem below provides the strong oracle property of $\hat{\boldsymbol{\beta}}_{HTR}$.

Theorem 4.1.2 *(Strong Oracle Property of* $\hat{\boldsymbol{\beta}}_{HTR}$*) Assume that conditions (RC1)-(RC4) hold. Then, as* $n \to \infty$*, we have*

$$Pr(\hat{\boldsymbol{\beta}}_{HTR} = \hat{\boldsymbol{\beta}}^\circ) \to 1. \tag{4.13}$$

Combining Proposition 2.1 and Theorem 2.2 yields the strong oracle property $\hat{\boldsymbol{\beta}}_{HTR}$, when we set $\hat{\beta}_{init} = (\mathbf{X}^T\mathbf{X} + \lambda_{init}\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$ in the first stage. Our result gives the strong oracle property under very mild conditions by only assuming $\lambda n^{-1/2} \to 0$ and $\lambda n^{1/2(\gamma - v(\gamma+1))} \to \infty$ in (RC3). We shall compare our result with adaptive Lasso in the fixed dimension setting. Adaptive lasso achieves oracle property by requiring $\lambda n^{1/2} \to 0$, or equivalently, the bias term λ goes to 0 with faster rate than $n^{-1/2}$. However, in HTR, the bias term λ can diverge to ∞ with no faster rate than $n^{1/2}$. This can further validate the superiority of HTR estimator: the thresholding level λ is only used to shut down the noise without introducing any bias term to the final estimator.

4.2 HTR under Ultra-High Dimensional Setting

4.2.1 Ultra-High Dimensional HTR

We discuss how to extend HTR for the ultra-high dimensional setting with $p >> n$. For instance, it is common to assume that p may grow at an exponential rate in n. In this case, the standard HTR in (4.1) may fail for $p >> n$. In particular, condition (RC1) fails, since $\lambda_{\min}(n^{-1}\mathbf{X}^T\mathbf{X}) = 0$ for $p > n$. Thus, we need to use a new covariance matrix of predictors $\mathbf{x}$, denoted by $\tilde{\Sigma}_X$, which is positive definite, to replace $n^{-1}\mathbf{X}^T\mathbf{X}$ in (4.1). The use of a positive-definite $\tilde{\Sigma}_X$ to replace $n^{-1}\mathbf{X}^T\mathbf{X}$ is also very common in the regularization literature. For instance, in Zou and Trevor (2005), the elastic net estimator for model (2.3) is defined as

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\boldsymbol{\beta}^T n\tilde{\Sigma}_X\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \lambda||\boldsymbol{\beta}||_1\}, \tag{4.14}$$

in which $n\tilde{\Sigma}_X = (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}_p)/(1+\lambda_2)$ for some $\lambda_2 > 0$.

Our new ultra-high dimensional HTR algorithm for $p \gg n$ is also a two-stage approach as follows.

1. Compute $\hat{\boldsymbol{\beta}}_{init}$, which satisfies the following estimation error bound

$$||\hat{\boldsymbol{\beta}}_{init} - \boldsymbol{\beta}^\circ||_2 \leq C_0\sqrt{\frac{s\log(p)}{n}} \tag{4.15}$$

in a large probability set $\mathcal{J}_0$, that is, $\Pr(\mathcal{J}_0) = 1 - \delta_{n,p,s} \to 1$ or $\delta_{n,p,s} = o(1)$. Specifically, we use the Lasso estimator of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}_{lasso}$, as a candidate of $\hat{\boldsymbol{\beta}}_{init}$, since it has been shown in Zhang and Huang (2008) that (4.15) holds for $\hat{\boldsymbol{\beta}}_{lasso}$ under the sparse Riesz condition. We may use other regularization estimators of $\boldsymbol{\beta}$, such as the Dantzig estimator, since the error bound (4.15) is widely available for them in the ultra-high dimensional framework.

2. Construct $\widehat{\mathbf{W}}$ and estimate $\hat{\boldsymbol{\beta}}_{HTR}$ according to

$$\hat{\boldsymbol{\beta}}_{HTR} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}||\widehat{\mathbf{W}}(\mathbf{X}^T\mathbf{y} - n\tilde{\Sigma}_X\boldsymbol{\beta})||_1 + \lambda||\boldsymbol{\beta}||_1. \tag{4.16}$$

We will show below that our ultra-high dimensional HTR is a general framework for carrying out screening, variable selection, and estimation. We first establish a connection between ultra-high dimensional HTR and Sure Independence Screening (SIS) when p is much larger than n. With a large dimensionality p, the computational cost and estimation accuracy are major difficulties for any statistical method. To overcome such difficulties, Fan and Lv (2008) introduced the SIS methodology to reduce dimensionality from a high p to a relatively large scale d_n with $d_n \leq n$. Specifically, let $\boldsymbol{\omega} = \mathbf{X}^T\mathbf{y} = (\tilde{\omega}_1, \ldots, \tilde{\omega}_p)^T$ be a $p \times 1$ vector of marginal correlations of predictors with the response variable. The standard SIS method is to select the features according to their marginal correlations with the response variable contained in $\boldsymbol{\omega}$, and then filter out those with weak marginal correlations with the response variable. This SIS procedure is equivalent to a special case of HTR by taking $\widehat{\mathbf{W}} = \operatorname{diag}(|\tilde{\omega}_1|, \ldots, |\tilde{\omega}_p|)$ and $\frac{1}{n}\tilde{\Sigma}_X = \operatorname{diag}\{\mathbf{X}\mathbf{X}^T\} = \mathbf{I}_p$ in (4.16). Thus, (4.16) reduces to

$$\hat{\beta}_{HTR,j} = \tilde{w}_j\mathbf{1}(|\tilde{w}_j| \geq \lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\sum_{j=1}^{p}[|\tilde{w}_j||\tilde{w}_j - \beta_j| + \lambda|\beta_j|]\}. \tag{4.17}$$

Without loss of generality, it is assumes that $|\tilde{w}_1| > |\tilde{w}_2| > \cdots > |\tilde{w}_p|$. For any given $q \in (0, 1)$, we can select the covariates corresponding to the first $[qn]$ largest $|\tilde{w}_j|$s by taking $\lambda = |\tilde{w}_{[qn]}|$ in (4.17), where $[qn]$ denotes the integer part of qn. Furthermore, we may combine the order of $\{\tilde{w}_j\}_j$ learned from SIS with HTR (SIS+HTR) to recalculate $\hat{\boldsymbol{\beta}}_{HTR}$.

Second, we show that our HTR procedure allows us to extend SIS to more complex settings, when predictors may be highly correlated. The incorporation of the correlation structure among predictors is critical for better variable selection and estimation in model (2.3). An important strategy is to balance between noise accumulation and correlation modeling. Without loss of generality, we assume that the true covariance matrix of $\mathbf{x}$, denoted

as $\Sigma_{\mathbf{x}}$, has a geometric decay structure and then we can use its regularized bandable convariance estimator, denoted as $\tilde{\Sigma}_X$, to approximate $\Sigma_{\mathbf{x}}$ (Bickel and Levina 2008b). Extensions to other covariance structures can also be done by using other regularized estimators in the literature (Cai et al. 2010, Lam and Fan 2009, Rothman et al. 2009, Fan et al. 2013). Specifically, we set $\tilde{\boldsymbol{\omega}} = \tilde{\Sigma}_X^{-1}\mathbf{X}^T\mathbf{y}$ and $\widehat{\mathbf{W}} = \text{diag}(|\tilde{\boldsymbol{\omega}}|)$. In this case, (4.16) reduces to

$$\hat{\boldsymbol{\beta}}_{HTR} = \underset{\boldsymbol{\beta}}{\text{argmin}}\{||\text{diag}(|\tilde{\boldsymbol{\omega}}|)(\frac{1}{n}\mathbf{X}^T\mathbf{y} - \tilde{\Sigma}_X\boldsymbol{\beta})||_1 + \lambda||\boldsymbol{\beta}||_1\}. \tag{4.18}$$

Since we explicitly account for the joint information of all covariates by regularizing their covariance matrix estimation through a de-correlation procedure instead of using the independence rule, we may call (4.18) as a Sure Correlation Screening (SCS) procedure, which could avoid the faithful assumption used in Fan and Lv (2008).

4.2.2 Theoretical Results

We formally investigate the strong oracle property of $\hat{\boldsymbol{\beta}}_{HTR}$ under the ultra-high dimensional scenario. We start with the following regularity condition on $\Sigma_{\mathbf{x}}$. Specifically, throughout the paper, it is assumed that $\Sigma_{\mathbf{x}}$ belongs to a well behaved covariance class $\mathcal{U}(\varepsilon_0, \alpha, C_1)$, which is defined as

$$\begin{aligned}\mathcal{U}(\varepsilon_0, \alpha, C_1) \quad = \quad & \{\Sigma = (\sigma_{jj'}) : \max_j \sum_{j'}\{|\sigma_{j'j}| : |j' - j| > k\} \leq C_1 k^{-\alpha} \text{ for all } k > 0 \\ & \text{and } 0 < \varepsilon_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\varepsilon_0\},\end{aligned}$$

where ε_0, C_1, and α are positive scalars. The condition $\Sigma_{\mathbf{x}} \in \mathcal{U}(\varepsilon_0, \alpha, C_1)$ basically requires that $\Sigma_{\mathbf{x}}$ be bandable. Such a condition on $\Sigma_{\mathbf{x}}$ can be relaxed by employing different covariance estimators (Bickel and Levina 2008b, Cai et al. 2010, Lam and Fan 2009, Rothman et al. 2009, Fan et al. 2013).

We also introduce the L_∞ Correlation Condition (LCC) for model identifiability. For a given set S with cardinality p_S and its complement $S^C = \{1, \cdots, p\}/S$ with cardinality

$p_{S^C} = p - p_S$, we consider a partition of the $p \times p$ matrix Σ according to (S, S^C) as follows:

$$\Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{SS^C} \\ \Sigma_{S^C S} & \Sigma_{S^C S^C} \end{pmatrix},$$

where $\Sigma_{S_1 S_2}$ is a $p_{S_1} \times p_{S_2}$ matrix corresponding to indices in S_1 and S_2, in which S_1 and S_2 are equal to either S or S^C. We say that (Σ, S) satisfies the L_∞ correlation condition, if there exists a $u_0(n, p, p_S) > 0$ such that

$$\min_{||\tau_S||_\infty = 1, ||\tau_{S^C}||_\infty = 1} ||\Sigma_{SS}\tau_S + \Sigma_{SS^C}\tau_{S^C}||_\infty > u_0(n, p, p_S), \tag{4.19}$$

where τ_S and τ_{S^C} are $p_S \times 1$ and $(p - p_S) \times 1$ vectors, respectively.

The L_∞ correlation condition is used to rule out the case of strong collinearity in the same spirit of condition 4 in Fan and Lv (2008). The sample version of LCC closely resembles the irrepresentable condition first proposed by Zhao and Yu (2006). The ir-representable condition is equivalent to putting a regularization constraint on the regression coefficients of the irrelevant covariates $\mathbf{X}_{S^C}$ on the relevant covariates $\mathbf{X}_S$, $||\Sigma_{SS}^{-1}\Sigma_{SS^C}||_1 \leq 1 - u_0(n, p, p_S)$ for some constant $u_0(n, p, p_S) > 0$. Similar to the ir-representable condition, if we put the constraint in the L_∞ norm rather than the L_1 norm, i.e. $||\Sigma_{SS}^{-1}\Sigma_{SS^C}||_\infty \leq 1 - u_0(n, p, p_S)$ and hold s fixed, this would imply the LCC condition by observing

$$\begin{aligned} \min_{\tau \in \Omega_0} ||\Sigma_{SS}\tau_S + \Sigma_{SS^C}\tau_{S^C}||_\infty &\geq \min_{\tau \in \Omega_0} ||\Sigma_{SS}^{-1}(\tau_S - \Sigma_{SS}^{-1}\Sigma_{SS^C}\tau_{S^C})||_\infty \\ &\geq \min_{\tau \in \Omega_0} \frac{1}{\sqrt{p_S}} \lambda_{min}(\Sigma_{SS}) ||\tau_S - \Sigma_{SS}^{-1}\Sigma_{SS^C}\tau_{S^C}||_\infty \\ &\geq \frac{\varepsilon_0}{\sqrt{p_S}} u_0(n, p, p_S), \end{aligned} \tag{4.20}$$

where $\Omega_0 = \{||\tau_S||_\infty = 1, ||\tau_{S^C}||_\infty = 1\}$. Generally, we allow $u_0(n, p, p_S)$ to diverge to 0.

We examine an example of Σ to show that for some Σ, the LCC condition holds, whereas the ir-representable condition does not. Specifically, we consider a specific Σ^0 with $\Sigma^0_{SS} = I_{p_S}$, $\Sigma^0_{S^C S^C} = I_{p-p_S}$, and $\Sigma^0_{SS^C} = (\Sigma^0_{S^C S})^T = [\mathbf{1}_{p_S}\rho/\sqrt{p_S}, \mathbf{0}, \cdots, \mathbf{0}]$, where $\mathbf{1}_{p_S}$

is a $p_S \times 1$ vector with all ones. Therefore, the LCC condition allows us to go beyond the ir-representable condition for consistent variable selection.

Proposition 4.2.1 *For $S = \{1, \ldots, p_S\}$ and Σ^0 defined as above, (Σ^0, S) satisfies the LCC condition, but not the ir-representable condition.*

We define the oracle estimator of $\boldsymbol{\beta}$ in the ultra-high dimensional setting as

$$\tilde{\boldsymbol{\beta}}^\circ = (\ \{(\tilde{\Sigma}_{SS})^{-1}\mathbf{X}_S^T\mathbf{y}\}^T, \mathbf{0}^T\ )^T, \tag{4.21}$$

where $\tilde{\Sigma}_{X,SS}$ denotes the submatrix of $\tilde{\Sigma}_X$ corresponding to the indices in the true active set S. Note that the difference between $\tilde{\boldsymbol{\beta}}^\circ$ and the oracle least squares estimate $\hat{\boldsymbol{\beta}}^\circ$ is very small, since $||\tilde{\Sigma}_{X,SS} - \Sigma_{X,SS}||_2^1 \leq ||\tilde{\Sigma}_X - \Sigma_X||_2^1 = O_p((\frac{\log(p)}{n})^{\alpha/(2(\alpha+1))})$ for $\Sigma_X \in \mathcal{U}(\varepsilon_0, \alpha, C_1)$ (Bickel and Levina 2008b). Moreover, if the ordinary least squares estimator is desirable, especially when s/n is moderate, we can first identify an initial active set, denoted as S_n, and then we can calculate $\hat{\boldsymbol{\beta}}_{ref} = (\mathbf{X}_{S_n}^T\mathbf{X}_{S_n})^{-1}\mathbf{X}_{S_n}^T\mathbf{y}$. Before we present the main results below, we let $\Sigma_{k_n} = B_{k_n}(\Sigma) = (\sigma_{ij}1_{(|i-j|\leq k_n)})$.

Theorem 4.2.2 *(Strong Oracle Property of $\hat{\boldsymbol{\beta}}_{HTR}$ under $p >> n$ with thresholded property) Suppose that $\Sigma_{\mathbf{x}} \in \mathcal{U}(\varepsilon_0, \alpha, C_1)$, (4.15) holds, and $(B_{k_n}(\Sigma_x), \mathcal{S}_n)$ satisfies the LCC condition. If the tuning parameter λ satisfies*

$$m < \lambda < M$$

for $k_n \asymp (\log p/n)^{-1/(2(\alpha+1))}$ and t^0 defined in Lemma 6.1, where

$$m \doteq C_0^\gamma(2k_n+1)(s\log(p)/n)^{\gamma/2}\max\{{\epsilon_0}^{-1}, \sqrt{\frac{2(\eta+1)}{\gamma(\epsilon_0,\delta)}}(\frac{\log p}{n})^{1/2}\},$$

and $M \doteq [u_0(n,p,s) - 2t^0 - 2k_n^\alpha](\min_{j\in S}|\beta_j^\circ| - C_0\sqrt{s_0\log p/n})^\gamma$, then with probability at least $1 - \delta_{n,p,s} - 3p^{-\eta}$, we have

$$\hat{\boldsymbol{\beta}}(\lambda) = \tilde{\boldsymbol{\beta}}^\circ. \tag{4.22}$$

Theorem 5.3.1 quantified our HTR estimator under the ultra-high dimensional scenario. Assuming that $\eta_0 \doteq \min_{j\in S}|\beta_j^\circ| > C_0\sqrt{s\log(p)/n}$ and $u_0(n,p,s)$ is fixed, we roughly require $\eta_0^\gamma \gtrsim \lambda \gtrsim (s\log(p)/n)^{\gamma/2}(2k_n+1)$. However, in Wang et al. (2013a), the calibrated CCCP method identifies the oracle estimator when $\eta \gg \lambda \gg s\sqrt{\log(p)/n}$. We point out an interesting phenomenon: within the range (m, M) with m and M defined in the above theorem, $\hat{\boldsymbol{\beta}}_{HTR}$ stays at oracle estimator $\tilde{\boldsymbol{\beta}}^\circ$. This agrees with our intuition that HTR's solution path has a piece-wise constant property. We mention that our result is not directly comparable with the calibrated CCCP method and any other method in the literature as we only require that $M > \lambda > m$ rather than $M \gg \lambda \gg m$. Finally, Theorem 5.3.1 is in line with the important theoretical properties of L_0 penalized regression considered in Zheng et al. (2013). This may further validate our HTR method.

