The Pennsylvania State University
The Graduate School

NEW STATISTICAL TOOLS FOR HIGH-DIMENSIONAL DATA MODELING

A Dissertation in
Statistics
by
Wanjun Liu

© 2019 Wanjun Liu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2019

The dissertation of Wanjun Liu was reviewed and approved* by the following:

Runze Li
Eberly Family Chair Professor
Dissertation Advisor, Chair of Committee

Lingzhou Xue
Associate Professor of Statistics

Xingyuan (Ethan) Fang
Assistant Professor of Statistics

Tao Yao
Associate Professor of Industrial and Manufacturing Engineering

Ephraim Hanks
Associate Professor of Statistics
Chair of Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

This dissertation consists of two parts. In the first part, we focus on the estimation of linear functional and its application to projection test for high-dimensional mean vector. We first study a general regularized quadratic programming with non-convex penalty and linear constraint. Deterministic error bounds are established for any stationary point that satisfies the necessary first order condition. We also propose an ADMM algorithm with local linear approximation to solve such the non-convex regularized quadratic programming. In particular, we study a special case of the regularized quadratic programming: estimation of linear functional. Furthermore, we apply the linear functional to perform projection test for high-dimensional data. Two projection tests are proposed. The first one is a projection test based on a data-splitting strategy, which achieves an exact t-test under normality assumption. The second one is a projection test based on an online framework, which updates the estimation of optimal projection direction when new observations arrive. This online projection test improves the power the data splitting approach. We derive the asymptotic normal distributions under both null and alternative for the online projection test. We conduct numerical studies to compare the finite sample performance of our proposed projection tests with several existing tests. The numerical results show that the proposed projection tests can keep the type I error rate well and are much more powerful than other existing tests.

In the second part, we focus on the model free feature screening for high-dimensional data via projection correlation. The idea of feature screening is to deliver a computationally efficient way to reduce the dimensionality of the feature space from a very high scale to a moderate one while retaining all the important features. The proposed method is based on ranking the projection correlations between features and response variable. This screening procedure does not require specifying any regression model and requires no moment conditions on both features and response variable. The theoretical analysis demonstrates the proposed method enjoys not only the sure screening property but also a stronger result called rank consistency property. The extensive simulated experiments show the proposed method wins the horse racing against its competitors on various scenarios.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to extend thanks to those who so generously contributed to the work presented in this dissertation.

First of all, I would like to express my sincere gratitude to my advisor Dr. Runze Li, not only for his tremendous academic support during my Ph.D. study, but also for the questions which incented me to widen my research from various perspectives. His guidance helped me through my research and writing of this dissertation. He has taught me more than I could ever give him credit here and his endless guidance is hard to forget in the rest of my life.

Besides my advisor, I would like to thank the rest of my committee: Dr. Lingzhou Xue, Dr. Xingyuan (Ethan) Fang, and Dr. Tao Yao, for their encouragement and helpful comments. My sincere thanks also goes to my collaborator Dr. Yuan Ke, who has always provided insightful discussions when we encounter difficulties. They all have played an important role in my Ph.D. study.

Last but not the least, thanks go to my parents, sister and girlfriend for their unbelievable support. They are the most important people in my world and I dedicate this dissertation to them.

This dissertation research was partly supported by National Science Foundation grants, DMS 1512422 and DMS 1820702, and National Institute on Drug Abuse, NIH grant P50 DA039838.

Dedication

I delicate this dissertation to my parents Xingming Liu and Juan Xue, my sister Xiaomin Liu and my girlfriend Xiufan Yu, who provided unconditional support and love and have always been there for me.

Chapter 1 Introduction

Rapid development of technology allows for a large number of features to be measured and collected. It is increasingly important to be able to solve such problems involving a very large number of variables. This type of data is typically known as high-dimensional data, where the number of variables p can be much larger than the sample size n. Such high-dimensional data arises in a broad spectrum of real applications such as genomics, finance, medical imaging, sensor network, etc. For example, advanced biotechnology now allows that thousands of genes or proteins can be measured. Financial data is also of a high-dimensional nature. Hundreds or thousands of financial instruments can be measured and tracked over time at very fine time intervals in high frequency trading. In this dissertation, we try to answer following two questions:

(1) Can we construct a powerful test for the mean of high-dimensional data?

(2) Can we select important features from thousands of variables without specifying a regression model and without moment assumptions?

To answer the first question, we construct a sparse projection test via an online framework for high-dimensional data. To answer the second question, we propose a model free feature screening procedure based on projection correlation.