4.3 Numerical Examples

4.3.1 Simulation Study

Continuous responses were generated according to model (2.3) with $\boldsymbol{\beta}^\circ = (3, 2, 0, 0, -1.5, \underbrace{0, \ldots, 0}_{p-5})^T$ and $n = 100$. Moreover, in model (2.3), $\mathbf{x}_i$ follows the $N(0, \Sigma_X)$ distribution with covariance matrix Σ_X and ϵ_i is independent of $\mathbf{x}_i$ and has a normal distribution with mean 0 and standard deviation $\sigma = 2$. Write $\Sigma_X = \sigma(\rho_{ij})$, we consider three different correlation structures of (ρ_{ij}) including

- Case 1: independent correlation design with $(\rho_{ij}) = \text{diag}(1, \cdots, 1)$;
- Case 2: weak correlation design with $\rho_{ij} = 0.30^{|i-j|}$;
- Case 3: relatively strong correlation design with $\rho_{ij} = 0.95^{|i-j|}$.

We consider both relatively high dimension $p = 40$ and ultra-high dimension case $p = 2000 \gg n$.

We investigate the sparsity recovery and estimation properties of the HTR estimator via numerical simulations. We compared the HTR estimator with the following estimators: the

oracle estimator which assumes the availability of the active set S_0; the adaptive lasso estimator proposed by Zou (2006); the smoothly clipped absolute deviation (SCAD) estimator (Fan and Li 2001); and the minimax concave penalty (MCP) estimator with $a = 3$ Zhang (2010). For SCAD, $n^{1/2}$–fold cross-validation was used to select the tuning parameter λ; for ALasso and HTR, sequential tuning in Bühlmann and Geer (2011) was used; and the MCP estimator was computed using the R package PLUS with the theoretically optimal tuning parameter value $\lambda = \sigma\sqrt{2/np}$. For the case $p = 30$, we also computed regularized estimators based on LAGS. To estimate the bandable covariance estimator $\tilde{\Sigma}_X$ in HTR, the banding parameter was selected by cross validation as described in Bickel and Levina (2008b). To further demonstrate the performance by using the regularized covariance matrix, we also compared the HTR estimates with the sample covariance matrix, and the independence covariance matrix and denoted them as HTR_{sam} and HTR_{ind}, respectively.

For each simulation setting, we generated 100 simulated data sets and applied different estimators to each dataset. Then, we calculated different statistics for each estimator and included them in Tables 4.1, 4.2, 4.3 and 4.4. We calculated the mean and median of $|\hat{\beta}_i| - |\beta_i|$ with $i = 1, 2, 3$ in order to measure the downward shrinkage bias. To measure the sparsity recovery, we calculated the mean and median of number of zero coefficients incorrectly estimated to be nonzero (i.e. false positive, denoted as FP) and the mean and median of number of nonzero coefficients correctly estimated to be nonzero (i.e. true positive, denoted by TP). To measure the estimation accuracy, we calculated the mean and median squared error (MSE) and the mean and median absolute error (MAE).

It is not surprising that Lasso always overfits. Other procedures improve the performance of Lasso by reducing the estimation bias and the false positive rate. The best overall performance is achieved by the HTR estimator with relatively small shrinkage bias, MSE, MAE, and FP. The MCP and SCAD also have overall fine performance. In the relatively high dimensional ($p = 30$) example, HTR outperforms LAGS in all three cases. When the dimension is 2000, in all cases, the HTR with sample covariance matrix encourages false selections and thus it has worse performance compared with that with regularized covariance

estimator. When the correlation structure gets stronger, ignoring the correlation structure would produce too sparse solution and miss true variables, which verifies our conjecture. This verifies the effectiveness of using regularized covariance matrix in the regression procedure.

Table 4.1: Mean of simulation results for $p = 40$: $|\hat{\beta}_1| - |\beta_1|$, $|\hat{\beta}_2| - |\beta_2|$, $|\hat{\beta}_3| - |\beta_3|$, MSE, MAE, TP, and FP. For each case, 100 simulated data sets were used.

Case	Methods	$\lvert\hat{\beta}_1\rvert - \lvert\beta_1\rvert$	$\lvert\hat{\beta}_2\rvert - \lvert\beta_2\rvert$	$\lvert\hat{\beta}_3\rvert - \lvert\beta_3\rvert$	MSE	MAE	TP	FP
1	Oracle	-0.0054	0.0219	-0.0017	0.0373	0.2670	NA	NA
	Lasso	-0.1004	-0.0788	-0.1399	0.1158	0.7599	3.00	10.76
	ALasso	-0.0090	0.0058	-0.0307	0.1061	0.7603	3.00	9.10
	SCAD	-0.0042	0.0215	-0.0054	0.0387	0.3199	3.00	5.91
	MCP	-0.0059	0.0214	-0.0031	0.0378	0.2743	3.00	5.18
	HTR	-0.0052	0.0174	-0.0097	0.0679	0.0679	3.00	5.58
	LAGS	-0.0158	0.0182	-0.0478	0.3770	1.0466	3.00	5.77
2	Oracle	-0.0054	0.0219	-0.0017	0.0373	0.2670	NA	NA
	Lasso	-0.1004	-0.0788	-0.1399	0.1158	0.7599	3.00	10.76
	ALasso	-0.0090	0.0058	-0.0307	0.1061	0.7603	3.00	9.10
	SCAD	-0.0042	0.0215	-0.0054	0.0387	0.3199	3.00	5.91
	MCP	-0.0059	0.0214	-0.0031	0.0378	0.2743	3.00	5.18
	HTR	-0.0052	0.0174	-0.0097	0.0679	0.0679	3.00	5.58
	LAGS	-0.0158	0.0182	-0.0478	0.3770	1.0466	3.00	5.77
3	Oracle	-0.0105	0.0073	0.0002	0.0487	0.2949	NA	NA
	Lasso	-0.1412	-0.1160	-0.1315	0.1880	1.0013	3.00	11.43
	ALasso	-0.0113	0.0002	-0.0350	0.1923	0.9959	3.00	9.46
	SCAD	-0.0072	0.0083	-0.0041	0.0183	0.3721	3.00	5.62
	MCP	-0.0104	0.0054	-0.0005	0.0487	0.2955	3.00	5.04
	HTR	-0.0212	-0.0027	-0.0103	0.0918	0.0918	3.00	5.60
	LAGS	0.0856	-0.0485	-0.0619	0.6434	1.3617	3.00	5.69

4.3.2 Bardet-Biedl syndrome gene expression study

We applied HTR to the Bardet Biedl syndrome gene expression study in Scheetz et al. (2006). For this data set, F1 animals were intercrossed and 120 twelve-week-old male offspring were selected for tissue harvesting from the eyes and for microarray analysis. The microarrays used to analyze the RNA from the eyes of these animals contain more than $31,042$ different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array). The intensity values were normalized using the RMA (robust multi-chip averaging, Bolstad et al. (2003), Irizarry et al. (2003)) method to obtain summary expression values for each probe set.

Table 4.2: Median of simulation results for $p = 40$: $|\hat{\beta}_1| - |\beta_1|$, $|\hat{\beta}_2| - |\beta_2|$, $|\hat{\beta}_3| - |\beta_3|$, MSE, MAE, TP, and FP. For each case, 100 simulated data sets were used.

Case	Methods	$\lvert\hat{\beta}_1\rvert - \lvert\beta_1\rvert$	$\lvert\hat{\beta}_2\rvert - \lvert\beta_2\rvert$	$\lvert\hat{\beta}_3\rvert - \lvert\beta_3\rvert$	MSE	MAE	TP	FP
1	Oracle	0.0086	-0.0171	0.0016	0.0221	0.2283	NA	NA
	Lasso	-0.1127	-0.1430	-0.1274	0.1072	0.7584	3.00	11.00
	ALasso	0.0045	-0.0304	-0.0272	0.0702	0.8080	3.00	8.50
	SCAD	0.0125	-0.0156	-0.0032	0.0241	0.2775	3.00	5.00
	MCP	0.0121	-0.0156	-0.0032	0.0221	0.2334	3.00	5.00
	HTR	0.0104	-0.0210	-0.0020	0.0521	0.0797	3.00	6.00
	LAGS	-0.0366	0.0097	0.0431	0.2264	0.9493	3.00	5.00
2	Oracle	-0.0054	0.0219	-0.0017	0.0270	0.2393	NA	NA
	Lasso	-0.1004	-0.0788	-0.1399	0.1062	0.6894	3.00	10.00
	ALasso	-0.0052	-0.0018	-0.0214	0.0736	0.7603	3.00	8.00
	SCAD	-0.0002	0.0270	0.0029	0.0188	0.3199	3.00	5.00
	MCP	0.0012	0.0246	-0.0003	0.0274	0.2743	3.00	5.00
	HTR	0.0051	0.0156	-0.0048	0.0482	0.0679	3.00	5.00
	LAGS	-0.0048	0.0121	-0.0458	0.3219	1.0466	3.00	5.00
3	Oracle	-0.0105	0.0073	0.0002	0.0350	0.2718	NA	NA
	Lasso	-0.1412	-0.1160	-0.1315	0.1460	0.8972	3.00	10.00
	ALasso	-0.0061	-0.0182	-0.0274	0.1167	0.9959	3.00	8.00
	SCAD	-0.0073	-0.0061	-0.0150	0.0082	0.3721	3.00	5.00
	MCP	-0.0116	-0.0112	-0.0087	0.0353	0.2955	3.00	5.00
	HTR	-0.0204	-0.0119	-0.0040	0.0484	0.0918	3.00	5.00
	LAGS	0.0317	-0.0250	-0.0874	0.5245	1.3617	3.00	5.00

Table 4.3: Mean of simulation results for $p = 2000$: we report $|\hat{\beta}_1|-|\beta_1|$, $|\hat{\beta}_2|-|\beta_2|$, $|\hat{\beta}_3|-|\beta_3|$, MSE, MAE, TP, and FP. For each case, 100 simulated data sets were used.

Case	Methods	$\lvert\hat{\beta}_1\rvert - \lvert\beta_1\rvert$	$\lvert\hat{\beta}_2\rvert - \lvert\beta_2\rvert$	$\lvert\hat{\beta}_3\rvert - \lvert\beta_3\rvert$	MSE	MAE	TP	FP
1	Oracle	0.0041	-0.0096	-0.0123	0.0316	0.2489	NA	NA
	Lasso	-0.2878	-0.3141	-0.3094	0.3895	1.7974	3.00	27.17
	ALasso	-0.1560	-0.1685	-0.1683	0.5123	1.8002	3.00	23.81
	SCAD	0.0046	-0.0095	-0.0186	0.0472	0.4175	3.00	9.08
	MCP	0.0040	-0.0104	-0.0106	0.0324	0.2562	3.00	5.13
	HTR	0.0030	-0.0122	-0.0122	0.0363	0.2621	3.00	5.03
	HTR_{sam}	0.0039	-0.0283	-0.0113	0.0812	0.2857	2.99	5.04
	HTR_{ind}	0.0035	-0.0106	-0.0122	0.0346	0.2563	3.00	5.02
2	Oracle	-0.0016	0.0058	-0.0087	0.0352	0.2565	NA	NA
	Lasso	-0.2340	-0.2153	-0.3020	0.2963	1.5513	3.00	25.46
	ALasso	-0.1185	-0.1008	-0.1580	0.4462	1.5513	3.00	22.23
	SCAD	-0.0017	0.0066	-0.0116	0.0462	0.4058	3.00	9.22
	MCP	-0.0015	0.0058	-0.0105	0.0363	0.2607	3.00	5.09
	HTR	0.0082	0.0077	-0.0397	0.0635	0.2798	2.99	5.01
	HTR_{sam}	-0.0021	0.0048	-0.0105	0.0392	0.2654	3.00	5.02
	HTR_{ind}	0.0098	0.0117	-0.0528	0.1096	0.3092	2.97	5.00
3	Oracle	0.0510	-0.0228	-0.0160	0.3832	0.8544	NA	NA
	Lasso	-0.0874	-0.1487	-0.3443	0.7433	1.7430	3.00	14.87
	ALasso	0.0389	-0.1478	-0.2206	0.9337	1.7430	3.00	12.47
	SCAD	-0.0112	0.1355	-0.2903	14.2929	5.7731	1.69	7.19
	MCP	1.4171	-2.0000	-0.4810	9.6553	5.4525	1.52	5.93
	HTR	0.1147	-0.1520	-0.1663	1.2495	1.4264	2.85	5.32
	HTR_{sam}	0.1851	-0.2627	-0.2174	1.7857	1.7192	2.75	5.40
	HTR_{ind}	2.3034	-1.1562	-1.5000	16.3439	6.5666	1.08	5.00

Table 4.4: Median of simulation results for $p = 2000$: we report $|\hat{\beta}_1| - |\beta_1|$, $|\hat{\beta}_2| - |\beta_2|$, $|\hat{\beta}_3| - |\beta_3|$, MSE, MAE, TP, and FP. For each case , 100 simulated data sets were used.

Case	Methods	$\lvert\hat{\beta}_1\rvert - \lvert\beta_1\rvert$	$\lvert\hat{\beta}_2\rvert - \lvert\beta_2\rvert$	$\lvert\hat{\beta}_3\rvert - \lvert\beta_3\rvert$	MSE	MAE	TP	FP
1	Oracle	-0.0041	-0.0075	-0.0148	0.0263	0.2364	NA	NA
	Lasso	-0.2787	-0.3063	-0.2874	0.3670	1.6493	3.00	24.00
	ALasso	-0.1621	-0.1744	-0.1787	0.5236	1.6493	3.00	22.00
	SCAD	-0.0017	-0.0073	-0.0121	0.0341	0.2891	3.00	6.00
	MCP	-0.0041	-0.0095	-0.0130	0.0266	0.2442	3.00	5.00
	HTR	-0.0088	-0.0075	-0.0148	0.0269	0.2416	3.00	5.00
	HTR_{sam}	-0.0041	-0.0071	-0.0130	0.0269	0.2416	3.00	5.00
	HTR_{ind}	-0.0068	-0.0075	-0.0148	0.0267	0.2399	3.00	5.00
2	Oracle	0.0052	0.0093	0.0011	0.0274	0.2424	NA	NA
	Lasso	-0.2317	-0.2061	-0.2826	0.2696	1.3675	3.00	19.00
	ALasso	-0.1190	-0.1076	-0.1580	0.4662	1.3675	3.00	18.00
	SCAD	-0.0065	0.0087	0.0001	0.0376	0.3320	3.00	6.00
	MCP	0.0052	0.0093	0.0011	0.0274	0.2506	3.00	5.00
	HTR	0.0070	0.0093	-0.0059	0.0274	0.2424	3.00	5.00
	HTR_{sam}	-0.0009	0.0078	0.0011	0.0277	0.2454	3.00	5.00
	HTR_{ind}	0.0088	0.0103	-0.0059	0.0274	0.2424	3.00	5.00
3	Oracle	0.0247	-0.0256	-0.0367	0.2431	0.7758	NA	NA
	Lasso	-0.0797	-0.1612	-0.3583	0.5996	1.6926	3.00	11.50
	ALasso	0.0197	-0.2193	-0.2073	0.8302	1.6926	3.00	10.00
	SCAD	1.4099	-2.0000	0.2486	7.1601	4.2030	2.00	6.00
	MCP	1.4253	-2.0000	-0.0403	7.9969	4.2880	2.00	6.00
	HTR	0.0997	-0.0912	-0.0647	0.4382	1.0802	3.00	5.00
	HTR_{sam}	0.0948	-0.0951	-0.0749	0.5165	1.1728	3.00	5.00
	HTR_{ind}	3.0845	-2.0000	-1.5000	16.0711	6.6339	1.00	5.00

The outcome of interest is the expression of TRIM32, corresponding to probe 1389163_at, a gene which has been shown to cause Bardet-Biedl syndrome (Chiang et al. 2006), which is a genetic disease of multiple organ systems including the retina. Following Scheetz et al. (2006), we focused on $18,957$ probes out of the $31,042$ probe sets on the array that exhibited a sufficient signal for reliable analysis and at least 2-fold variation in expression.

The aim of this data analysis is to find the genes, whose expressions are correlated with that of gene TRIM32. We used model (2.3) to address this problem and applied different regularization methods in the analysis. We first standardized the probes so that they have mean zero and standard deviation 1. As in Huang et al. (2008), we focused on 3000 probes with the largest variances among the $18,975$ covariates and considered two approaches. The first approach is to regress on the $p = 3000$ probes. The second approach is to regress on the 200 probes among the 3000 with the largest marginal correlation coefficients with TRIM32. We randomly partitioned the data 100 times, each with a training set of size 80 and a test set of 40 observations. The prediction mean squared error was computed within the test set, while the scaled estimators and the lasso estimator with a fixed penalty level λ were computed based on the training set.

In addition, we compared the prediction performance of all the estimators mentioned in the simulations. In each replication, we computed all the regularization estimators based on the training set of 80 observations. The penalty level is selected by 5-fold cross validation over the training data set. Table 4.5 includes the median of downward prediction bias (DPBias), defined as $\sum_{i=1}^{\#\text{test sample}} |(\hat{y}_i)| - |y_i|$, median of the mean squared prediction error (MSPE), and the average selected model size in the 100 replications for $p = 300$ and 2000. For MCP, the tuning parameters were selected by cross validation since the standard deviation of the random error is unknown. HTR works at least as good as, if not better than, ALasso, SCAD, and MCP with much sparser models and small prediction errors. It is worth pointing out that the HTR procedure produces the sparsest solution yet with a well controlled prediction error. Moreover, HTR controls the downward prediction bias well. The performance of the MCP procedure is satisfactory but its optimal performance depends on

another tuning parameter a. In screening or diagnostic applications, it is often important to develop an accurate diagnostic test using as few features as possible in order to control the cost. The same consideration also matters when selecting target genes in gene therapies.