1.1 Linear functional and its applications

In the first part of this dissertation, we focus on the estimation of the linear functional of the form $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ and apply it to perform projection test for

high-dimensional mean vector. This linear functional $\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ has many interesting statistical applications including

- **(Projection test for mean vectors)** Let us consider a one-sample mean vector test problem in high-dimensional data. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from a p-dimensional distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Of interest is to test

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \text{ verses } H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

  where $\boldsymbol{\mu}_0$ is some given vector. Traditional methods such as Hotelling's T^2 test is not directly applicable when $p > n$ since the sample covariance matrix is not invertible. The idea of projection test is to reduce the dimension by projecting the high-dimensional vector $\mathbf{x}_i$ to a space of lower dimension. Li et al. (2015) shows that the optimal projection direction is $\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ with $\boldsymbol{\eta} = \boldsymbol{\mu}$. Then the high-dimensional $\mathbf{x}_i$'s can be projected to a 1-dimensional space by left multiply $(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^\top$. Similar conclusion holds for two-sample mean vectors test where the optimal projection direction is $\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ with $\boldsymbol{\eta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors for the two populations respectively and $\boldsymbol{\Sigma}$ is the common covariance matrix.

- **(Linear Discriminant Analysis (LDA))** Consider the linear discriminant analysis for classification problem. Assume we have two p-dimensional normal distributions $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ (class 1) and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ (class 2) with the same covariance matrix. Let $\mathbf{z}$ be a random observation that is drawn from one the these two populations with equal probabilities. The well-known Fisher's linear discriminant rule is characterized by the linear functional $\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ with $\boldsymbol{\eta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. The new observation $\mathbf{z}$ is classified into class 1 if and only if

$$(\mathbf{z} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta} \geq 0.$$

- **(Markowitz portfolio allocation problem)** We consider the Markowitz portfolio allocation problem. Suppose we have p assets $\mathbf{x} = (X_1, \ldots, X_p)^\top$ to invest, where $\mathbf{x}$ have mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For a given amount of expected return m, we want find a allocation plan such that the investment

risk is minimized. The optimal portfolio allocation is proportional to $\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ with $\boldsymbol{\eta} = \boldsymbol{\mu}$ (Chen et al. 2015).

This linear functional is typically not directly applicable in real application as it is usually unknown and needs to be estimated from data. To estimate $\boldsymbol{\beta}$, traditional approaches take two steps: (i) we first construct an estimate $\widehat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ and then (ii) estimate $\boldsymbol{\beta}$ using $\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}$ or $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\eta}}$ if $\boldsymbol{\eta}$ is unobserved, where $\widehat{\boldsymbol{\eta}}$ is an estimator of $\boldsymbol{\eta}$. Although the two-step estimator is usually asymptotically consistent in the low-dimensional setting, it may no longer be a consistent estimator in high dimensions. First, the estimate $\widehat{\boldsymbol{\Sigma}}$ is typically not invertible when $p > n$ and thus the plug-in estimator is not well-defined. A naive approach is ignoring the dependence among the variables and replace $\widehat{\boldsymbol{\Sigma}}$ by some diagonal matrix such as identity matrix $\mathbf{I}$ or $\mathrm{diag}(\widehat{\boldsymbol{\Sigma}})$, where $\mathrm{diag}(\widehat{\boldsymbol{\Sigma}})$ is a diagonal matrix with elements being the diagonal elements of $\widehat{\boldsymbol{\Sigma}}$. This approach works well when the true covariance matrix is approximately a diagonal matrix but may fail when the correlation among the variables are strong. In the setting of linear discriminant analysis, Mai et al. (2012) pointed out that ignoring correlations among variables could produce misleading feature selection and inferior classification.

Second, it is not possible to obtain a consistent estimator of $\boldsymbol{\Sigma}$ or its inverse when the dimension is high unless structural assumptions such as sparsity or low rank are imposed on $\boldsymbol{\Sigma}$ or its inverse (Fan, Liao and Liu 2016). To estimate the covariance matrix, thresholding methods are proposed to obtain a sparse estimator for $\boldsymbol{\Sigma}$ by setting small estimated elements to zero (Bickel and Levina 2008, Cai and Liu 2011*a*, Rothman et al. 2009). Regularized methods are widely used to estimate sparse precision matrix $\boldsymbol{\Sigma}^{-1}$. For example, Yuan and Lin (2007) and Lam and Fan (2009) studied the estimation of sparse precision matrix using penalized likelihood. Cai et al. (2011) and Yuan (2010) proposed to estimate the precision matrix through column-by-column regressions.

Third, consistent estimation of $\boldsymbol{\Sigma}$ or its inverse does not automatically guarantee the consistency of $\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\eta}$ or $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\eta}}$. Fan and Fan (2008) demonstrated that even in the ideal scenario where the true covariance matrix $\boldsymbol{\Sigma}$ is an identity matrix and is assumed to be known, the classical Fisher's classification rule is no better than random guessing when p is sufficiently large, due to noise accumulation in estimating $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Instead of imposing structural conditions on $\boldsymbol{\Sigma}$ or its inverse, one can impose

sparsity structure on the linear functional $\boldsymbol{\beta}$ itself. This is a plausible assumption in real applications such as portfolio selection, linear discriminant analysis and optimal direction estimation for projection test, etc. Fan and Fan (2008) demonstrated that the classification rule using all features can perform as poorly as the random guessing. Thus it is necessary to select a subset of important features for high-dimensional classification problem. Under the sparsity assumption, direct approaches based on regularization methods can be applied to estimate the linear functional. Mai et al. (2012) studied the linear discriminant analysis problem using the penalized least squares with L_1 penalty. Dantzig-type selectors are also studied in linear discriminant analysis, Markowtiz portfolio allocation problem and time series settings (Cai and Liu 2011*b*, Chen et al. 2015). However, the resulting estimators by these methods are usually biased. Nonconvex penalties such as the SCAD (Fan and Li 2001) and the MCP (Zhang 2010) attract more attention recently. Comparing with the L_1 penalty, the estimator given by nonconvex penalties enjoys more desirable statistical properties such as asymptotic unbiasedness and oracle property. Penalized linear regression with nonconvex penalty has been well studied, see Fan and Li (2001), Zhang (2010) and Wang et al. (2013). To avoid the drawbacks of traditional methods and direct approaches using L_1 penalty and Dantzig selector, we propose an approach to estimate the linear functional based on regularized quadratic programming with nonconvex penalty and linear constraint.

In chapter 3, we study a general regularized quadratic programming with nonconvex penalty and linear constraint. Under the assumption that the quadratic form satisfies the restricted strong convexity (RSC) condition, we establish the deterministic L_1 and L_2 error bounds for our estimator. Our theory applies to any stationary point that satisfies the first order necessary conditions to be a local minimum. Further assuming the strict dual feasibility, we also show that the stationary point is unique and establish the support recovery and L_∞ error bound. In addition, we consider three applications of the regularized quadratic programming: (1) estimation of linear functional, (2) F-type test for regression coefficients and (3) sparse linear discriminant analysis. We establish the convergence rates in terms of L_1 and L_2 norms for stationary points. It is quite challenging to solve such noncovnex optimization problem especially in the high-dimensional setting. We propose an ADMM algorithm with local linear approximation (LLA), which approximates the nonconvex penalty function by its first order expansion.

It is guaranteed that the estimator converges to a local minimum and thus the numerical solution fits in our theory. This ADMM algorithm can naturally handle the linear equality constraint. A BIC-type criteria is proposed to select the tuning parameter in the penalty function.

In chapter 4, we propose two projection tests for high-dimensional mean vector. Assuming that the optimal projection direction is sparse, we apply the regularized quadratic programming discussed in chapter 3 to estimate optimal projection direction. The first test is the data splitting projection test. The entire dataset is partitioned into two sets. We use the first set to estimate the optimal projection direction and perform the test only using the data in the second set. This data splitting projection test achieves an exact t-test under normality assumption. The second test is the online projection test. We update the estimation of optimal projection direction whenever a new observation arrives. We establish the asymptotic normal distribution of the proposed test statistic under both null hypothesis and alternative hypothesis, based on which we derive the power function under alternative. We also propose a mini-batch version of the online projection test, which updates the estimation of the optimal projection direction when a batch of new observations arrive and thus reduces the computational burden. Both of the online projection tests improve the power of the data splitting projection test. We also conduct numerical studies to compare the finite sample performance of our proposed projection tests with several existing tests including sum-of-squares-type tests, maximum-type tests and other projection tests. The numerical results show that the proposed projection tests can keep the type I error rate well and are much more powerful than other existing tests.

1.2 Robust and model free feature screening

Datasets with ultra-high dimensional features characterize many contemporary research problems in machine learning, statistics, engineering, social science, finance and so on. When the features contain redundant or noisy information, estimating their functional relationship with the response can become quite challenging in terms of computational expediency, statistical accuracy and algorithmic stability (Fan et al. 2009). To overcome such challenges caused by ultra-high dimensionality, Fan and Lv (2008) proposed a sure independence screening (SIS) procedure which

aims to screen out the redundant features by ranking their marginal Pearson correlations. The SIS method is named after the sure independence screening property which states that the selected subset of features contains all the active ones with probability approaching one. The promising numerical performance soon made SIS popular among ultra-high dimensional studies. The sure screening idea has be applied to many important statistical problems including generalized linear model (Fan and Song 2010), multi-index semi-parametric models (Zhu et al. 2011), nonparametric models (Fan et al. 2011, Liu et al. 2014), quantile regression (He et al. 2013, Wu and Yin 2015) and compress sensing (Xue and Zou 2011) among others.