Table 4.5: Gene Expression Data Analysis

p	Method	MSPE	DPBias	avg model size
300	Lasso	0.1511	-0.1412	26.23
	Alasso	0.1573	-0.0934	17.36
	SCAD	0.4728	0.5114	11.23
	MCP	0.4475	0.5874	5.55
	HTR	0.2618	-0.0789	3.94
2000	Lasso	0.2120	-0.1591	33.00
	Alasso	0.1736	-0.0667	22.84
	SCAD	0.2017	-0.1498	12.97
	MCP	0.2699	-0.0588	6.99
	HTR	0.1999	-0.0520	6.42

4.4 Conclusions and Further Discussions

The main contribution of this paper is two fold. First, we have offered a new perspective to achieve unbiased estimation instead of non-convex penalized regression, which can be formulated as a linear programming and thus is computational tractable. The global optimal solution is assured. Secondly, we have proposed a new framework to incorporate the covariance estimator into the regression procedure for better trade off between noise accumulation and correlation modeling and leave the possibility of relaxing conditions for consistent variable selection.

4.5 Proofs

We present the proofs of all theoretical results below.

Proof of Proposition 4.1.1. Note that

$$\hat{\boldsymbol{\beta}}^{ridge}(\lambda) - \boldsymbol{\beta}^{\circ} = -\lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}^{\circ} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}. \tag{4.23}$$

Then, it follows from (RC1) that

$$\begin{aligned}
& E[||\hat{\boldsymbol{\beta}}^{ridge}(\lambda) - \boldsymbol{\beta}^{\circ}||_2^2|\mathbf{X}] = E[|| - \lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}^{\circ} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}||_2^2|\mathbf{X}] \\
\leq\ & 2\lambda^2\{\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda\}^{-2}||\boldsymbol{\beta}^{\circ}||_2^2 + 2\{\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda\}^{-2}E[\boldsymbol{\varepsilon}^T\mathbf{X}\mathbf{X}^T\boldsymbol{\varepsilon}] \\
=\ & 2\{\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda\}^{-2}\{\lambda^2||\boldsymbol{\beta}^{\circ}||_2^2 + \mathrm{Tr}(\mathbf{X}^T\mathbf{X})\sigma^2\} \\
\leq\ & 2\frac{\lambda^2||\boldsymbol{\beta}^{\circ}||_2^2 + \sigma^2 p\lambda_{max}(\mathbf{X}^T\mathbf{X})}{(\lambda_{min}(\mathbf{X}^T\mathbf{X}) + \lambda)^2} \\
\leq\ & 2\frac{\lambda^2||\boldsymbol{\beta}^{\circ}||_2^2 + \sigma^2 npB}{(nb + \lambda)^2} \leq 2\frac{\lambda^2||\boldsymbol{\beta}^{\circ}||_2^2 + \sigma^2 npB}{n^2b^2},
\end{aligned}$$

which yields the proof of Proposition 4.1.1.

Proof of Theorem 5.2.2. The proof of Theorem 5.2.2 consists of two steps. The first step is to show the exact support recovery as

$$\lim_{n\to\infty} \mathrm{P}(S \subset S_n) = 1, \tag{4.24}$$

where $S_n = \{j|\hat{\beta}_{HTR,j} \neq 0\}$. The second step is to show

$$\lim_{n\to\infty} \mathrm{P}(S_n \subseteq \mathcal{S}) = 1, \tag{4.25}$$

We prove (4.24) as follows. It is easy to show that the Karush-Kuhn-Tucker (KKT) conditions of (4.1) lead to

$$\widehat{\mathbf{W}}(\frac{1}{n}\mathbf{X}^T\mathbf{X})\mathrm{sign}(\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{HTR})) = \lambda \times \mathrm{sign}(\hat{\boldsymbol{\beta}}_{HTR}), \tag{4.26}$$

where $\text{sign}(x)$ is the signum function of x. Thus, if $\hat{\beta}_{HTR,j} \neq 0$, then we have

$$\hat{w}_j[\frac{1}{n}\mathbf{X}^T\mathbf{X} \times \text{sign}(\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))]_{(j)} = \lambda \times \text{sign}(\hat{\beta}_{HTR,j}),$$

where $[\mathbf{a}]_{(j)}$ denotes the j-th component of any vector $\mathbf{a}$. Since $|\frac{1}{n}[\mathbf{X}^T\mathbf{X} \times \text{sign}(\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}))]_{(j)}| \leq ||\frac{1}{n}\mathbf{X}^T\mathbf{X}||_\infty$, we have

$$\lambda|\text{sign}(\hat{\beta}_i^{(n)})| \leq \hat{w}_j||\frac{1}{n}\mathbf{X}^T\mathbf{X}||_\infty. \tag{4.27}$$

Therefore, to prove (4.24), it suffices to show that as $n \to \infty$, we have

$$\text{P}(\cup_{j \in S^C}\{\hat{w}_j||\frac{1}{n}\mathbf{X}^T\mathbf{X}||_\infty > \lambda\}) \to 0. \tag{4.28}$$

Write $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$. We now bound the left-hand side (LHS) of (4.28) as follows:

$$\begin{aligned} &\text{P}(\cup_{j \in S^c}\{\hat{w}_j||\hat{\Sigma}||_\infty > \lambda\}) \leq \sum_{j \in S^c} \text{P}(\hat{w}_j||\hat{\Sigma}||_\infty \leq \lambda) \\ \leq &\sum_{j \in S^c} \text{P}(\hat{\beta}^2_{init,j} \geq (\frac{\lambda}{||\hat{\Sigma}||_\infty})^{2/\gamma}) \leq \frac{E||\hat{\boldsymbol{\beta}}_{init} - \boldsymbol{\beta}^\circ||_2^2}{(\lambda/||\hat{\Sigma}||_\infty)^{2/\gamma}} = O((\frac{1}{\lambda n^{\gamma/2 - v/2(\gamma+1)}})^{2/\gamma}), \end{aligned} \tag{4.29}$$

We prove (4.25) as follows. Rewrite the KKT conditions as following,

$$\text{sign}(\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{HTR})) = \lambda \times (\hat{\Sigma})^{-1}\widehat{\mathbf{W}}^{-1}\text{sign}(\hat{\boldsymbol{\beta}}_{HTR}). \tag{4.30}$$

Therefore, to prove (4.25), it suffices to show that as $n \to \infty$, we have

$$\text{P}(\lambda||(\hat{\Sigma})^{-1}||_\infty\{\max \hat{w}_j^{-1}\} < 1 \ \forall j \in S) \to 1. \tag{4.31}$$

The LHS of (4.31) is bounded by

$$\text{(LHS)} \geq \text{P}(\min_{j \in S}(\hat{w}_j) > \lambda||(\hat{\Sigma})^{-1}||_\infty) = \text{P}(\min_{j \in S}(\left|\hat{\beta}_{init,j}\right|) > [\lambda||(\hat{\Sigma})^{-1}||_\infty]^{1/\gamma}). \tag{4.32}$$

Since $\min_{j\in S}\left|\hat{\beta}_{init,j}\right| \geq \min_{j\in S}|\beta_j^\circ| - ||\hat{\beta}_{init,S} - \beta_S^\circ||_\infty \geq \min_{j\in S}|\beta_j^\circ| - ||\hat{\beta}_{init,S} - \beta^\circ||_2$, we have

$$\text{RHS of (4.32)} \geq \mathrm{P}(\min_{j\in S}|\beta_j^\circ| > [\lambda||(\hat{\Sigma})^{-1}||_\infty]^{1/\gamma} + ||\hat{\beta}_{init} - \beta^\circ||_2) \tag{4.33}$$

where $E(||\hat{\beta}_{init} - \beta^\circ||_2^2) = O(\frac{p}{n})$. Further, by assumption RC3, we have

$$\eta > (\frac{B\lambda}{\sqrt{n}}\sqrt{\frac{p}{n}})^{1/\gamma} + \sqrt{\frac{p}{n}}O_p(1) \geq [\frac{\lambda_n}{n}||(\hat{\Sigma})^{-1}||_\infty]^{1/\gamma} + \sqrt{\frac{p}{n}}O_p(1), \tag{4.34}$$

yielding $\lim_{n\to\infty}\mathrm{P}(S_n = S) = 1$.

Denote the event $\{S_n = S\}$ as $\mathcal{J}$. In $\mathcal{J}$, we have

$$X^T(y - X_S\hat{\beta}_S) = \begin{pmatrix} X_S^T(y - X_S\hat{\beta}_S) \\ X_{S^c}^T(y - X_S\hat{\beta}_S) \end{pmatrix}. \tag{4.35}$$

The KKT conditions yield

$$\hat{\beta}_{S_n} = \hat{\beta}_S = (X_S^T X_S)^{-1}X_S^T y = \hat{\beta}_S^\circ. \tag{4.36}$$

It should note that the consistent selection property actually implies the strong oracle property, unlike the dilemma involved in the Lasso. This finishes the proof of Theorem 5.2.2.

Proof of Proposition 4.2.1. It can be easily shown that $||\Sigma_{11}^{-1}\Sigma_{12}||_1 = 1$, thus the Irrepresentable Condition fails. On the other hand, we have

$$\begin{aligned} \min_{\Omega_)}||\Sigma_{11}\boldsymbol{\tau}_1 + \Sigma_{12}\boldsymbol{\tau}_2||_\infty &\geq \min_{\boldsymbol{\tau}\in\Omega_0}||\Sigma_{11}^{-1}(\boldsymbol{\tau}_1 - \Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{\tau}_2)||_\infty \\ &\geq \min_{\boldsymbol{\tau}\in\Omega_0}\frac{1}{\sqrt{s}}\lambda_{min}(\Sigma_{11})||\boldsymbol{\tau}_1 - \Sigma_{11}^{-1}\Sigma_{12}\boldsymbol{\tau}_2||_\infty \\ &\geq \frac{\rho}{s}, \end{aligned} \tag{4.37}$$

i.e., the *LCC* condition holds with $u_0(n,p,s) = \frac{\rho}{s}$.

Proof of Theorem 5.3.1 Suppose, we are on $\mathcal{J}_0$, then we have $||\hat{\boldsymbol{\beta}}_{init} - \beta^\circ||_2^1 \leq C_0\sqrt{\frac{s\log p}{n}}$. As we stated before, it suffices for us to show the support recovery, i.e., $S_n = S$ with large probability, as it implies the strong oracle property by the KKT conditions. First, we show $S_n \subseteq S$ with large probability. It suffices to show that

$$\exists j \in S^c, \hat{w}_j||\tilde{\Sigma}||_\infty < \lambda \text{ with large probability.} \tag{4.38}$$

We note that,

$$\begin{aligned}
\Pr(\exists j \in S^c, \hat{w}_j||\tilde{\Sigma}||_\infty < \lambda) &\leq \sum_{j\in S^c} Pr(\hat{w}_j||\tilde{\Sigma}||_\infty \leq \lambda) \\
&\leq \sum_{j\in S^c} Pr(|\tilde{\beta}_j| \geq (\frac{\lambda}{||\tilde{\Sigma}||_\infty})^{1/\gamma}) \\
&\leq p \times Pr(||\tilde{\Sigma}||_\infty \geq \frac{\lambda}{||\tilde{\beta}-\beta||_2^\gamma}) \\
&\leq p \times Pr(||\tilde{\Sigma} - B_k(\Sigma)||_{max} \geq \frac{1}{2k_n}\frac{\lambda}{(C_0\sqrt{s\log p/n})^\gamma}) \\
&+ p \times Pr(||B_k(\Sigma)||_{max} \geq \frac{1}{2k_n}\frac{\lambda}{(C_0\sqrt{s\log p/n})^\gamma}) \\
&= (R1) + (R2)
\end{aligned} \tag{4.39}$$

We consider the probability bound for terms $(R1)$ and $(R2)$ in the above inequality separately. For $(R2)$,

$$\begin{aligned}
(R2) &\leq Pr(||\Sigma||_{max} \geq \frac{1}{2k_n}\frac{\lambda}{(C_0\sqrt{s\log p/n})^\gamma}) \\
&\leq Pr(||\Sigma||_2 \geq \frac{1}{2k_n}\frac{\lambda}{(s\log p/n)^{\gamma/2}}) \\
&= 0
\end{aligned} \tag{4.40}$$

since $\lambda > \frac{C_0^\gamma}{\epsilon_0}(2k_n+1)(s\log p/n)^{\gamma/2}$.

Next, we bound $(R1)$. By Lemma 6.1, for $\frac{1}{2k_n+1}\frac{\lambda}{(C_0\sqrt{s\log p/n})^\gamma} \geq \sqrt{2(\eta+1)\frac{1}{\gamma(\epsilon_0,\delta)}}\sqrt{\frac{\log p}{n}}$,

we have

$$(R1) \leq 3p^{-\eta}. \tag{4.41}$$

This indicates that $S_n \supseteq S$. The second step is based on the first step, where we shall use a proof by contradiction to show that $S_n \subseteq S$ with the additional assumption $\lambda \geq (u_0 - 3C(\frac{\log p}{n})^{\frac{\alpha}{2(\alpha+1)}})(\eta - C\sqrt{s \log p/n})$. Define $\mathcal{J}_1 = \{||\tilde{\Sigma} - B_{k_n}(\Sigma)||_\infty \leq t^0\}$, with t^0 defined in Lemma 6.1 ; $\eta = \min_{j \in S} \beta_j$ and $\boldsymbol{\tau}_1$. Then it suffices to show that

$$||\boldsymbol{\tau}_S||_\infty < 1, \text{ with large probability.} \tag{4.42}$$

If not, then we would have $\boldsymbol{\tau} \in \Omega_0$.

Combining the KKT conditions and the LCC condition gives us that, conditional on the event $\mathcal{J}_0 \cap \mathcal{J}_1$,

$$\begin{aligned} \frac{\lambda}{\hat{\eta}^\gamma} &\geq ||\Sigma_{k_n,11}\boldsymbol{\tau}_1 + \Sigma_{k_n,12}\boldsymbol{\tau}_2||_\infty - ||\tilde{\Sigma}_{k_n,11} - \Sigma_{k_n,11}||_\infty - ||\tilde{\Sigma}_{k_n,12} - \Sigma_{k_n,12}||_\infty \\ &\geq u_0(n,p,s) - 2C_1 k_n^\alpha - 2||\tilde{\Sigma}_{k_n} - \Sigma_{k_n}||_\infty. \end{aligned} \tag{4.43}$$

Further define $\mathcal{J} = \mathcal{J}_0 \cap \mathcal{J}_1$. In the event $\mathcal{J}$,

$$\hat{\eta}^\gamma \leq (\frac{\lambda}{u_0(n,p,s) - 2C_1 k_n^\alpha - 2t^0}). \tag{4.44}$$

On the other hand,

$$\begin{aligned} \hat{\eta} &\geq \eta - ||\hat{\boldsymbol{\beta}}_{init} - \boldsymbol{\beta}^\circ||_\infty \\ &\geq \eta - C_0\sqrt{s_0 \log p/n}. \end{aligned} \tag{4.45}$$

Combining (4.44) and (4.45) together leads to $\lambda \geq (u_0(n,p,s) - 2C_1 k_n^\alpha - 2t^0)(\eta - C_0\sqrt{s_0 \log p/n})^\gamma$, which is a contradiction.

In conclusion, with probability at least $\Pr(\mathcal{J}) \geq 1 - \delta_{n,p,s} - 3p^{-\eta}$, we have

$$S_n = S, \tag{4.46}$$

and further we have

$$\hat{\beta}_{\text{HTR}} = \tilde{\beta}^{\circ}. \tag{4.47}$$

Lemma 4.5.1 *For all $t \geq t^0$, we have*

$$P(\mathcal{J}_2) \geq 1 - 3(p \vee n)^{-\eta}. \tag{4.48}$$

Proof First, it follows from Lemma A.3 of Bickel and Levina (2008b) that

$$\begin{aligned} P(||\tilde{\Sigma} - B_{k_n}(\Sigma)||_{\infty} \geq t) &\leq (2k_n + 1)p \exp\{-n(t^0)^2 \gamma(\varepsilon_0, \delta)\} \\ &\leq (2k_n + 1)(p \vee n) \exp\{-2n(\eta + 1)\frac{1}{\gamma(\varepsilon_0, \delta)} \frac{\log(p \vee n)}{n} \gamma(\varepsilon_0, \delta))\} \\ &\leq 3((p \vee n)k_n) \exp\{-(\eta + 1) \log((p \vee n)k_n))\} \\ &\leq 3((p \vee n)k_n)^{-(\eta+1)+1} \leq 3(p \vee n)^{-\eta}, \end{aligned}$$

where $t^0 = \sqrt{2(\eta + 1)\frac{1}{\gamma(\epsilon_0, \delta)}}\sqrt{\frac{\log(p \vee n)}{n}}$.

CHAPTER5: SPARSE MULTICATEGORY DISCRIMINANT ANALYSIS

Many supervised machine learning tasks can be cast as multi-class classification problems. Linear discriminant analysis has been well studied in two class classification problems and can be easily extended to multi-class cases. For high dimensional classification, traditional linear discriminant analysis fails due to diverging spectra and accumulation of noise. Therefore, researchers have proposed penalized LDA (Fan et al. 2012, Witten and Tibshirani 2011). However, most available methods for high dimensional multi-class LDA are based on an iterative algorithm, which is computationally expensive and not theoretically justified. In this paper, we present a new framework for sparse multicategory discriminant analysis (SMDA) for high dimensional multi-class classification by simultaneous extracting the discriminant directions. Our SMDA can be cast as an convex programming which distinguishes itself from other state-of-the-art method. We evaluate the performances of the resulting methods on the extensive simulation study and a real data analysis.