The idea of screening is to deliver a computationally efficient way to reduce the dimensionality of the feature space from a very high scale to a moderate one. The researchers will then benefit both computationally and statistically from learning the data in a much reduced feature space. Besides the sure screening property, we argue an appealing screening method should satisfy the following two properties. First, the screening method should be model free, which means that the screening method can be implemented without specifying a regression model. In ultra-high dimensional regime, it is challenging if not impossible to specify a correct regression model with existence of the huge number of redundant features. Hence the model free property is desired as it guarantees the effectiveness of the screening method in the presence of model mis-specification. The model free screening method becomes a hot research topic in recent years, see Zhu et al. (2011), Li et al. (2012), Mai and Zou (2015), He et al. (2013), Cui et al. (2015) and the references therein. The second property is robustness which means the screening method should be insensitive to outliers. Assumption like sub-Gaussianity is usually not realistic in ultra-high dimensional applications. Even when the sub-gaussian assumption is satisfied on the population level, they can be easily violated in the realized sample simply due to ultra-high dimensionality. Therefore the screening method which is sensitive to outliers may perform poorly in real applications. The robust screening method also draws certain amount of attention recently. He et al. (2013), Wu and Yin (2015) and Ma et al. (2017) among others considered quantile based screening which adapts to heavy-tailed data. Wang (2012) and Fan, Ke and Wang (2016) developed screening methods for strongly correlated features.

In chapter 5, we propose a model free feature screening method. The proposed

method is based on ranking the projection correlations between features and the response. The projection correlation, proposed by Zhu et al. (2017), is a measure of dependence between two random vectors which enjoys several nice probability properties. For example, the projection correlation is well-defined for any two random vectors of any dimensions and no moment conditions are required for the two random vectors. In addition, the estimation is free of tuning parameters. The proposed screening procedure does not require specifying any regression model and is insensitive to outliers. As the projection correlation is dimension free to both random vectors, the proposed screening method can be applied to multi-task learning problems (Caruana 1997). The theoretical analysis demonstrates the proposed method enjoys not only the sure screening property but also a stronger result called rank consistency property. The only condition required is a minimum signal gap between active and inactive features. The extensive simulated experiments show the proposed method wins the horse racing against its competitors on various scenarios.

1.3 Organization of this dissertation

The rest of this dissertation is organized as follows. In chapter 2, we provide literature review on topics that are related to this dissertation, including regularized methods in high-dimensional linear model, hypothesis testing for high-dimensional mean vector and feature screening for high-dimensional data. In chapter 3, we first study a general regularized quadratic programming with nonconvex penalty and linear constraint. Then we further three applications of the regularized quadratic programming: (1) estimation of linear functional, (2) F-type test for regression coefficients and (3) sparse linear discriminant analysis. In chapter 4, we propose two projection tests for high-dimensional mean vector using the optimal projection direction. One is based on a data splitting approach and the other one is based on an online framework for such projection test. In chapter 5, we propose a model free feature screening via the projection correlation. In chapter 6, we conclude this dissertation and discuss future work.

Chapter 2 Literature Review

2.1 Variable selection via regularization

2.1.1 An overview

Variable selection has become a popular and fundamental problem in high dimensional regression where the underlying model has a sparse representation. A large number of predictors are usually collected to reduce possible modeling biases. However, sparse models are preferable because of the simplicity and interpretability. In addition, identifying important predictors can improve the prediction accuracy. Therefore, it is necessary to select important predictors and only include these important predictors in the model. Over the past two decades, many model selection methods have been developed. A majority part of them are based on the regularized M-estimation including the Lasso (Tibshirani 1996), the SCAD (Fan and Li 2001), th elastic net (Zou and Hastie 2005), and the Dantzig selector (Candes and Tao 2007), among others. These methods have attracted a large amount of theoretical and algorithmic studies. See Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Fan and Lv (2008), Zou and Li (2008), Bickel et al. (2009), Zhang (2010), and references therein.

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^p$ is the response vector, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times p}$ is the covariate matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the unknown regression coefficient,

and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^p$ are independent and identically distributed random errors. When $p < n$, the ordinary least squares estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

In high-dimensional setting where $p \geq n$, $\widehat{\boldsymbol{\beta}}_{\text{ols}}$ is not well-defined since $\mathbf{X}^\top \mathbf{X}$ is not invertible. A common assumption is that the true parameter $\boldsymbol{\beta}^\star = (\beta_1^\star, \ldots, \beta_p^\star)^\top$ is sparse, meaning that most elements in $\boldsymbol{\beta}^\star$ are zero. Regularized least squares are widely used to select important variables and estimate the regression coefficient simultaneously. The regularized least squares takes the following form

$$\frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.1}$$

where $p_\lambda(\cdot)$ is some penalty function. The regularized least squares estimator is defined as the minimizer of (2.1). Various penalty functions have been proposed and their theoretical properties and numeric performances are well studied, see Tibshirani (1996), Fan and Li (2001), Zou (2006), Zou and Hastie (2005) and Zhang (2010). These penalty functions can be categorized into two classes: convex penalty and nonconvex penalty. Convex penalty such as Lasso penalty is very popular due to its attractive computation advantage. However, the resulting estimator of the Lasso penalty is usually biased. More recently, nonconvex penalties such as the SCAD and the MCP attract more attention since it has more desirable statistical properties (Fan and Li 2001, Zhang 2010, Xue et al. 2012, Fan, Xue and Zou 2014). The computation with a nonconvex penalty can be very challenging since we need to solve a nonconvex optimization problem. In the rest of this section, we will focus on the two most popular penalties: the Lasso penalty and the folded-concave penalty.

2.1.2 Variable selection with L_1 penalty

Tibshirani (1996) first introduced the Lasso regression, which minimizes the least squares with L_1 penalty

$$\widehat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.2}$$

The Lasso regression has the following attractive properties. The Lasso shrinks the resulting estimator $\widehat{\boldsymbol{\beta}}_{\text{Lasso}}$ towards 0, and some of the coefficients are estimated exactly to be 0. Therefore, Lasso regression can select important variables and estimate the coefficients simultaneously. Lasso regression is also computationally attractive since it only needs to solve a convex optimization problem. The Lasso estimator depends on the choice of the tuning parameter λ, which controls the amount of shrinkage applied to the estimator. A large λ would shrink many elements of the coefficient to be 0.

Before introducing the theoretical properties of the Lasso, we first introduce some notations and definitions. Recall that $\boldsymbol{\beta}^\star$ is the true regression coefficient and is assumed to be sparse. Let $\mathcal{A}^\star = \{j : \beta_j^\star \neq 0\}$ be the index set of nonzero components in $\boldsymbol{\beta}^\star$. Let $\widehat{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}^\star$ and $\mathcal{A}(\widehat{\boldsymbol{\beta}}) = \{j : \widehat{\beta}_j \neq 0\}$ be the index set of nonzero components in $\widehat{\boldsymbol{\beta}}$. An estimator $\widehat{\boldsymbol{\beta}}$ is (estimation) consistent if $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^\star$ in probability. An estimator $\widehat{\boldsymbol{\beta}}$ is model consistent if $\lim_{n\to\infty} \mathrm{P}(\mathcal{A}(\widehat{\boldsymbol{\beta}}) = \mathcal{A}^\star) = 1$. An estimator $\widehat{\boldsymbol{\beta}}$ is sign consistent if $\mathrm{P}(\mathrm{sgn}(\widehat{\boldsymbol{\beta}}) = \mathrm{sgn}(\boldsymbol{\beta}^\star)) \to 1$.

Under the conditions that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed random variables with mean 0 and variance σ^2 and $\frac{1}{n}\mathbf{X}^\top\mathbf{X} \to \mathbf{C}$, where $\mathbf{C}$ is a positive definite matrix, Knight and Fu (2000) proved that the Lasso estimator is estimation consistent when p is fixed. Without loss of generality, assume that $\mathcal{A}^\star = \{1, 2, \ldots, s\}$ and the complement set $\mathcal{A}^{\star c} = \{s+1, s+2, \ldots, p\}$. Let

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix},$$

where $\mathbf{C}_{11}$ is a $s \times s$ matrix. If $\lambda/\sqrt{n} \to \lambda_0 \geq 0$, then

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{Lasso}} - \boldsymbol{\beta}^\star) \xrightarrow{d} \arg\min\ V(\mathbf{u}),$$

where $\xrightarrow{d}$ means convergence in distribution, $\mathbf{u} = (u_1, \ldots, u_p)^\top \in \mathbb{R}^p$,

$$V(\mathbf{u}) = -2\mathbf{u}^\top\mathbf{W} + \mathbf{u}^\top\mathbf{C}\mathbf{u} + \lambda_0 \sum_{j=1}^{p} [u_j \mathrm{sgn}(\beta_j^\star) I(\beta_j^\star \neq 0) + |u_j| I(\beta_j^\star = 0)],$$