5.1 Fisher's Linear Discriminant Analysis

Suppose the random variables representing two classes $\mathcal{C}_1$ and $\mathcal{C}_2$ follow p-variate normal distributions $\mathbf{X}|Y = 1 \sim N(\boldsymbol{\mu}_1, \Sigma)$ and $\mathbf{X}|Y = 2 \sim N(\boldsymbol{\mu}_2, \Sigma)$ respectively. For any linear discriminant rule

$$\delta_{\mathbf{w}}(\mathbf{X}) = 1\{\mathbf{w}^T(\mathbf{X} - \boldsymbol{\mu}_a) > 0\}, \tag{5.1}$$

where $\boldsymbol{\mu}_a = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ and 1 denotes the indicator function. Define $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, then the misclassification rate of the classifier $\delta_{\mathbf{x}}$ is

$$W(\delta_{\mathbf{w}}) = 1 - \Phi\{\mathbf{w}^T\boldsymbol{\mu}_d/(\mathbf{w}^T\Sigma\mathbf{w})^{\frac{1}{2}}\}. \tag{5.2}$$

The mission is to find a good data projection direction $\mathbf{w}$. Note that the Fisher discriminant

$$\delta_F(\mathbf{X}) = 1\{(\Sigma^{-1}\boldsymbol{\mu}_d)^T(\mathbf{X} - \boldsymbol{\mu}_a) > 0\} \tag{5.3}$$

corresponds to the Bayes rule, which minimizes the misclassification error, or equivalently, solves the following constrained optimization problem,

$$\underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbf{w}^T\Sigma\mathbf{w} \text{ s.t. } \mathbf{w}^T\boldsymbol{\mu}_d = 1. \tag{5.4}$$

An extension of Fisher's LDA is possible by considering above formulation. Suppose there are K classes and, for $j = 1, \ldots, K$, the jth class has mean $\boldsymbol{\mu}_j$ and common covariance structure Σ. Fisher's reduce rank approach to multi-class classification problem is to find $r \leq K - 1$ discriminant directions $(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$ such that separate the population centroid the most in the projected space $\mathcal{S} = \text{span}(\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_r)$. Then the population centroids and new observation $\mathbf{X}$ are both projected to $\mathcal{S}$. The observation $\mathbf{X}$ will be assigned to the class whose projected centroid is closest to the projection of $\mathbf{X}$ on $\mathcal{S}$. It is not necessary to compute all $K - 1$ discriminant directions (DDs) whose span is that of all K population centroids; the process can stop as long as the projected population centroids are well spread out in $\mathcal{S}$. In light of this procedure, Fisher's LDA sequentially solve

$$\underset{\mathbf{w}_k}{\operatorname{argmax}} \frac{\mathbf{w}_k^T B\mathbf{w}_k}{\mathbf{w}_k^T\Sigma\mathbf{w}_k} \text{ s.t. } \mathbf{w}_k^T\Sigma\mathbf{w}_j = 0, \forall j < k, \tag{5.5}$$

where $B = \mathbf{U}^T\mathbf{U} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_a, \ldots, \boldsymbol{\mu}_K - \boldsymbol{\mu}_a)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_a, \ldots, \boldsymbol{\mu}_K - \boldsymbol{\mu}_a)^T$ is the between class covariance and $\boldsymbol{\mu}_a = \frac{\sum_{i=1}^K \boldsymbol{\mu}_i}{K}$. The resulting solution, $\mathbf{w}_k$, is referred as the k-th discriminant direction.

In order to solve (5.5), we can be reformulate it as the following constraint form,

$$\underset{\mathbf{w}_k}{\operatorname{argmax}}\, \mathbf{w}_k^T\mathbf{B}\mathbf{w}_k \text{ s.t. } \mathbf{w}_k^T\Sigma\mathbf{w}_k \leq 1 \text{ and } \mathbf{w}_k^T\Sigma\mathbf{w}_j = 0, \forall j < k. \tag{5.6}$$

5.2 Sparse Multicategory Discriminant Analysis

In high dimensions, there are several reasons that why Fisher's linear discriminant rule does not lead to a suitable classifier in high dimensions,

1 $\hat{\Sigma}$ is singular and fails to converge to Σ in high dimensions;

2 The sample population centroids are contaminated by the noise accumulation effect when p is large;

3 The classifier results non-interpretable discriminant by using all features.

Some work has been done to modify Fisher's linear discriminant rule to appreciate for high dimensional issues. Duintjer Tebbens and Schlesinger (2007) required the solution does not lie in the null space of $\mathbf{B}$. Others have proposed to modify problem (5.5) by using a positive definite estimate of Σ, see Friedman (1989), Dudoit et al. (2002), Bickel and Levina (2004) among many others. More recently, Fan et al. (2012) has proposed the regularized optimal affine discriminant (ROAD) method; Cai and Liu (2011c) proposed a direct estimation approach for sparse linear discriminant analysis. However, their method focuses on binary classification problem and extension to multi-class problem is unavailable. Witten and Tibshirani (2011) reformulate the problem (5.5) as

$$\mathbf{w}_k^{\circ} = \underset{\mathbf{w}_k}{\operatorname{argmax}}\, \mathbf{w}_k^T B \mathbf{w}_k \text{ subject to } \mathbf{w}_k^T \Sigma \mathbf{w}_k \leq 1, \tag{5.7}$$

where $B^k = \frac{1}{n}\mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1/2}P_k^{\perp}(\mathbf{Y}^T\mathbf{Y})^{-1/2}\mathbf{Y}^T\mathbf{X}$ with $P_k^{\perp}$ defined as an orthogonal projection matrix into the space that is orthogonal to $(\mathbf{Y}^T\mathbf{Y})^{-1/2}\mathbf{Y}^T\mathbf{X}\hat{\mathbf{w}}_i$ for all $i < k$.

Then they propose the k-th penalized discriminant direction $\hat{\mathbf{w}}_k$ to be the solution to

$$\underset{\mathbf{w}_k}{\operatorname{argmax}}\{\mathbf{w}_k^T B^k \mathbf{w}_k - P_k(\mathbf{w}_k)\} \text{ subject to } \mathbf{w}_k^T \Sigma \mathbf{w}_k \leq 1. \tag{5.8}$$

However, in general, problem (5.8) is not a convex programming even if lasso penalty is used, i.e. $P_k(\mathbf{w}_k) = \lambda_k ||\mathbf{w}_k||_1$, because it involves maximizing an objective function that is

not concave. Thus solving (5.8) sequentially is computationally intractable. Moreover, the k-th discriminant estimator depends on all the previous discriminant estimator, and thus the estimation error could be accumulated and enlarged.

In this paper, we propose a unified framework for linear discriminant analysis, which has a very close connection with the ROAD classifier and Witten's penalized LDA framework. We shall make use of Theorem 5.2.2, which provides a reformulation of criterion (5.5). We start with a representation of the between correlation matrix $\mathbf{B}$ in the following proposition

Proposition 5.2.1 *We can decompose* $\mathbf{B}$ *as* $\Psi^T\Psi$*, such that* $\Psi^T = (\boldsymbol{\mu}_1 + \frac{1}{\sqrt{K}-1}(\boldsymbol{\mu}_K - \sqrt{K}\boldsymbol{\mu}_a), \ldots, \boldsymbol{\mu}_{K-1} + \frac{1}{\sqrt{K}-1}(\boldsymbol{\mu}_K - \sqrt{K}\boldsymbol{\mu}_a))$.

The above proposition gives a full rank representation of $\mathbf{B}$ such that $\mathbf{B} = \Psi^T\Psi$, where Ψ is a full rank matrix. We exploit such a representation in our procedure by providing the following reformulation of criterion (5.5).

Theorem 5.2.2 *The solution* $\mathbf{W}^\circ = (\mathbf{w}_1^\circ, \ldots, \mathbf{w}_{K-1}^\circ)$ *to problem (5.5) also solves*

$$\text{argmin}\{\frac{1}{2}Tr(\mathbf{W}^T\Sigma\mathbf{W}) - Tr(\mathbf{L}^T\mathbf{W})\}, \tag{5.9}$$

where $\mathbf{L}^T = \mathbf{P}^T\Psi$ *and* $\mathbf{P}$ *is the eigen-matrix of* $\Psi\Sigma^{-1}\Psi^T$*, i.e.* $\Psi\Sigma^{-1}\Psi^T = \mathbf{P}\Lambda\mathbf{P}^T$ *with* Λ *the diagonal matrix.*

Given the above theorem, we are ready to propose the unified framework, sparse multi-category discriminant analysis (SMDA). We would like to add a penalty function for capacity control. As our primary interest is classification error control (risk control), Lasso penalty is added for regularization. We define the sparse discriminant directions (SDDs) to be the solution to

$$\mathbf{W}^\circ = \text{argmin}\{\frac{1}{2}\text{Tr}(\mathbf{W}^T\Sigma\mathbf{W}) - \text{Tr}(\mathbf{L}^T\mathbf{W}) + \sum_{k=1}^{K-1}\lambda_k||\mathbf{w}_k||_1\}. \tag{5.10}$$

Corollary 5.2.3 ***(Binary Case)*** *In binary classification setting, problem (5.10) reduces to*

$$\mathbf{w}^\circ = \text{argmin}\{\frac{1}{2}\mathbf{w}^T\Sigma\mathbf{w} - \mathbf{w}^T\boldsymbol{\ell} + \lambda||\mathbf{w}||_1\}, \tag{5.11}$$

where $\boldsymbol{\ell} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

The above optimization procedure was first proposed by Sun et al. (2014) and can be obtained by first considering a reformulation of criterion (5.4) in binary classification setting as

$$\mathbf{w}_0 = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} - \mathbf{w}^T\boldsymbol{\ell}, \tag{5.12}$$

then transfer (5.12) to its corresponding penalized version

$$\mathbf{w}_{0,\lambda} = \underset{\mathbf{w}}{\operatorname{argmin}}\, f(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} - \mathbf{w}^T\boldsymbol{\ell} + \lambda||\mathbf{w}||_1. \tag{5.13}$$

Therefore, by Corollary 5.2.3, we indeed have a unified procedure by including their result as special case. Moreover, we discuss some connections between formulation (5.11) and the optimization problem considered in Fan et al. (2012) for performing high dimensional binary classification . However, rather than recasting the problem as in (5.11), they formulate it as

$$\mathbf{w}_c = \underset{||\mathbf{w}||_1 \le c, \mathbf{w}^T\boldsymbol{\ell}=1}{\operatorname{argmin}} \mathbf{w}^T\Sigma_R\mathbf{w},$$

which can further be reformulated as

$$\mathbf{w}_\lambda = \underset{\mathbf{w}^T\boldsymbol{\ell}=1}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} + \lambda||\mathbf{w}||_1. \tag{5.14}$$

Since (5.14) involves a linear equality constraint, they replace it by a quadratic penalty as

$$\mathbf{w}_{\lambda,\gamma} = \operatorname{argmin} \frac{1}{2}\mathbf{w}^T\Sigma_R\mathbf{w} + \lambda||\mathbf{w}||_1 + \frac{1}{2}\gamma(\mathbf{w}^T\boldsymbol{\ell} - 1)^2. \tag{5.15}$$

Their formulation requires the simultaneously tuning of λ and γ, which can be computationally intensive. However, in Fan et al. (2012), they stated that the solution to (5.15) is not sensitive to γ, since solution is always in the direction of $\Sigma_R^{-1}\boldsymbol{\ell}$ when $\lambda = 0$, as validated

by simulations. Their formulation (5.14) is close to the formulation (5.13). This result sheds some light on why $\mathbf{w}_{\lambda,\gamma}$ is not sensitive to γ. The estimation procedure (5.11) also enjoys other nice properties, for example, the solution path of (5.11) enjoys the piecewise linear property. We refer reader to Sun et al. (2014) for more details.

5.2.1 A Vector-Wise Coordinate Descent Algorithm

We develop a fast computational algorithm to problem (5.33) by using the co-ordinate descent. What makes the co-ordinate descent algorithm particularly attractive for problem (5.33) is that there is an closed form formula for each or-ordinate. We write $\mathbf{W} = (\mathbf{w}^1, \ldots, \mathbf{w}^p)$, where $\mathbf{w}^i$ is the i-th column of $\mathbf{W}$. Without of generality, suppose that $\tilde{\mathbf{w}}^j$ for all $j \geq 2$ are given, and we need to optimize (5.33) with respect to $\mathbf{w}^1$. In this case, the objective function (5.33) becomes

$$g(\mathbf{w}_1) = \frac{1}{2}\mathrm{Tr}\left((\mathbf{w}^1, \widetilde{\mathbf{W}}^{2:p})\Sigma(\mathbf{w}^1, \widetilde{\mathbf{W}}^{2:p})^T\right) - \mathrm{Tr}\left(\mathbf{L}(\mathbf{w}^1, \widetilde{\mathbf{W}}^{2:p})^T\right) + ||\boldsymbol{\lambda} \odot \mathbf{w}^1||_1 + \sum_{k=2}^{K-1} ||\boldsymbol{\lambda} \odot \tilde{\mathbf{w}}^k||_1, \tag{5.16}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{K-1})^T$ and $\odot$ denotes the hadamard product, i.e. element-wise product. We take the derivative of $g(\mathbf{w}_1)$ over $\mathbf{w}_1$,

$$\begin{aligned} g'(\mathbf{w}_1) &= \frac{1}{2}\left\{\sum_{k\neq 1}\sigma_{1k}\tilde{\mathbf{w}}^k + \sum_{k\neq 1}\sigma_{1k}\tilde{\mathbf{w}}^k\right\} + \sigma_{11}\mathbf{w}^1 - \ell^1 + \mathrm{diag}(\boldsymbol{\lambda})\Gamma \\ &= \sum_{k\neq 1}\sigma_{1k}\tilde{\mathbf{w}}^k + \sigma_{11}\mathbf{w}^1 - \ell^1 + \mathrm{diag}(\boldsymbol{\lambda})\Gamma. \end{aligned} \tag{5.17}$$

By simple calculation, we can construct the co-ordinate update as

$$\mathbf{w}^1 = \frac{S(\ell^1 - \sum_{k\neq 1}\sigma_{1k}\tilde{\mathbf{w}}^k, \boldsymbol{\mu})}{\sigma_{11}}, \tag{5.18}$$

where $S(\cdot,\cdot)$ is the vector-wise soft thresholding operator, that is,

$$S(\mathbf{x}, \boldsymbol{\lambda}) = \mathrm{sign}(\mathbf{x})\max(|\mathbf{x}| - \boldsymbol{\lambda}, \mathbf{0}). \tag{5.19}$$

Based on this result, we can obtain a coordinate descent algorithm as follows

Algorithm

(a) Initialize $\mathbf{W}$ at a starting point $\mathbf{W}_{(0)}$ and set $m = 0$.

(b) Repeat:

- (b.1) Increase m by 1: $m \leftarrow m + 1$
- (b.2) for $j \in 1, \cdots, p$, if $\tilde{\mathbf{w}}^j_{(m-1)} = \mathbf{0}$, then set $\tilde{\mathbf{w}}^j_{(m)} = \mathbf{0}$;
  otherwise: $\tilde{\mathbf{w}}^j_{(m)} = \operatorname{argmin} g(\tilde{\mathbf{w}}^1_{(m)}, \cdots, \tilde{\mathbf{w}}^{j-1}_{(m)}, \mathbf{w}^j, \tilde{\mathbf{w}}^{j+1}_{(m-1)}, \cdots, \tilde{\mathbf{w}}^p_{(m-1)})$.

(c) Until numerical convergence: we require $||\mathbf{W}_{(m)} - \mathbf{W}_{(m-1)}||$ to be sufficiently small.

5.2.2 Implementation of SMDA

Let $\widetilde{\Sigma}$ and $\widehat{\mathbf{L}}$ be, respectively, estimators of Σ and $\mathbf{L}$. Here we use $\tilde{\Sigma}$ to denote any positive covariance estimator other than sample covariance matrix $\hat{\Sigma}$. Then the sample version of the problem reduces to

$$\hat{\mathbf{W}} = \operatorname{argmin}\{\frac{1}{2}\operatorname{Tr}(\mathbf{W}^T\widetilde{\Sigma}\mathbf{W}) - \operatorname{Tr}(\widehat{\mathbf{L}}\mathbf{W}) + \sum_{k=1}^{K-1}\lambda_k||\mathbf{w}_k||_1\}. \tag{5.20}$$

Let $\widetilde{\Sigma}$ be a regularized covariance estimator of Σ, which will be discussed in detail in section 5.2.3. Further more, we take $\widehat{\Psi}^T = (\hat{\boldsymbol{\mu}}_1 + \frac{1}{\sqrt{K}-1}(\hat{\boldsymbol{\mu}}_K - \sqrt{K}\hat{\boldsymbol{\mu}}_a), \ldots, \hat{\boldsymbol{\mu}}_{K-1} + \frac{1}{\sqrt{K}-1}(\hat{\boldsymbol{\mu}}_K - \sqrt{K}\hat{\boldsymbol{\mu}}_a))$ be an estimator of Ψ^T, where $\hat{\boldsymbol{\mu}}_k = \sum_{j\in\mathcal{C}_k}\frac{\mathbf{x}_j}{n_k}$, $\mathcal{C}_k$ is the index set of k-th class and $n_k = \sharp\{\mathcal{C}_k\}$, the cardinality of $\mathcal{C}_k$ for $1 \le k \le K-1$. We further decompose $\widehat{\Psi}\widetilde{\Sigma}^{-1}\widehat{\Psi}^T$ as $\widehat{\mathbf{P}}\widehat{\Lambda}\widehat{\mathbf{P}}^T$, i.e., $(\widehat{\Lambda}, \widehat{\mathbf{P}})$ is the eigen-pair of $\widehat{\Psi}\widetilde{\Sigma}^{-1}\widehat{\Psi}^T$, we then take $\widehat{\mathbf{L}}$ as $\widehat{\mathbf{P}}^T\widehat{\Psi}$.

We remind the reader that, as Fisher's linear discriminant analysis, it is usually not necessary to compute all $K-1$ discriminant directions, the procedure can stop as long as the data is well separated in the projection space. Moreover, we point out the connection between the discriminant directions and a eigen problem. By Theorem 5.2.2, we know $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{K-1})$, where $\lambda_1 \ge \ldots \ge \lambda_{K-1}$ is the eigen-value matrix to $\Sigma^{-1}\mathbf{B}$,

which is composed of objective values corresponding to the discriminant rules of (5.5). Note that the eigen-vetors that corresponds to the smaller eigenvalues will tend to be very sensitive to the exact choice of training data and express high variability in risk estimation. Based on this observation, we propose a "eigen-cut" procedure to achieve a reduced rank projection, which is implemented as following:

- Compute $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{K-1})$;
- Calculate the sample cdf of λ_i's: $\hat{F}_n(\lambda_i) = \sum_{k=1}^{i} \hat{\lambda}_k / \sum_{k=1}^{K-1} \hat{\lambda}_k$, for $1 \leq i \leq K-1$;
- Cue the discriminant directions corresponding to the eigenvalue which satisfies: $\hat{F}_n(\lambda_i) < \alpha_{cut}$, for $1 \leq i \leq K-1$.

The above procedure only preserves the discriminant directions that account for $1-\alpha$ of the separation and ignores the ones that contribute little. Such procedure is also commonly used in factor analysis. We recommend to use $\alpha_{cut} = 0.10$ based on the our simulations, which works for most of the settings.

5.2.3 Estimation of Covariance Matrices

Our procedure requires to estimate a positive covariance estimator and this section is contributed to this issue. We discuss four commonly assumed covariance structures of $\mathbf{X}$ and provide corresponding estimators.

Shrunken Covariance Matrices

Friedman (1989) first proposes the regularized discriminant analysis (RDA) by shrinking the sample covariance matrix to an identity matrix such that the variance of associated with the sample based estimate at the expense of potentially increased bias. Friedman considers to shrink the covariance estimator to the identity matrix, i.e.

$$\tilde{\Sigma}_\gamma = (1-\gamma)\hat{\Sigma}_n + \gamma I, \tag{5.21}$$

where γ is the regularization parameter to control the shrinkage toward an identity matrix, I, which can be chosen by cross validation. Further, Ledoit and Wolf (2004) shows that the above estimator is consistent when p/n is bounded, while, at the same time, enjoys very good computational property. We explore this estimator in both simulations and real data analysis. Moreover, we would like to point reader to Zou and Trevor (2005), where the elastic net estimator for a linear model can be recast as the solution to

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\{\boldsymbol{\beta}^T n\tilde{\Sigma}_n\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \lambda||\boldsymbol{\beta}||_1\}, \tag{5.22}$$

in which $\tilde{\Sigma}_\gamma$ is defined as in (5.21).

Sparse Precision Matrices

Precision matrix estimation is strongly connected to the estimation of graphical models. To be more specific, for Gaussian distributions, recovering the structure of the graph G is equivalent to estimating the support of the precision matrix (Lauritzen 1996). In this setting, it is natrual to assume a sparse graph structure and thus a sparse precision matrix. Cai et al. (2011) proposes the constrained l1-minimization for inverse matrix estimation (CLIME) which enjoys very attractive computational efficiency for high dimensional data and is adopted by our method. Specifically, the CLIME estimator $\hat{\Omega} = (\hat{\omega}_{ij})$ is defined as $\hat{\Omega} = (\hat{\omega}_{ij})$ with $\hat{\omega}_{ij} = \hat{\omega}_{ji} = \hat{\omega}^1_{ij} I\{|\hat{\omega}^1_{ij}| \leq \hat{\omega}^1_{ji}| + \hat{\omega}^1_{ji} I\{|\hat{\omega}^1_{ji}| \leq \hat{\omega}^1_{ij}|\}$, where $\hat{\Omega}^1$ is estimated by

$$\min ||\Omega||_1 \text{ subject to:} \tag{5.23}$$

$$|\hat{\Sigma}\Omega - I|_\infty \leq \gamma, \quad \Omega \in \mathbb{R}^{p\times p}. \tag{5.24}$$

In the end, $\tilde{\Sigma}$ is taken as $\hat{\Omega}^{-1}$.

Bandable Covariance Matrix

Motivated by applications such as climatology and spectroscopy, where there is a natural metric on the index set and $|i-j|$ large implies near independence or conditional independence of X_i and X_j, Bickel and Levina (2008b) proposes to regularize large covariance matrix through banding or tapering over a well behaved class of matrices: the bandable class of covariance matrices, i.e.

$$\begin{aligned}\mathcal{U}(\varepsilon_0, \alpha, C) &= \{\Sigma = (\sigma_{jj'}) : \max_j \sum_{j'} \{|\sigma_{j'j}| : |j'-j| > k\} \leq Ck^{-\alpha} \text{ for all } k > 0 \\ &\quad \text{and } 0 < 1/M \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M\}. \end{aligned} \tag{5.25}$$

Cai and Liu (2011b) further relaxes the above assumption by only requiring that the eigenvalue of Σ is bounded from above, that is, $\lambda_{\max}(\Sigma) \leq M$. They then propose a new tapering estimator and show estimators by tapering the maximum likelihood estimator achieves minimax risk spectral norm rate $\min\{n^{-2\alpha/(2\alpha+1)+\frac{\log p}{n}}, \frac{p}{n}\}$. Specifically, the tapering estimator is defined as $\tilde{\Sigma} = (w_{ij}\hat{\sigma}_{ij})$, where $\hat{\sigma}_{ij}$ is the (i,j) element of sample covariance matrix $\hat{\Sigma}_n$ and w_{ij} has the following form

$$w_{ij} = \begin{cases} 1, & \text{when} \quad |i-j| \leq k/2 \\ 2 - \frac{2|i-j|}{k}, & \text{when} \quad k/2 < |i-j| < k \\ 0, & \text{otherwise.} \end{cases} \tag{5.26}$$

Sparse Covariance Matrices

In many applications, there is no natural order on the features space like we assumed for bandable covariance matrices. In this setting, permutation-invariant estimators are favored and general sparsity assumption is usually imposed on the whole covariance matrix, i.e. most of entries in each row/column of covariance matrix are zero or negligible. We apply a hard thresholding procedure proposed in Bickel and Levina (2008a) under this setting. The thresholding estimator $\widetilde{\Sigma} = (\widetilde{\sigma}_{ij})$ is given by $\widetilde{\sigma}_{ij} = \widehat{\sigma}_{ij} I(|\widehat{\sigma}_{ij}| \geq \gamma\sqrt{\frac{\log p}{n}})$ for some tuning

parameter γ, which can be chosen by cross validation.

5.2.4 Tuning Parameter Selection

Regularization Parameter λ

In this section, we consider to select tuning parameters λ_k for the SMDA problem (5.33). The simplest approach would be to take $\lambda_k = \lambda$, i.e. the same tuning parameter value for all components. However, as mentioned in Witten and Tibshirani (2011), this results in effectively penalizing each discriminant direction more than previous discriminant since the loss function corresponding to the k-th discriminant direction is equal to k-th largest eigenvalue of $\Psi\Sigma^{-1}\Psi^T$, denoted as $\lambda_k(\Psi\Sigma^{-1}\Psi^T)$. Thus, instead of having k distinctive tuning parameters, we set

$$\lambda_k = \lambda \times \lambda_k(\Psi\Sigma^{-1}\Psi^T), \tag{5.27}$$

where λ is a single tuning parameter. By reducing the number of tuning parameter from k to 1, we significantly reduce the computation burden of this procedure. Finally, the tuning parameter can be chosen by cross validation.

Regularization Parameter in Covariance Estimation

In this section, we talk about the choice of regularization parameter in covariance matrix estimation. Inspired by Bickel and Levina (2008b), we propose a cross validation methods to choose the regularization parameter in a very general setting. We take the shrunken covariance estimator Σ_γ as an example and state that our method can be generalized to all other settings. We propose to select the tuning parameter γ by minimizing the risk

$$R(\gamma) = E||\hat{\Sigma}_\gamma - \Sigma||_{(1,1)} \tag{5.28}$$

with the oracle γ given by

$$\gamma^\circ = \underset{\gamma}{\operatorname{argmin}}\, R(\gamma). \tag{5.29}$$

We choose the l_1 to l_1 matrix norm over than other matrix norms mainly for computational issues. We shall note that the selection of γ is not sensitive to the choice to norm. We next propose a N-fold cross validation scheme to estimate the risk and thus γ°: randomly partition the original sample into N equal size subsamples, choose a single subsample as the validation sample and the remaining $N-1$ subsamples as the construction data. We use the sample covariance matrix of the validation data as the target to choose the best γ for the construction sample. The cross validation is repeated N times and denote $\hat{\Sigma}_\gamma^{c,k}$, $\hat{\Sigma}^{v,k}$ as the shrunken covariance matrix estimator of construction data and the sample covariance matrix of the validation data from the v-th split respectively. Then the risk (5.28) can be estimated by

$$\hat{R}(\gamma) = \frac{1}{N}\sum_{k=1}^{N} ||\hat{\Sigma}_\gamma^{c,k} - \hat{\Sigma}^{v,k}||_{(1,1)} \tag{5.30}$$

and γ is selected as

$$\hat{\gamma} = \underset{k}{\operatorname{argmin}}\, \hat{R}(\gamma). \tag{5.31}$$

Generally we found little sensitivity to the choice of N and use 3-fold cross validation through out this paper.

5.2.5 Covariance Structure Selection

In this section, we talk about how we can adapt to the covariance structure of Σ. We propose a simple criterion based on prediction performance of the classifier. Write f the prediction accuracy and $\hat{f}$ the estimated prediction accuracy. Suppose we have a finite set of covariance structure, denoted as $\Theta =$ to choose from: shrunken covariance matrices, sparse precision matrices, bandable covariance matrices and sparse covariance matrices in this paper, then we propose to maximize the following criterion for covariance structure selection

$$\underset{k\in\Theta}{\operatorname{argmax}}\{f_k(\lambda,\gamma)\}, \tag{5.32}$$

where λ and γ are regularization parameters in optimization and covariance matrix estimation respectively. For a proper estimation of f_k's, subsampling method can be used as discussed in last subsection.

5.3 Theoretical Investigation

In this section, we investigate the theoretical property of SMDA and its associated estimator. By substituting $\widetilde{\Sigma}$ and $\hat{\mathbf{L}}$, we can calculate an estimate of $\mathbf{W}^\circ$ as

$$\hat{\mathbf{W}}_\lambda = \text{argmin}\{\frac{1}{2}\text{Tr}(\mathbf{W}^T\widetilde{\Sigma}\mathbf{W}) - \text{Tr}(\hat{\mathbf{L}}^T\mathbf{W}) + \sum_{k=1}^{K-1}\lambda_k||\mathbf{w}_k||_1\}. \tag{5.33}$$

For two vetors $\mathbf{a}, \mathbf{b}$, a natural way to measure the discrepancy of their directions is the L^2 norm distance as L^2 convergence indicates the direction consistence. For two set of vectors, $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$ and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_K\}$, we consider the the following loss function,

$$||A - B||_{2,\infty} \doteq \max_k ||\mathbf{a}_k - \mathbf{b}_k||_2. \tag{5.34}$$

We focus on the scenario that there are a few nonzero components in $\mathbf{w}_{j,0}$, the j-th column of $\mathbf{W}^\circ$, that is, a few response variables are associated with the covariates of interest in each projection direction. Such a scenario is common in many large-scale problems. Let $S_j = \{i : w_{i,j,0} \neq 0\}$ be the active set of $\mathbf{w}_{j,0} = (w_{1,j,0}, \cdots, w_{p,j,0})^T$ and s_j is the number of elements in S_j. Further we make the following assumptions.

Assumptions

1 We assume there exist constants (m, M) and $(\sigma_{\min}, \sigma_{\max})$, such that

$$0 < m \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M < \infty \text{ and}$$

$$0 < \sigma_{\min} \leq \inf_{j\in\{1,\ldots,p\}} \sigma_j \leq \sup_{j\in\{1,\ldots,p\}} \sigma_j \leq \sigma_{\max} < \infty.$$

2 Assume that $0 < p_0 \leq \inf \frac{n_j}{n} \leq \sup \frac{n_j}{n} \leq p_1 < 1$.

3 We require the estimators $\widetilde{\Sigma}$ and $\widetilde{\Psi\Omega\Psi^T}$ is consistent in the sense that

$$||\widetilde{\Sigma} - \Sigma||_2^1 = a_n \tag{5.35}$$

with probability at least $1 - \delta_{1,n,p}$; and,

$$||\widetilde{\Psi\Omega\Psi^T} - \Psi\Omega\Psi^T||_2^1 = b_n \tag{5.36}$$

with probability at least $1 - \delta_{2,n,p}$. Moreover we assume that $\Psi\Omega\Psi^T$ has distinctive eigenvalues.

Now we are ready to present the main result.

Theorem 5.3.1 *Assume that assumptions 1-3 holds,and*

$$\lambda = \max\{a_n||\mathbf{W}^\circ||_{2,\infty}, t_1^0\} \asymp a_n||\mathbf{W}^\circ||_{2,\infty} \vee b_n||\Psi||_{\infty,2} \vee \log p/n \tag{5.37}$$

where $t_1^0 := C_b b_n||\Psi||_{\infty,2} \vee \frac{\eta_0}{c_K}\log p/n$*, in which,* c_K *and* C_b *does not depend on* n, p, s_0*. Then with probability at least* $1 - (K-1)p^{-\eta_0} - \delta_{1,n,p} - \delta_{2,n,p}$*, we have*

$$||\hat{\mathbf{W}}_\lambda - \mathbf{W}^\circ||_{2,\infty} \doteq \max_j\{||\hat{\mathbf{w}}_{j,\lambda} - \mathbf{w}_{j,0}||_2\} \leq C\sqrt{s}\lambda, \tag{5.38}$$

where C *is constant not depending* n *and* p*.*

Theorem 5.3.1 gives an oracle inequality and the column-wise L_2 convergence rate of $\hat{\mathbf{W}}_\lambda$ in the sparse case, which indicates column-wise direction consistency. This result has several important implications. If $\sqrt{s_0}\lambda = o(1)$, then $||\hat{\mathbf{W}}_\lambda - \mathbf{W}^\circ||_2$ converges to zero in probability. Therefore, our SMDA should perform well for the sparse cases with $s_0 << n$. This is extremely important in practice, since the extremely sparse cases are common for many large-scale problems.

5.4 Simulation Studies

In this section, we present extensive simulation study and compare our SMDA method with several other methods including independence rule (Naive Bayes), RDA (Friedman 1989) and Penalized LDA (Witten and Tibshirani 2011) , denoted as PLDA, and DSDA in binary classification setting. We denote SMDA with bandable covariance matrix estimator, shrunken covariance matrix estimator, sparse covariance matrix estimator and sparse precision matrix as SMDA-Ba, SMDA-Sh, SMDA-SC, SMDA-SP respectively. We also consider the SMDA framework by using the sample covariance matrix $\hat{\Sigma}_n$, which is singular when $p > n$. To remedy this issue, we add a small constant η (e.g. $\eta = 10^{-6}$) to all diagonal entries of the matrix $\hat{\Sigma}_n$ and denote such estimator SMDA-Sa. In all simulations studies, we consider the number of features $p = 1,000$ and the sample size of the training and testing data is $n = 100$ for each class.

Setting 1 (Sparse Strong Signal and Dense Weak Signal With Independent Features) In the first setting, we consider two classes, $\mathcal{C}_1$ and $\mathcal{C}_2$. We take $\mathbf{x}_i^k (\in \mathcal{C}_k) \sim N(\mu_k, \Sigma), \forall k = 1, 2,$ and $1 \leq i \leq n$,where, $\Sigma = (\sigma_{ij})$ is taken such that $\sigma_{ij} = 1$ for $i = j$; $\sigma_{ij} = 0$ for $i \neq j$. Further more, we consider three different cases. In case 1, we consider a sparse strong signal setting and set $\boldsymbol{\mu}_1 = (\mathbf{1}_{10}, \mathbf{0}_{p-10}), \boldsymbol{\mu}_2 = \mathbf{0}$ by introducing a mean shift $\mathbf{1}_{10}$. In case 2 and 3, we consider a relatively dense but weak signal setting and set $\boldsymbol{\mu}_1 = (\mathbf{1}_{200}/\sqrt{10}, \mathbf{0}_{p-200}), \boldsymbol{\mu}_2 = \mathbf{0}$ and $\boldsymbol{\mu}_1 = (\mathbf{1}_{500}/\sqrt{30}, \mathbf{0}_{p-500}), \boldsymbol{\mu}_2 = \mathbf{0}$ respectively.

Corresponding result is presented in Table 5.1. Regularizing the covariance matrix helps improve the classification accuracy, especially when the signal becomes relatively denser and weaker. Moreover, when the signal is sparse and strong, regularizing the covariance matrix helps reduce the variance of classification error and number of selected features.

Setting 2 (Sparse Signal With Power Decay Correlation) There are three classes: $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ with $\mathbf{x}_i^k (\in \mathcal{C}_k) \sim N(\mu_k, \Sigma), \forall k = 1, 2, 3,$ and $1 \leq i \leq n$. In this setting, we introduce a mean shift such that $\boldsymbol{\mu}_1 = (\mathbf{1}_5, -\mathbf{1}_5, \mathbf{0}_{p-10}), \boldsymbol{\mu}_2 = (-\mathbf{1}_5, \mathbf{1}_5, \mathbf{0}_{p-10})$ and $\boldsymbol{\mu}_3 = (\mathbf{1}_1 0, \mathbf{0}_{p-10})$. $\Sigma = (\sigma_{ij})$ is taken such that $\sigma_{ij} = 1$ for $i = j$; $\sigma_{ij} = \rho^{|i-j|}$ for $i \neq j$, with ρ varies from 0 to

0.9.

Corresponding result is presented in Table 5.2 and 5.3. We see that SMDA-Ba has the overall fine performance over other methods especially when ρ is large (the bandable structure is strong), which is not surprising as it coincides with our intuition. SMDA-Sh performs rather well when ρ is small ($\rho \leq 0.6$).