$\mathbf{W}$ has a $N(\mathbf{0}, \sigma^2\mathbf{C})$ distribution and $I(\cdot)$ is the indicator function. This theorem shows that the Lasso estimator is root-n estimation consistent. Under the same

conditions, Zou (2006) showed that

$$\limsup_{n\to\infty} \mathrm{P}(\mathcal{A}(\widehat{\boldsymbol{\beta}}_{\text{Lasso}}) = \mathcal{A}^\star) \le c < 1,$$

where c is a constant depending on the true model. This result shows that when $\lambda/\sqrt{n} \to \lambda_0$, the Lasso estimator is not model consistent. The optimal estimation of Lasso regression is achieved only when $\lambda = O(\sqrt{n})$, however it leads to inconsistent model selection. Zou (2006) and Zhao and Yu (2006) proposed a necessary condition for Lasso estimator to be model consistent, which states that there exists some sign vector $\mathbf{d} = (d_1, \ldots, d_s)^\top$ with $d_j = 1$ or -1, such that

$$\|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{d}\|_\infty \le 1.$$

This type of condition is referred to as the irrepresentable condition. The irrepresentable condition closely resembles a regularization constraint on the regression coefficients of the irrelevant covariates on the relevant covariates. This irrepresentable condition is almost necessary and sufficient for a Lasso estimator to be model consistent. However this irrepresentable condition is nontrivial. Zou (2006) constructed an interesting example in which the irrepresentable condition fails. Under the irrepresentable condition, Zhao and Yu (2006) showed that if $\lambda_n/n \to 0$ and $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ for some $0 \le c < 1$, then

$$\mathrm{P}(\mathrm{sgn}(\widehat{\boldsymbol{\beta}}_{\text{Lasso}}) = \mathrm{sgn}(\boldsymbol{\beta}^\star)) = 1 - o(e^{-n^c}).$$

The Lasso regression forces all the coefficients to be equally penalized even when $|\beta_j^\star|$ is large. As a consequence, the Lasso estimator is usually biased and cannot achieve the oracle property. Zou (2006) proposed the adaptive Lasso by assigning different weights to different coefficients

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|,$$

where $\mathbf{w} = (w_1, \ldots, w_p)^\top$ is a known weight vector. The insight is that if $|\beta_j^\star|$ is large, it should be penalized less and if $|\beta_j^\star|$ is small, it should be penalized more. Suppose $\tilde{\boldsymbol{\beta}}$ is a root-n consistent estimator of $\boldsymbol{\beta}^\star$, one can construct the weight

vector by $\widehat{\mathbf{w}} = 1/|\tilde{\boldsymbol{\beta}}|^{\gamma}$ with $\gamma > 0$. In practice, the root-n consistent estimator can be obtained from the ordinary least squares estimator. The adaptive lasso estimator $\widehat{\boldsymbol{\beta}}_{\text{ada}}$ is given by

$$\widehat{\boldsymbol{\beta}}_{\text{ada}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j |\beta_j|.$$

With a proper choice of λ, the adaptive lasso estimator enjoys the following (weak) oracle property. Suppose that $\lambda/\sqrt{n} \to 0$ and $\lambda n^{(\gamma-1)/2} \to \infty$, then

$$\begin{aligned} &\lim_{n\to\infty} \mathrm{P}(\mathcal{A}(\widehat{\boldsymbol{\beta}}_{\text{ada}}) = \mathcal{A}^{\star}) = 1, \\ &\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{ada},\mathcal{A}^{\star}} - \boldsymbol{\beta}^{\star}_{\mathcal{A}^{\star}}) \to N(\mathbf{0}, \sigma^2 \mathbf{C}_{11}^{-1}), \end{aligned}$$

where $\boldsymbol{\beta}_{\mathcal{A}^{\star}}$ represents the subvector of $\boldsymbol{\beta}$ corresponding to the subset $\mathcal{A}^{\star}$. The oracle property implies that the estimating procedure can estimate the zero coefficients as exact zero with probability approaching one, and estimate the nonzero coefficients as efficiently as if the true sparsity pattern is known in advance. The adaptive Lasso is essentially a convex optimization with L_1 penalty and efficient algorithms for solving the Lasso can be used to compute the adaptive Lasso estimator.