Setting 3 (Sparse Signal With Equal Correlation) There are three classes: $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$ with $\mathbf{x}_i^k (\in \mathcal{C}_k) \sim N(\mu_k, \Sigma), \forall k = 1, 2, 3,$ and $1 \leq i \leq n$. In this setting, we introduce a mean shift such that $\boldsymbol{\mu}_1 = (\mathbf{1}_5, -\mathbf{1}_5, \mathbf{0}_{p-10}), \boldsymbol{\mu}_2 = (-\mathbf{1}_5, \mathbf{1}_5, \mathbf{0}_{p-10})$ and $\boldsymbol{\mu}_3 = (\mathbf{1}_1 0, \mathbf{0}_{p-10})$. $\Sigma = (\sigma_{ij})$ is taken such that $\sigma_{ij} = 1$ for $i = j$; $\sigma_{ij} = \rho$ for $i \neq j$, with ρ varies from 0 to 0.9.

Result is summarized in Table 5.4 and 5.5. We see that SMDA-Sh and SMDA-Sa has the overall fine performance over other methods, but SMDA-Sh helps with the variable selection and its variance.

Setting 4 (Sparse Signal With Block Diagonal Correlation) In this example, we follow the set up as in the above example, except that the covariance matrix is taken to be block diagonal. There are 5 blocks with each of dimension 200×200. We further consider two separate cases: in the first case, we take each block as a power decay correlation matrix, i.e., the (i, j) element in each block is taken to be $(\rho^{|i-j|}$; in the second case, we take each block as a equal correlated matrix with pairwise correlation ρ, or in other words, $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = 1$, if $i = j$; ρ, otherwise.

Result with the block diagonal setting where each block is taken to be a $AR(1)$ covariance matrix is summarized in Table 5.6 and 5.7. It shares similar pattern with setting 2. We omit the result with the block diagonal setting where each block is taken to be a equal correlation matrix as it shares similar spirit with setting 3.

Setting 5 (Sparse Signal With Sparse Correlation Matrix) In this example, we follow the basic set up as in setting 2, while the correlation matrix is taken to be very sparse. We consider the $AR(1)$ population correlation model, $\Sigma = [\sigma_{ij}] = [\rho^{|i-j|}]$ with $\rho = 0.5$. The

value of 0.5 was chosen so that the matrix is very sparse. The resulting matrix is then permuted at random to have sparse correlation but not bandable structure.

Setting 6 (Sparse Signal With Sparse Precision Matrix) In this example, we follow the basic set up in setting 5, instead of sparse correlation, we consider a sparse precision matrix (inverse of correlation matrix) Ω as the underlying mechanism. We take the precision matrix as the diagonal block matrix with block size 5 where each block has off-diagonal entries equal to 0.5 and diagonal 1. Finally the resulting matrix is then randomly permuted.

Results from above two settings are summarized in Table 5.8. In setting 5, SMDA-SC out performs all the other methods as the population covariance matrix has sparse covariance structure. In setting 6, SMDA-Sh and SMDA-SC performs the best and then followed by SMDA-SP. The reason is that that corresponding covariance matrix also has a sparse covariance matrix structure. In linear discriminant analysis, as pointed by Friedman (1989), the prediction accuracy can often be improved by replacing $\hat{\Sigma}_n$ by a shrunken estimate. Likewise, Zou and Trevor (2005) conjectured that whenever ridge regression improves on OLS (ordinary least square), elastic net will improve lasso by incorporating the shrunken estimation of covariance matrix. We view our result as a partial validation and a generalization of Friedman and Zou's conjecture. However, we show by simulation that the type of regularization is better to adapt to the structure of covariances matrices, shrunken estimate does not always give the best performance in all cases. For example, when the correlation between covariates has a bandable structure, which is appropriate for applications such as climatology, spectroscopy and GWA studies, bandable estimation of covariance matrices has the best performance in terms of risk estimation.

5.5 An Application To Cancer Research Study

In cancer research study, a reliable and precise classification of tumors is essential for successful treatment of cancer. cDNA microarrays and high-density oligonucleotide chips have allowed us monitoring the expression levels for thousands of genes simultaneously

Table 5.1: Setting 1: independent features setting. We report the Median Testing Classification Error (MTE) in percentage, the Median of number of nonzero coefficients (denoted as s) and their standard deviations (in parentheses).

	SMDA-Sa	SMDA-Ba	SMDA-Sh	SMDA-SC	SMDA-SP	DSDA	PLDA	NB
MTE	7.5 (2.29)	**6.5(2.20)**	6.8(1.80)	**6.5(1.83)**	**6.5(1.91)**	7.5(2.27)	21.8(3.11)	24.0(3.60)
s	15(24.2)	13(21.6)	12(15.5)	13(16.9)	12(18.1)	12 (10.2)	1000 (0)	1000(0)
MTE	10.5(2.59)	5.5(2.10)	**5.0(1.55)**	5.3(1.58)	5.5(1.58)	15.5(3.31)	5.5(1.62)	10.0(1.90)
s	398(89.7)	695(248.7)	571(209.6)	671(205.9)	565(196.4)	87(14.2)	1000 (0)	1000(0)
MTE	17.5(3.06)	**9.0(2.30)**	**9.0(2.32)**	**9.0(2.38)**	**9.0(2.14)**	29.0(3.92)	**9.0(1.90)**	14.0(2.81)
s	446(96.0)	970.5(109.2)	903(116.3)	969(117.1)	884(116.6)	97 (18.6)	1000 (0)	1000(0)

Table 5.2: Setting 2: Sparse Signal with Power Decay Correlation. We report the Median Testing Classification Error in percentage and its standard deviations (in parentheses).

rho	SMDA-Sa	SMDA-Ba	SMDA-Sh	SMDA-SC	SMDA-SP	PLDA	NB	RDA
0	3.00(1.04)	2.33(1.13)	**2.00(0.89)**	2.33(0.93)	2.67(2.03)	7.33(1.68)	13.00(2.13)	66.67(0)
1	4.33(1.19)	3.67(2.48)	**2.67(0.93)**	3.33(1.07)	3.33(2.08)	9.17(1.97)	14.00(2.28)	66.67(0)
2	6.00(1.37)	5.00(1.58)	**4.33(1.33)**	**4.33(1.42)**	5.17(2.55)	10.33(1.74)	15.17(2.3)	66.67(0)
3	7.67(1.41)	6.33(1.57)	**5.33(1.39)**	6.00(1.63)	6.33(2.82)	11.83(1.79)	16.67(2.11)	66.67(0)
4	8.67(1.72)	8.00(1.90)	**7.33(1.57)**	**7.33(1.74)**	8.00(2.12)	14.17(2.00)	18.33(2.22)	66.67(0)
5	10.33(1.76)	8.67(1.72)	**8.33(1.80)**	**8.33(1.50)**	9.00(3.00)	15.00(2.97)	20.67(2.36)	66.67(0)
6	11.67(2.04)	**9.83(1.83)**	10.67(1.73)	10.00(1.78)	11.00(4.23)	18.67(2.74)	24.00(2.67)	66.67(0)
7	12.00(1.90)	**10.17(1.91)**	12.17(1.41)	11.00(2.47)	12.33(2.93)	21.67(3.49)	26.33(3.04)	66.67(0)
8	11.00(2.31)	**9.33(1.91)**	12.33(1.89)	10.67(2.11)	12.67(2.96)	32.00(3.09)	31.00(3.26)	66.67(0)
9	7.33(2.41)	**6.50(2.61)**	9.67(2.84)	9.00(4.83)	10.00(5.10)	32.50(1.79)	39.00(3.51)	66.67(0)

Table 5.3: Setting 2:Sparse Signal with Power Decay Correlation. We report the Median of number of nonzero coefficients and its standard deviations (in parentheses).

ρ	SMDA-Sa		SMDA-Ba		SMDA-Sh		SMDA-SC		SMDA-SP	
	s1	s2	s1	s2	s1	s2	s1	s2	s1	s2
0	10 (8.5)	13(21.2)	10(2.5)	10(3.6)	10(1.5)	10(16.9)	10(2.7)	10(6.3)	10(0.8)	14(1.9)
1	10(10.7)	13(23.2)	10(2.4)	9(3.2)	10(2.1)	10(15.3)	10(2.0)	10(4.4)	10(0.6)	17(1.9)
2	10(6.3)	12(17.3)	10(2.1)	9(2.7)	10(2.2)	10(18.7)	10(1.6)	10(4.1)	10(0.6)	12(2.2)
3	10(2.5)	12(11.2)	10(2.3)	9(3.2)	10(1.7)	10(17.0)	10(1.6)	10(3.8)	10(0.9)	15(2.5)
4	10(1.1)	10(11.1)	9(1.8)	8(3.0)	10(1.2)	10(14.8)	10(2.6)	10(5.9)	10(0.7)	14(2.0)
5	9(3.8)	10(15.6)	9(2.0)	7(3.5)	10(3.7)	10(13.4)	10(1.1)	10(3.1)	10(0.8)	14(2.3)
6	8(2.3)	9(16.3)	8(1.7)	7(3.2)	10(1.1)	10(15.3)	10(1.2)	8(3.8)	10(1.2)	13(2.5)
7	8(2.8)	7(16.3)	8(1.3)	6(2.3)	10(0.8)	10(11.7)	10(1.4)	7(3.7)	10(1.3)	11(2.4)
8	6(1.7)	3(12.1)	7(1.4)	5(2.5)	10(1.4)	10(11.5)	9(3.2)	6(8.3)	10(1.9)	10(8.7)
9	6(1.7)	3(12.7)	6(2.3)	4(4.3)	10(0.2)	9(10.2)	8(7.9)	7(18.3)	10(1.9)	10(7.2)

Table 5.4: Setting 3: Sparse Signal With Equal Correlation. We report the Median Testing Classification Error in percentage and its standard deviations (in parentheses).

ρ	SMDA-Sa	SMDA-Ba	SMDA-Sh	SMDA-SC	SMDA-SP	PLDA	NB	RDA
0	3.00(1.10)	2.67(1.05)	2.00(0.95)	2.33(1.03)	3.00(1.10)	7.83(2.04)	13.00(2.07)	66.67(0.00)
1	3.33(1.08)	4.33(1.33)	3.00(1.10)	4.83(2.17)	3.00(0.99)	21.50(9.13)	26.67(9.12)	66.67(0.00)
2	2.83(0.92)	3.50(1.19)	2.67(0.99)	2.67(3.39)	3.33(1.18)	42.50(8.22)	38.33(9.82)	66.67(0.00)
3	2.00(0.75)	2.33(1.11)	2.00(0.84)	2.00(2.09)	3.33(2.32)	45.67(5.50)	46.67(8.72)	66.67(0.00)
4	1.17(0.63)	1.67(1.03)	1.33(0.84)	1.33(2.54)	3.00(4.29)	44.67(10.42)	51.67(8.38)	66.67(0.00)
5	0.67(0.61)	1.00(0.69)	0.67(0.64)	1.33(6.72)	2.67(4.19)	47.50(12.38)	55.33(7.80)	66.67(0.00)
6	0.33(0.34)	0.50(0.44)	0.33(0.50)	0.33(4.58)	5.67(8.36)	55.83(9.76)	57.67(6.77)	66.67(0.00)
7	0.00(0.16)	0.00(0.25)	0.33(0.37)	0.00(5.41)	17.33(17.23)	56.33(8.65)	60.50(6.00)	66.67(0.00)
8	0.00(0.03)	0.00(0.16)	0.00(0.32)	0.00(0.44)	25.50(19.04)	57.17(5.86)	61.67(4.87)	66.67(0.00)
9	0.00(0.00)	0.00(0.00)	0.00(0.23)	0.00(0.00)	41.17(19.34)	55.83(9.42)	62.67(5.31)	66.67(0.00)

Table 5.5: Setting 3: Sparse Signal With Equal Correlation. We report the Median of number of nonzero coefficients and its standard deviations (in parentheses).

ρ	SMDA-Sa		SMDA-Ba		SMDA-Sh		SMDA-SC		SMDA-SP	
	s1	s2	s1	s2	s1	s2	s1	s2	s1	s2
0	10(13.0)	13(21.6)	10(2.7)	10(3.1)	10(1.3)	10(16.2)	10(2.8)	11(5.8)	10(2.6)	13(25.3)
1	13(15.5)	70(24.5)	12(8.1)	43(21.2)	12(17.7)	78(30.6)	53(26.7)	43(30.6)	10(2.7)	11(25.1)
2	18(29.3)	89(31.7)	14(9.2)	53(19.8)	13(14.2)	85(27.1)	16(17.6)	135(68.0)	10(5.0)	12(39.3)
3	22(26.9)	95(30.0)	14(12.2)	52(18.6)	11(21.7)	85(34.1)	11(25.1)	113(97.8)	10(5.0)	10(41.4)
4	22(30.2)	101(32.2)	16(10.8)	59(17.6)	10(9.1)	80(21.7)	16(20.0)	103(93.8)	10(3.4)	11(50.8)
5	26(34.5)	111(36.9)	15(12.6)	61(17.7)	10(3.3)	75(13.6)	45(37.4)	90(104.6)	10(3.3)	11(52.7)
6	48(28.4)	136(33.8)	23(17.1)	64(24.9)	10(2.1)	69(11.8)	85(39.8)	100(66.0)	12(6.2)	16(52.1)
7	41(11.9)	133(16.5)	29(19.4)	66(23.9)	10(1.7)	53(9.6)	73(35.9)	117(64.1)	12(11.1)	20(98.6)
8	16(3.9)	108(7.8)	58(30.8)	103(36.4)	10(1.5)	42(7.3)	17(16.3)	110(24.1)	9(14.5)	17(135.0)
9	11(2.7)	97(7.7)	66(13.9)	96(13.8)	10(0.8)	31(9.7)	10(2.6)	106(23.2)	6(24.6)	15(92.8)

Table 5.6: Setting 4: Sparse Signal With Block Diagonal Correlation. We report the Median Testing Classification Error in percentage and its standard deviations (in parentheses).

ρ	SMDA-Sa	SMDA-Ba	SMDA-Sh	SMDA-SC	SMDA-SP	PLDA	NB	RDA
0	3.00(0.99)	2.33(1.08)	2.00(0.75)	2.33(0.80)	2.33(0.87)	8.50(1.80	13.33(2.24)	66.67(0.00)
1	4.33(1.11)	3.67(1.32)	3.00(1.00)	3.00(1.35)	4.00(1.16)	8.17(1.40	14.00(2.07)	66.67(0.00)
2	6.00(1.72)	5.33(2.25)	4.00(1.25)	4.33(1.48)	5.00(1.39)	9.83(1.49	15.33(2.22)	66.67(0.00)
3	7.33(1.44)	6.67(1.77)	5.67(1.42)	5.67(1.61)	6.33(1.45)	10.67(1.68	16.83(2.31)	66.67(0.00)
4	8.67(1.64)	7.33(1.59)	7.00(1.50)	7.00(1.48)	8.00(1.54)	13.00(1.14	18.50(2.45)	66.67(0.00)
5	10.50(1.52)	9.00(1.73)	8.67(1.67)	8.67(1.86)	9.33(1.76)	15.33(2.12	20.83(2.30)	66.67(0.00)
6	12.00(1.95)	10.00(1.93)	10.33(1.76)	10.00(1.74)	11.00(1.88)	16.83(1.57)	24.00(2.40)	66.67(0.00)
7	12.17(2.26)	10.33(2.17)	11.67(1.93)	11.33(2.06)	12.00(1.92)	21.17(2.65)	27.17(2.75)	66.67(0.00)
8	11.00(2.53)	9.67(2.26)	12.33(2.07)	11.00(2.30)	12.33(2.10)	27.67(3.71)	30.67(3.38)	66.67(0.00)
9	7.17(2.23)	7.00(2.47)	9.33(2.18)	9.00(3.06)	10.00(2.23)	33.83(2.58)	38.33(3.66)	66.67(0.00)

Table 5.7: Setting 4: Sparse Signal With Block Diagonal Correlation. We report the Median of number of nonzero coefficients and its standard deviations (in parentheses).

ρ	SMDA-Sa		SMDA-Ba		SMDA-Sh		SMDA-SC		SMDA-SP	
	s1	s2	s1	s2	s1	s2	s1	s2	s1	s2
0	10(3.0)	15(17.1)	10(3.7)	10(5.0)	10(0.8)	10(13.6)	10(1.8)	11(4.8)	10(2.8)	13(21.9)
1	10(8.2)	13(23.7)	10(2.1)	9(2.5)	10(1.1)	11(13.6)	10(1.8)	10(4.4)	10(4.5)	18(27.0)
2	10(10.5)	13(22.7)	10(2.4)	9(3.0)	10(0.3)	10(6.8)	10(1.3)	10(3.6)	10(2.3)	17(25.4)
3	10(2.1)	11(11.9)	9(2.5)	9(4.1)	10(3.8)	11(23.7)	10(2.0)	10(4.1)	10(4.1)	18(28.7)
4	10(4.1)	11(13.4)	9(1.7)	8(3.2)	10(0.7)	10(11.4)	10(1.4)	10(4.7)	10(4.8)	15(27.8)
5	9(1.8)	10(12.1)	8(1.6)	7(2.1)	10(1.4)	10(16.8)	10(0.9)	9(3.4)	10(3.8)	18(27.7)
6	8(1.4)	8(8.3)	8(1.9)	7(2.7)	10(0.4)	10(12.3)	10(1.1)	8(3.8)	10(3.6)	12(26.5)
7	7(1.6)	7(13.4)	7(1.5)	6(2.4)	10(1.6)	10(14.4)	10(0.9)	7(2.5)	10(3.9)	10(26.1)
8	6(2.0)	5(13.5)	7(1.3)	6(2.3)	10(0.7)	10(11.8)	9(1.3)	7(4.7)	10(1.4)	10(14.7)
9	5(1.4)	2(12.4)	6(1.5)	4(2.4)	10(0.3)	9(7.9)	8(2.5)	6(7.9)	10(0.6)	10(6.2)

Table 5.8: Setting 5& 6: Sparse Signal with Sparse Correlation or Sparse Precision Matrix. We report the Median Testing Classification Error (MTCE) in percentage, the Median of number of nonzero coefficients in both projection directions (denoted as s_1 and s_2 respectively) and their standard deviations (in parentheses) in both of Sparse Correlation (SC) setting and Sparse Precision (SP) setting.

		SMDA-Sa	SMDA-Ba	SMDA-Sh	SMDA-SC	SMDA-SP	PLDA	NB	RDA
SC	MTE	2.67(0.91)	2.33(1.14)	**1.67(0.75)**	2.00(0.79)	3.00(1.03)	9.83(1.55)	16.33(2.31)	66.67(0)
	s1	18(29.5)	10(3.3)	10(0.7)	10(2.2)	10(1.6)	1000(0)	1000(0)	1000(0)
	s2	55(46.6)	10(5.0)	10(13.6)	12(7.0)	11(20.6)	1000(0)	1000(0)	1000(0)
SP	MTE	2.67(1.05)	2.33(1.17)	**2.00(0.87)**	**2.00(0.76)**	2.33(0.94)	8.00(1.92)	13.67(2.17)	66.67(0)
	s1	10(17.8)	10(2.6)	10(1.0)	10(1.5)	10(3.3)	1000(0)	1000(0)	1000(0)
	s2	13(28.7)	10(3.8)	11(13.1)	11(4.6)	12(25.8)	1000(0)	1000(0)	1000(0)

and lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more informative classification. The ability to successfully distinguish between tumor classes (already known or yet to be discovered) using gene expression data is an important task for better diagnostics and treatment.

We examine the performance of Pen-MDA on human breast tumor microarray datasets publicly available at `https://genome.unc.edu/cgi-bin/SMD/publication/viewPublication.pl?pub_no=107&23820` and described in Harrell et al. (2011). The data consisted of a combined microarray data set of four studies taken from the public domain. We utilized the microarray as presented in the following breast cancer datasets: GSE2034, GSE12276, GSE2603 and the NKI295. The clinical data from these patients was obtained from previous studies (Bos et al. 2009; Zhang et al. 2009). NC60 cell line microarray data was obtained from `http://genome-www.stanford.edu/nci60/`.Additional microarrays from the GEO for the MDA-MB-231 cells were downloaded from GSE12237 and GSE2603. Probes in these external sets were assigned to Entrez Gene identifiers and replicate gene names were collapsed to the median. The data from the four tumor datasets were then combined using Distance Weighted Discrimination (Benito et al. 2004) to remove the systematic biases present in different microarray datasets. In all datasets, samples were standardized to zero mean and unit variances before other analyses were performed. The samples have 4 subtypes: luminal A/B, Her2-enriced, basal-like and normal-like. We note that the normal-like subtype is much rarer We delete the normal-like subtype since they are much rarer than others.

We are interested in the classification of tumors based on the gene expression profiles. We further focus on 1000 genes with the largest variances among the $18,975$ covariates. We randomly partition the data 100 times, each with a training set of size $2/3$ of the original sample and a test set of $1/3$ of the observations. We report the training error, testing error and number of genes selected for SMDA, NB, RDA and penalized LDA, and summarize the result in Table 5.9. We shall notice that SMDA-Sh and SMDA-Ba has similar performance and out performs all other comparative methods in terms of classification error. All of PLDA, NB and RDA make use of 1000 genes and yet produce worse performance, especially, RDA

performs the worst and is similar to random guessing. Overall, the SMDA by incorporating covariance structure is a validated classification technique.

Table 5.9: Real data analysis: We report Median Test Classification Error (MTE) and Median of number of nonzero coefficients.

Methods	MTE	s1	s2	s3
SMDA-Sa	0.2458	52	52	73
SMDA-Sh	**0.1620**	46	49	52
SMDA-Ba	**0.1654**	145	189	226
SMDA-SC	0.1844	190	190	190
SMDA-SP	0.2430	74	322	499
PLDA	0.1899	1000	1000	1000
NB	0.1899	1000	1000	1000
RDA	0.6983	1000	1000	1000

5.6 Conclusions and Discussions

In this paper, we introduce a unified framework, Penalized Multiple Discriminant Analysis (SMDA), for linear discriminant analysis in high dimensional multi-class classification setting. Our SMDA has very close connection with the ROAD methodology proposed by Fan et al. (2012) when considered in binary classification setting. We also proposed to incorporate the regularization covariance estimator into the classification setting to improve the risk estimation by trade-off between noise accumulation and correlation modeling, and we demonstrate its effectiveness in various simulation settings and a real data example. Further, we propose a simple method to choose the covariance structure based on the classification error. Both theory and numerical examples have shown the superiority of our SMDA framework over other methods.

5.7 Appendix

Proof of Proposition 5.2.1 Denote $\tilde{\boldsymbol{\mu}}_a = \frac{K\boldsymbol{\mu}_a - \boldsymbol{\mu}_K}{K-1}$, then we have

$$\begin{aligned}\Psi^T &= (\boldsymbol{\mu}_1 + \frac{1}{\sqrt{K}-1}(\boldsymbol{\mu}_K - \sqrt{K}\boldsymbol{\mu}_a), \ldots, \boldsymbol{\mu}_{K-1} + \frac{1}{\sqrt{K}-1}(\boldsymbol{\mu}_K - \sqrt{K}\boldsymbol{\mu}_a)) \\ &= (\boldsymbol{\mu}_1 - \tilde{\boldsymbol{\mu}}_a + \frac{\sqrt{K}}{K-1}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_a), \ldots, \boldsymbol{\mu}_{K-1} - \tilde{\boldsymbol{\mu}}_a + \frac{\sqrt{K}}{K-1}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_a)).\end{aligned} \tag{5.39}$$

Write $\mathbf{a} = \frac{\sqrt{K}}{K-1}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_a)$, then we end up with

$$\begin{aligned}\Psi^T\Psi &= \sum_{k=1}^{K-1}(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_a + \mathbf{a})(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_a + \mathbf{a})^T \\ &= \sum_{k=1}^{K-1}(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_a)(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_a)^T + \frac{K}{K-1}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_a)(\boldsymbol{\mu}_K - \boldsymbol{\mu}_a)^T \\ &= \sum_{k=1}^{K-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_a)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_a)^T - 2\sum_{k=1}^{K-1}\frac{1}{K-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_a)(\boldsymbol{\mu}_a - \boldsymbol{\mu}_K)^T \\ &+ \frac{1}{K-1}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_a)(\boldsymbol{\mu}_K - \boldsymbol{\mu}_a)^T + \frac{K}{K-1}(\boldsymbol{\mu}_K - \boldsymbol{\mu}_a)(\boldsymbol{\mu}_K - \boldsymbol{\mu}_a)^T \\ &= \sum_{k=1}^{K}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_a)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_a)^T \\ &= \mathbf{B}.\end{aligned} \tag{5.40}$$

Proof completed.

In order to arrive at the main theorem, we need the following lemmas.

Lemma 5.7.1 *If* $(\lambda, \mathbf{v})$ *is an eigen-pair of* $\Sigma^{-1/2}\mathbf{B}\Sigma^{-1/2}$ *then* $(\lambda, \Sigma^{-1/2})$ *is the eigen-pair of* $\Sigma^{-1}\mathbf{B}$*; vice, versa.*

Proof Suppose $(\lambda, \mathbf{v})$ is an eigen-pair of $\Sigma^{-1/2}\mathbf{B}\Sigma^{-1/2}$, then we have

$$\Sigma^{-1/2}\mathbf{B}\Sigma^{-1/2}\mathbf{v} = \lambda\mathbf{v}, \tag{5.41}$$

multiplying $\Sigma^{-1/2}$ on both side, we conclude that $(\lambda, \Sigma^{-1/2}\mathbf{v})$ is the eigen-pair of $\Sigma^{-1}\mathbf{B}$; vice, versa.

Since $\Sigma^{-1/2}\mathbf{B}\Sigma^{-1/2}$ and $\Sigma^{-1}\mathbf{B}$ have the same rank, we conclude that the eigen-pair of $\Sigma^{-1/2}\mathbf{B}\Sigma^{-1/2}$ and $\Sigma^{-1}\mathbf{B}$ has one-to-one correspondence with the same eigenvalues.

Lemma 5.7.2 *The eigen-pairs of* $\Sigma^{-1}\mathbf{B}$ *solves the Fisher's linear discriminant rule (5.5).*

Proof Before we state the proof, we start by some notation and definitions. Suppose $M_{q\times q}$ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q$. Denote $\lambda_j(M)$ as the j-th large eigen value of M, i.e. λ_j.

First, we observe that problem (5.5) can reduce to the following problem,

$$\mathbf{v}_1^\circ = \text{argmax}\, \frac{\mathbf{v}_1^T\mathbf{B}\mathbf{v}_1}{\mathbf{v}_1^T\mathbf{v}_1} \text{ and} \tag{5.42}$$

$$\mathbf{v}_k^\circ = \text{argmax}\, \frac{\mathbf{v}_k^T\mathbf{B}\mathbf{v}_k}{\mathbf{v}_k^T\mathbf{v}_k} \text{ s.t. } \mathbf{v}_k \perp \mathbf{v}_j = 0, \forall 1 \leq j < k, 1 \leq k \leq K-1, \tag{5.43}$$

$$\mathbf{w}_k^\circ = \Sigma^{-1/2}\mathbf{v}_k^\circ, \forall 1 \leq k \leq K-1, \tag{5.44}$$

where $(\lambda_k, \mathbf{w}_k^\circ), 1 \leq k \leq K-1$ are the solutions to problem (5.5). Then

$$\mathbf{w}_k^\circ = \Sigma^{-1/2}\mathbf{v}_k \text{ where } (\lambda_k(\Sigma^{-1/2}\mathbf{B}\Sigma^{-1/2}), \mathbf{v}_k) \text{ is the eigen-pair of } \Sigma^{-1/2}\mathbf{B}\Sigma^{-1/2}, \forall 1 \leq k \leq K-1. \tag{5.45}$$

Thus applying Lemma 5.7.1, we conclude the result.

Proof of **Theorem 5.2.2** Write

$$Q(\mathbf{W}) = \frac{1}{2}\text{Tr}(\mathbf{W}^T\Sigma\mathbf{W}) - \text{Tr}(\mathbf{L}\mathbf{W}), \tag{5.46}$$

where, $\mathbf{L} = \mathbf{P}^T\Psi$.

Denote the minimizer of $Q(\mathbf{W})$ as $\mathbf{W}^\circ$, then $\mathbf{W}^\circ$ satisfies

$$\partial Q(\mathbf{W}^\circ) = \Sigma\mathbf{W}^\circ - \mathbf{L}^T = \mathbf{0}. \tag{5.47}$$

Thus, we conclude $\mathbf{W}^\circ = \Sigma^{-1}\Psi^T\mathbf{P}$. In order to show that the problem 5.9 gives the solution to Fisher's linear discriminant rule (5.5), by Lemma 5.7.2, we only need to show that the

solution is composed of eigen-vectors of $\Sigma^{-1}\mathbf{B}$. More formally, we need to verify that $\mathbf{W}^\circ$ is indeed the eigen-matrix of $\Sigma^{-1}\mathbf{B}$, i.e. each column of $\mathbf{W}^\circ$ is a eigen-vector of $\Sigma^{-1}\mathbf{B}$. Note that we have the following equality holds

$$\Psi\Sigma^{-1}\Psi^T = \mathbf{P}\Lambda\mathbf{P}^T. \tag{5.48}$$

Left multiply $\Sigma^{-1}\Psi^T$ and right multiply $\mathbf{P}^T$ on both side of 5.48, we end up with

$$\Sigma^{-1}\Psi^T\Psi\Sigma^{-1}\Psi^T\mathbf{P} = \Sigma^{-1}\Psi^T\mathbf{P}\Lambda, \tag{5.49}$$

i.e.

$$\Sigma^{-1}\mathbf{B}\mathbf{W}^\circ = \mathbf{W}^\circ\Lambda. \tag{5.50}$$

Proof completed.

Lemma 5.7.3 *We have the following basic inequality*

$$\begin{aligned}
&\frac{1}{2}Tr\{(\widehat{\mathbf{W}}_\lambda - \mathbf{W}_0)^T\widetilde{\Sigma}(\widehat{\mathbf{W}}_\lambda - \mathbf{W}_0)\} + \sum_{k=1}^{K-1}\lambda_k||\hat{\mathbf{w}}_{k,\lambda}||_1 \\
&\leq Tr\{(\mathbf{W}_0^T(\Sigma - \widetilde{\Sigma}) + (\hat{\mathbf{L}} - \mathbf{L})^T)(\hat{\mathbf{W}}_\lambda - \mathbf{W}_0)\} + \sum_{k=1}^{K-1}\lambda_k||\mathbf{w}_{k,0}||_1.
\end{aligned} \tag{5.51}$$

Proof We first rewrite the optimization problem as

$$\text{argmin}\,\frac{1}{2}\text{Tr}\{(\widehat{\mathbf{W}}_\lambda - \widetilde{\Sigma}^{-1}\widehat{\mathbf{L}})^T\widetilde{\Sigma}(\widehat{\mathbf{W}}_\lambda - \widetilde{\Sigma}^{-1}\widehat{\mathbf{L}})\} + \sum_{k=1}^{K-1}\lambda_k||\hat{\mathbf{w}}_{k,\lambda}||_1. \tag{5.52}$$

Thus, we have

$$\begin{aligned}
&\frac{1}{2}\text{Tr}\{(\widehat{\mathbf{W}}_\lambda - \widetilde{\mathbf{W}}^{-1}\widehat{\mathbf{L}})^T\widetilde{\Sigma}(\widehat{\mathbf{W}}_\lambda - \widetilde{\mathbf{W}}^{-1}\widehat{\mathbf{L}})\} + \sum_{k=1}^{K-1}\lambda_k||\hat{\mathbf{w}}_{k,\lambda}||_1 \\
&\leq \frac{1}{2}\text{Tr}\{(\mathbf{W}_0 - \widetilde{\Sigma}^{-1}\widehat{\mathbf{L}})^T\widetilde{\Sigma}(\mathbf{W}_0 - \widetilde{\mathbf{W}}^{-1}\widehat{\mathbf{L}})\} + \sum_{k=1}^{K-1}\lambda_k||\mathbf{w}_{k,0}||_1,
\end{aligned} \tag{5.53}$$

which yields,

$$\begin{aligned}
&\frac{1}{2}\mathrm{Tr}\{(\widehat{\mathbf{W}}_\lambda - \mathbf{W}_0)^T\widetilde{\Sigma}(\widehat{\mathbf{W}}_\lambda - \mathbf{W}_0)\} + \sum_{k=1}^{K-1}\lambda_k||\hat{\mathbf{w}}_{k,\lambda}||_1 \\
&\leq \mathrm{Tr}\{(\hat{\mathbf{L}} - \widetilde{\Sigma}\mathbf{W}_0)^T(\widehat{\mathbf{W}}_\lambda - \mathbf{W}_0)\} + \sum_{k=1}^{K-1}\lambda_k||\mathbf{w}_{k,0}||_1 \qquad (5.54)\\
&\leq \mathrm{Tr}\{(\mathbf{W}_0^T(\Sigma - \widetilde{\Sigma}) + (\mathbf{L} - \widehat{\mathbf{L}})^T)(\widehat{\mathbf{W}}_\lambda - \mathbf{W}_0)\} + \sum_{k=1}^{K-1}\lambda_k||\mathbf{w}_{k,0}||_1,
\end{aligned}$$

in which we have used $\widehat{\mathbf{L}} = \Sigma\mathbf{W}_0 + \widehat{\mathbf{L}} - \mathbf{L}_0$ in the last equality.

Lemma 5.7.4 *There exists a constant c_K such that for any $t > C_b||\Psi||_{\infty,2}b_n$, we have*

$$Pr(||\hat{\ell}_j - \ell_j||_\infty > t) \leq 2p\exp\{-c_K n t^2\} \text{ for } 1 \leq j \leq K-1. \qquad (5.55)$$

Proof First, we define $\hat{A} = \widetilde{\Psi\Omega\Psi^T}$ and $A = \Psi\Omega\Psi^T$. By assumptions and matrix perturbation theory, we know that

$$|\lambda_j(\hat{A}) - \lambda_j(A)| = O(c_n) \text{ for } 1 \leq j \leq K-1,$$

and

$$||\hat{\phi}_j - \phi_j|| \leq O(\sum_{i\neq j}\frac{||\Delta A||}{\lambda_j - \lambda_i}) \leq C_b b_n,$$

where C_b is a constant.