Theoretical properties of Lasso estimator in the high-dimensional setting are studied in Bickel et al. (2009). In order to establish the error bound, Bickel et al. (2009) imposed the following restricted eigenvalue (RE) assumption. This assumption plays an important role in the analysis of Lasso estimator with high dimension. For some integer s such that $1 \le s \le p$ and a positive number c_0, the design matrix $\mathbf{X}$ satisfies $\mathrm{RE}(s, c_0)$ condition if

$$\min_{J \subseteq \{1,\ldots,p\}, |J| \le s} \quad \min_{\boldsymbol{\delta} \neq \mathbf{0}, \|\boldsymbol{\delta}_{J^c}\|_1 \le c_0 \|\boldsymbol{\delta}_J\|_1} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\sqrt{n}\|\boldsymbol{\delta}_J\|_2} = \kappa(s, c_0) > 0. \tag{2.3}$$

This condition is known to be relatively mild on the design matrix for high-dimensional data and is much weaker than the irrepresentable condition. Let condition $\mathrm{RE}(s, 3)$ be satisfied and standardize the $\mathbf{X}$ such that all the diagonal elements of the matrix $\mathbf{X}^{\top}\mathbf{X}/n$ be equal to 1. Let $\lambda = A\sigma\sqrt{\log p/n}$ with $A > 2\sqrt{2}$,

then with probability at least $1 - p^{1-A^2/8}$, we have

$$\|\widehat{\boldsymbol{\beta}}_{\text{Lasso}} - \boldsymbol{\beta}^\star\|_1 \leq \frac{16A}{\kappa(s,3)}\sigma s\sqrt{\frac{\log p}{n}}.$$

Dantzig selector proposed by Candes and Tao (2007) is also very popular when $p > n$. The Dantzig estimator is defined as the solution to the following problem

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \\ &\text{subject to } \|\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \leq \lambda. \end{aligned} \tag{2.4}$$

Clearly, the Dantzig selector can be efficiently solved by linear programming. If assumption RE$(s, 1)$ is satisfied and choose $\lambda = A\sigma\sqrt{\log p/n}$ for some $A > \sqrt{2}$, then with probability at least $1 - p^{1-A^2/2}$, we have

$$\|\widehat{\boldsymbol{\beta}}_{\text{Dantzig}} - \boldsymbol{\beta}^\star\|_1 \leq \frac{8A}{\kappa(s,1)}\sigma s\sqrt{\frac{\log p}{n}}.$$

The Dantzig selector is closely related to the Lasso estimator and has similar performance to the Lasso estimator under the sparsity scenario (Bickel et al. 2009).

2.1.3 Variable selection with nonconvex penalty

The Lasso estimator is usually biased, which motivates researchers to use other type of penalty. Fan and Li (2001) argued that a good penalty function should result in an estimator with the following three properties:

1. *Unbiasedness*: The resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modeling bias.

2. *Sparsity*: The resulting estimator is a thresholding rule, which sets small estimated coefficients to zero to reduce model complexity.

3. *Continuity*: The resulting estimator is continuous in data to avoid instability in model prediction.

Note that none of the L_q penalty $p_\lambda(|t|) = \lambda|t|^q$ satisfy all the three properties. When $q < 1$, the resulting estimator is not continuous in data. When $q > 1$, it

does not produce a sparse solution. When $q = 1$, the resulting estimator is usually biased. Fan and Li (2001) proposed the Smoothly Clipped Absolute Deviation (SCAD) penalty that satisfies all the three properties. The first derivative of the SCAD penalty is given by

$$p'_{\text{SCAD},\lambda}(|t|) = \lambda \left\{ I(|t| \leq \lambda + \frac{(a\lambda - |t|)_+}{(a-1)\lambda}) I(|t| > \lambda) \right\}, \tag{2.5}$$

for some $a > 2$, where $I(\cdot)$ is the indicator function and $a_+ = aI(a > 0)$ is the positive part of a. Consider the penalized least squares with the SCAD penalty

$$\widehat{\boldsymbol{\beta}}_{\text{SCAD}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p_{\text{SCAD},\lambda}(|\beta_j|). \tag{2.6}$$

Fan and Li (2001) proved that the SCAD estimator $\widehat{\boldsymbol{\beta}}_{\text{SCAD}}$ is root-n consistent and has the oracle property when p is fixed. More specifically, if $\max\{p''_{\text{SCAD},\lambda}(|\beta_j^\star|) : \beta_j^\star \neq 0\} \to 0$, then

$$\|\widehat{\boldsymbol{\beta}}_{\text{SCAD}} - \boldsymbol{\beta}^*\|_2^2 = O_p(n^{-1/2} + a_n),$$

where $p''_{\text{SCAD},\lambda}(\cdot)$ is the second derivative of $p_{\text{SCAD},\lambda}(\cdot)$ and $a_n = \max\{p'_{\text{SCAD},\lambda}(|\beta_j^\star|) : \beta_j^\star \neq 0\}$. Further assume

$$\liminf_{n\to\infty} \liminf_{t\to 0+} p'_{\text{SCAD},\lambda}(|t|)/\lambda > 0,$$

$\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ as $n \to \infty$, then with probability approaching to 1, the SCAD estimator enjoys the oracle property.