Thus, we have

$$\begin{aligned}
\Pr(|\widehat{\Psi}_{\cdot i}^T\hat{\boldsymbol{\phi}}_j - \Psi_{\cdot i}^T\boldsymbol{\phi}_j| > t) &\leq \Pr(|\widehat{\Psi}_{\cdot i}^T\hat{\boldsymbol{\phi}}_j - \Psi_{\cdot i}^T\hat{\boldsymbol{\phi}}_j| > t/3) + \Pr(|\Psi_{\cdot i}^T\hat{\boldsymbol{\phi}}_j - \Psi_{\cdot i}^T\boldsymbol{\phi}_j| > t/3)\\
&\leq \Pr(||\widehat{\Psi}_{\cdot i}^T - \Psi_{\cdot i}^T||_2 > t/3) + \Pr(Cb_n||\Psi_{\cdot i}^T||_2^1 > t/3)\\
&\leq \Pr(\sqrt{K-1}||\widehat{\Psi}_{\cdot i}^T - \Psi_{\cdot i}^T||_\infty > t/3)\\
&\leq 2\exp\{-c_{i,K}nt^2\},
\end{aligned}$$

where

$$
\begin{aligned}
c_{i,K} &= \frac{1}{18\sigma_i^2[\sum_{k\neq j,K}(\frac{\sqrt{K}}{\sqrt{K}-1}\frac{n_k}{n})^2 + (1-\frac{\sqrt{K}}{\sqrt{K}-1}\frac{n_j}{n}) + \frac{(1-\sqrt{K}n_K/n)^2}{(\sqrt{K}-1)^2}]} \\
&\geq \frac{1}{18\sigma_{\max}^2(K-1)[(K-2)(\frac{\sqrt{K}}{\sqrt{K}-1}p_1)^2 + (1-\frac{\sqrt{K}}{\sqrt{K}-1}p_0) + \frac{(1-\sqrt{K}p_0)^2}{(\sqrt{K}-1)^2}]} \\
&\doteq c_K.
\end{aligned}
$$

Thus, by union-sum inequality, we have

$$
\Pr(||\hat{\ell}_j - \ell_j||_\infty > t) \leq 2p\exp\{-c_K n t^2\}. \tag{5.56}
$$

Proof of Theorem 5.3.1

Frist, we define $\mathcal{J}_1 = \cap_{j=1}^{K-1}\{||\hat{\ell}_j - \ell_j|| \leq t_1\}$. Taking $t_1 = t_1^0 = \max\{\eta_0/c_K\sqrt{\frac{\log p}{n}}, C_b||\Psi||_{\infty,2}b_n\}$, by lemma 5.7.4, we have

$$
\Pr(\mathcal{J}_1) \geq 1 - (K-1)p^{-\eta_0}. \tag{5.57}
$$

Further define $\mathcal{J}_2 = \{||\tilde{\Sigma} - \Sigma|| \leq a_n\}$ and $\mathcal{J}_3 = \{||\Psi\tilde{\Sigma}\Psi^T - \Psi\Sigma\Psi^T|| \leq b_n\}$ and let $\mathcal{J}_0 = \cap_{j=1}^3\mathcal{J}_j$. Thus

$$
\Pr(\mathcal{J}_1) \geq 1 - (K-1)p^{-\eta_0} - \delta_{1,n,p} - \delta_{2,n,p}. \tag{5.58}
$$

On the set $\mathcal{J}_0$, by taking $\lambda_j = \max\{a_n||\mathbf{w}_{j,0}||_2, t_1^0\}$ and using the basic inequality, we have

$$
\begin{aligned}
&\frac{1}{2}\mathrm{Tr}\{(\widehat{\mathbf{W}}_\lambda - \mathbf{W}_0)^T\widetilde{\Sigma}(\widehat{\mathbf{W}}_\lambda - \mathbf{W}_0)\} + \sum_{j=1}^{K-1}\lambda_j||\hat{\mathbf{w}}_{j,\lambda}||_1 \\
&\leq \mathrm{Tr}\{(\mathbf{W}_0^T(\Sigma - \widetilde{\Sigma}) + (\mathbf{L} - \widehat{\mathbf{L}})^T)(\widehat{\mathbf{W}}_\lambda - \mathbf{W}_0)\} + \sum_{j=1}^{K-1}\lambda_j||\mathbf{w}_{j,0}||_1 \\
&\leq \sum_{j=1}^{K-1}\{||\Sigma - \widetilde{\Sigma}||_2||\mathbf{w}_{0,j}||_2||\hat{\mathbf{w}}_{j,\lambda} - \mathbf{w}_{j,0}||_2 + ||\hat{\ell}_j - \ell||_\infty||\hat{\mathbf{w}}_{j,\lambda} - \mathbf{w}_{j,0}||_1 + \sum_{j=1}^{K-1}\lambda_j||\mathbf{w}_{j,0}||_1 \\
&\sum_{j=1}^{K-1}\{\lambda_j||\hat{\mathbf{w}}_{j,\lambda} - \mathbf{w}_{j,0}||_1 + \lambda_j||\mathbf{w}_{j,0}||_1\}
\end{aligned} \tag{5.59}
$$

Let $\mathbf{w}_{0,S_0} = [w_{0,j} I(j \in S_0)]$, where $w_{0,j}$ is the $j-$th component of $\mathbf{w}_0$. Then the above equation can be rewritten as

$$\frac{1}{2}\sum_{j=1}^{K-1}\{(\hat{\mathbf{w}}_{j,\lambda} - \mathbf{w}_{j,0})^T(\widetilde{\Sigma} - \Sigma + \Sigma)(\hat{\mathbf{w}}_{j,\lambda} - \mathbf{w}_{j,0}) + \lambda_j||\hat{\mathbf{w}}_{j,\lambda,S_j}||_1 + \lambda_j||\hat{\mathbf{w}}_{j,\lambda,S_0^c}||_1\}$$
$$\leq \sum_{j=1}^{K-1}\{\lambda_j||\hat{\mathbf{w}}_{j,\lambda,S_j} - \mathbf{w}_{j,0,S_j}||_1 + \lambda_j||\mathbf{w}_{j,0,S_j}||_1 + \lambda_j||\hat{\mathbf{w}}_{j,\lambda,S_j^c}||_1\}$$

which yields

$$\sum_{j=1}^{K-1}(m - O(a_n))||\hat{\mathbf{w}}_{j,\lambda} - \mathbf{w}_{j,0}||_2^2 \leq 2\sum_{j=1}^{K-1}\lambda_j\sqrt{s_j}||\hat{\mathbf{w}}_{j,\lambda} - \mathbf{w}_{j,0}||_2.$$

Finally, we obtain the following inequality, by taking $\lambda = \max\{\lambda_1, \ldots, \lambda_{K-1}\}$ and $s_0 = \max\{\lambda_1, \ldots, \lambda_{K-1}\}$

$$||\hat{\mathbf{W}}_\lambda - \mathbf{W}_0||_{2,\infty} \leq \frac{2\lambda\sqrt{s_0}}{\lambda_{min} - O(1)a_n} \leq C\lambda\sqrt{s_0},$$

which finishes the proof.

REFERENCE

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," *2nd International Symposium on Information Theory*, 267–281.

Amos, C. I., Elston, R. C., Bonney, G. E., Keats, B. J. B., and Berenson, G. S. (1990), "A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype," *Am. J. Hum. Genet.*, 47, 247–254.

Amos, C. I. and Laing, A. E. (1993), "A comparison of univariate and multivariate tests for genetic linkage," *Genetic Epidemiology*, 84, 303–310.

Barron, A., Birgé, L., and Massart, P. (1999), "Risk bounds for model selection via penalization," *Probability theory and related fields*, 113, 301–413.

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

— (2008a), "Covariance regularization by thresholding," *The Annals of Statistics*, 2577–2604.

— (2008b), "Regularized estimation of large covariance matrices," *The Annals of Statistics*, 36, 199–227.

Bolstad, B., Irizarry, R., Å strand, M., and Speed, T. (2003), "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, 19, 185–193.

Breiman, L. and Friedman, J. (1997), "Predicting multivariate responses in multiple linear regression," *Journal of the Royal Statistical Society, Series B*, 59, 3–54.

Bühlmann, P. and Geer, S. V. D. (2011), "Statistics for high-dimensional data: methods, theory and applications," .

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

— (2011b), "Adaptive thresholding for sparse covariance matrix estimation," *Journal of the American Statistical Association*, 106, 672–684.

— (2011c), "A direct estimation approach to sparse linear discriminant analysis," *Journal of the American Statistical Association*, 106.

Cai, T., Liu, W., and Luo, X. (2011), "A constrained âĎŞ 1 minimization approach to sparse precision matrix estimation," *Journal of the American Statistical Association*, 106, 594–607.

Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010), "Optimal rates of convergence for covariance matrix estimation," *The Annals of Statistics*, 38, 2118–2144.

Candes, E. and Tao, T. (2007), "The Dantzig selector: Statistical estimation when p is much larger than n," *The Annals of Statistics*, 35, 2313–2351.

Chatterjee, A. and Lahiri, S. (2011), "Bootstrapping lasso estimators," *Journal of the American Statistical Assocation*, 106, Bootstrapping Lasso Estimators.

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

Chiang, A., Beck, J., and al., E. (2006), "Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a BardetâĂŞBiedl syndrome gene (BBS11)," *Proceedings of the National Academy of Sciences*, 103, 6287–6292.

Chiang, M. C., Barysheva, M., Toga, A. W., Medland, S. E., Hansell, N. K., James, M. R., McMahon, K. L., de Zubicaray, G. I., Martin, N. G., Wright, M. J., and Thompson, P. M. (2011), "BDNF gene effects on brain circuitry replicated in 455 twins," *NeuroImage*, 55, 448–454.

Chun, H. and Keles, S. (2010), "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *Journal of Royal Statistical Society, Series B*, 72, 3–25.

Cook, R., Li, B., and Chiaromonte, F. (2010), "Envelope models for parsimonious and efficient multivariate linear regression," *Statist. Sinica*, 20, 927–1010.

Cook, R. D., Helland, I. S. and Su, Z. (2013), "Envelopes and Partial Least Squares Regression," *Journal of Royal Statistical Society, Series B, To appear*, 75, 851–877.

Donoho, D. L. and Johnstone, I. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Pickard, D. (1995), "Wavelet Shrinkage: Asymptopia? (with discussion)," *Journal of the Royal Statistical Society, Ser. B*, 57, 301–369.

Donoho, D. L. et al. (2000), "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math Challenges Lecture*, 1–32.

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

Duintjer Tebbens, J. and Schlesinger, P. (2007), "Improving implementation of linear discriminant analysis for the high dimension/small sample size problem," *Computational Statistics & Data Analysis*, 52, 423–437.

Efron, B. (2010), "Correlated z-values and the accuracy of large-scale statistical estimates," *Journal of the American Statistical Association*, 105, 1042–1055.

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

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

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

— (2006), "Statistical challenges with high dimensionality: Feature selection in knowledge discovery," *arXiv preprint math/0602133*.

Fan, J., Liao, Y., and Mincheva, M. (2013), "Large covariance estimation by thresholding principal orthogonal complements," *Journal of Royal Statistical Society, Series B, To appear*, 75, 603–68.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space (with discussion)," *Journal of Royal Statistical Society, Series B*, 70, 849–911.

— (2011), "Nonconcave Penalized Likelihood With NP-Dimensionality," *Information Theory, IEEE Transactions on*, 57, 5467–5484.

Fan, J. and Peng, H. (2004), "On non-concave penalized likelihood with diverging number of parameters," *Annals of Statistics*, 32, 928–961.

Formisano, E., Martino, F. D., and Valente, G. (2008), "Multivariate analysis of {fMRI} time series: classification and regression of brain responses using machine learning," *Magnetic Resonance Imaging*, 26, 921–934.

Friedman, J. H. (1989), "Regularized discriminant analysis," *Journal of the American statistical association*, 84, 165–175.

Greenshtein, E., Ritov, Y., et al. (2004), "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization," *Bernoulli*, 10, 971–988.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), "The Elements of Statistical Learning," .

Huang, J., Ma, S., and Zhang, C. (2008), "Adaptive Lasso for sparse high-dimensional regression models," *Statistica Sinica*, 18, 1063–1618.

Huo, X. and Ni, X. (2007), "When do stepwise algorithms meet subset selection criteria?" *The Annals of Statistics*, 35, 870–887.

Irizarry, R., Hobbs, B., and Collin, F. (2003), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, 4, 249–264.

James, G. M. and Radchenko, P. (2009), "A generalized Dantzig selector with shrinkage tuning," *Biometrika*, 96, 323–337.

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

Kherif, F., Poline, J. B., Flandin, G., Benali, H., Simon, O., Dehaene, S., and Worsley, K. J. (2002), "Multivariate model specification for fMRI data," *Neuroimage*, 16, 1068–1083.

Kim, Y., Choi, H., and Oh, H.-S. (2008), "Smoothly clipped absolute deviation on high dimensions," *Journal of American Statistical Association*, 103, 1665–1673.

Kim, Y. and Kwon, S. (2012), "Global optimality of nonconvex penalized estimators," *Biometrika*, 99, 315–325.

Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008), "Pleiotropy and Principle Components of Heritability Combine to Increase Power for Association," *Genetic Epidemiology*, 32, 9–19.

Knickmeyer, R. C., Gouttard, S., Kang, C., Evans, D., Wilber, K., Smith, J. K., Hamer, R. M., Lin, W., Gerig, G., and Gilmore, J. H. (2008), "A structural MRI study of human brain development from birth to 2 years," *J Neurosci.*, 28, 12176–12182.

Krishnan, A., Williams, L. J., McIntosh, A. R., and Abdi, H. (2011), "Partial least squares (PLS) methods for neuroimaging: a tutorial and review," *Neuroimage*, 56, 455–475.

Lam, C. and Fan, J. (2009), "Sparsistency and Rates of Convergence in Large Covariance Matrix Estimation." *The Annals of statistics*, 37, 4254–4278.

Lange, K., Hunter, D. R., and Yang, I. (2000), "Optimization transfer using surrogate objective functions," *Journal of Computational and Graphical Statistics*, 9, 1–20.

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

Leek, J. T. and Storey, J. D. (2008), "A general framework for multiple testing dependence," *Proceedings of the National Academy of Sciences*, 105, 18718–18723.

Leng, C., Lin, Y., and Whaba, G. (2004), "A note on the Lasso and related procedures in model selection," *Statistica Sinica*, 16, 1273–1284.

Lenroot, R. K. and Giedd, J. N. (2006), "Brain development in children and adolescents: insights from anatomical magnetic resonance imaging." *Neurosci Biobehav Rev.*, 30, 718–729.

Lin, D., Foster, D. P., and Ungar, L. H. (2008), "A risk ratio comparison of l0 and l1 penalized regressions," *Technical Report*.

Lin, J.-a., Zhu, H., Knickmeyer, R., Styner, M., Gilmore, J., and Ibrahim, J. G. (2012), "Projection Regression Models for Multivariate Imaging Phenotype," *Genetic epidemiology*, 36, 631–641.

Lopes, M., Jacob, L., and Wainwright, M. (2011), "A more powerful two-sample test in high dimensions using random projection," *arXiv preprint arXiv:1108.2401*.

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

Ott, J. and Rabinowitz, D. (1999), "A principle-components approach based on heritability for combining phenotype information," *Hum Heredity*, 49, 106–111.

Paus, T. (2010), "Population neuroscience: Why and how," *Human Brain Mapping*, 31, 891–903.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R., and Wang, P. (2010), "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," *Annals of Applied Statistics*, 4, 53–77.

Peper, J. S., Brouwer, R. M., Boomsma, D. I., Kahn, R. S., and Pol, H. E. H. (2007), "Genetic influences on human brain structure: A review of brain imaging studies in twins," *Human Brain Mapping*, 28, 464–473.

Rosset, S. and Zhu, J. (2007), "Piecewise linear regularized solution paths," *The Annals of Statistics*, 35, 1012–1030.

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

ROWE, D. B. and Hoffmann, R. G. (2006), "Multivariate Statistical Analysis in fMRI." *IEEE Eng Med Biol Med*, 25, 60–64.

Scharinger, C., Rabl, U., Sitte, H. H., and Pezawas, L. (2010), "Imaging genetics of mood disorders," *NeuroImage*, 53, 810–821.

Scheetz, T., Kim, K., and al., E. (2006), "Regulation of gene expression in the mammalian eye and its relevance to eye disease," *Proceedings of the National Academy of Sciences*, 103, 14429–14434.

Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

Sentana, E. (2009), "The econometrics of mean-variance efficiency tests: a survey," *The Econometrics Journal*, 12, C65–C101.

Sun, Q., Zhu, H., Liu, Y., and Ibrahim, J. G. (2014), "SPReM: Sparse Projection Regression Model For High-dimensional Linear Regression," *Journal of the American Statistical Association*, In Press.

Teipel, S. J., Born, C., Ewers, M., Bokde, A. L. W., Reiser, M. F., Moller, H. J., and Hampel, H. (2007), "Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment," *NeuroImage*, 38, 13–24.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288.

van de Geer, S., Bühlmann, P., and Ritov, Y. (2013), "On asymptotically optimal confidence regions and tests for high-dimensional models," *arXiv preprint arXiv:1303.0518*.

Vounou, M., Janousova, E., Wolz, R., Stein, J., Thompson, P., Rueckert, D., Montana, G., and ADNI (2012), "Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease," *Neuroimage*, 60, 700–716.

Wang, B. L., Kim, Y., and Li, R. (2013a), "Calibrating non-convex penalized regression in ultra-high dimension," *The Annals of Statistics*, 41, 2505–2682.

Wang, X., Jiang, Y., Huang, M., and Zhang, H. (2013b), "Robust variable selection with exponential squared loss," *Journal of the American Statistical Association*, 108, 632–643.

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

Xiong, H., Goulding, E. H., Carlson, E. J., Tecott, L. H., McCulloch, C. E., and Sen, Ś. (2011), "A flexible estimating equations approach for mapping function-valued traits," *Genetics*, 189, 305–316.

Yang, K. (2012), "Least Absolute Gradient Selector: Statistical Regression via Pseudo-Hard Thresholding," *arXiv preprint arXiv:1204.2353*.

Yap, J. S., Fan, J., and Wu, R. (2009), "Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci," *Biometrics*, 65, 1068–1077.

Zhang, C. H. (2010), "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, 38, 894–942.

Zhang, C. H. and Huang, J. (2008), "The sparsity and bias of the lasso selection in high-dimensional linear regression," *The Annals of Statistics*, 36, 1567–1594.

Zhang, C.-H. and Zhang, S. (2011), "Confidence intervals for low-dimensional parameters with high-dimensional data," *arXiv preprint arXiv:1110.2563*.

Zhang, C.-H. and Zhang, T. (2012), "A general theory of concave regularization for high dimensional sparse estimation problems," *Statistical Science*, 27, 576–593.

Zhao, P. and Yu, B. (2006), "On model selection consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563.

Zheng, Z., Fan, Y., and Lv, J. (2013), "High dimensional thresholded regression and shrinkage effect," *Journal of the Royal Statistical Society: Series B*, to appear.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Trevor, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society, Series B*, 67, 301–320.