The nonconvex SCAD penalty brings extra computational burden since we need to solve a nonconvex optimization problem. Fan and Li (2001) proposed a unified algorithm using local quadratic approximation (LQA) to solve (2.6). The idea is to approximate the penalty function $p_\lambda(\cdot)$ by its second order Taylor expansion. If β_{j0} is close to β_j, then

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}(\beta_j^2 - \beta_{j0}^2).$$

By the local quadratic approximation, the SCAD estimator can be obtained by solving a quadratic programming. Instead of using the second order Taylor ex-

pansion, Zou and Li (2008) proposed anther algorithm which approximates the penalty function by its first order Taylor expansion. This algorithm is known as the local linear approximation (LLA) algorithm. If β_{j0} is close to β_j, then the penalty function can be approximated by

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + p'_\lambda(|\beta_{j0}|)(|\beta_j| - |\beta_{j0}|).$$

The LLA algorithm is distinguished from the LQA algorithm in that the final estimates naturally adopt a sparse representation. The LLA algorithm inherits the good features of Lasso in terms of computational efficiency and can be solved by efficient algorithms. In particular, Zou and Li (2008) proposed the one-step LLA estimator and showed that this one-step LLA estimator enjoys the oracle property. Let the initial estimate $\boldsymbol{\beta}^{(0)}$ be ordinary least squares estimator. Then the one-step LLA estimator is obtained by

$$\boldsymbol{\beta}^{(1)} = \arg\min \ \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(0)}|)|\beta_j|.$$

Similar to the SCAD penalty, the minimax concave penalty (MCP) function is defined as

$$p'_{\mathrm{MCP}}(|t|) = a^{-1}(a\lambda - |t|)_+, \tag{2.7}$$

for some $a > 0$. Zhang (2010) showed that the resulting MCP estimator is sign consistent with high probability without assuming the irrepresentable condition and attains the minimax rate. The penalized linear unbiased selection (PLUS) algorithm was introduced in Zhang (2010) to obtain the MCP estimator. Fan, Xue and Zou (2014) systematically studied the family of folded concave penalty function $p_\lambda(t)$ under a general framework. The $p_\lambda(t)$ satisfies the following conditions

(1) $p_\lambda(t)$ is increasing and concave in $t \in [0, \infty)$ with $p_\lambda(0) = 0$;

(2) $p_\lambda(t)$ is differentiable in $t \in (0, \infty)$ with $p'_\lambda(0) := p'_\lambda(0+) \geq a_1\lambda$;

(3) $p'_\lambda(t) \geq a_1\lambda$ for $t \in (0, a_2\lambda]$;

(4) $p'_\lambda(t) = 0$ for $t \in [a\lambda, \infty)$ with $a > a_2$,

where a, a_1, a_2 are fixed positive constants. The folded concave penalty includes the SCAD and the MCP. Fan, Xue and Zou (2014) considered a general problem

taking the following form,

$$\min_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.8}$$

where $\ell_n(\boldsymbol{\beta})$ is a convex loss function and $p_\lambda(\cdot)$ is a folded concave penalty. Define the oracle estimator as we know the true support set $\mathcal{A}^\star$ in advance,

$$\widehat{\boldsymbol{\beta}}^{(o)} = (\widehat{\boldsymbol{\beta}}^{(o)}_{\mathcal{A}^\star}, \mathbf{0}) = \underset{\boldsymbol{\beta}, \boldsymbol{\beta}_{\mathcal{A}^{\star c}} = \mathbf{0}}{\arg\min} \ \ell_n(\boldsymbol{\beta}).$$

For example, the oracle estimator in the linear regression setting with $\ell_n(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ is characterized by $(\widehat{\boldsymbol{\beta}}^{(o)}_{\mathcal{A}^\star}, \mathbf{0})$ with

$$\widehat{\boldsymbol{\beta}}^{(o)}_{\mathcal{A}^\star} = (\mathbf{X}_{\mathcal{A}^\star}^\top \mathbf{X}_{\mathcal{A}^\star})^{-1} \mathbf{X}_{\mathcal{A}^\star}^\top \mathbf{y},$$

where $\mathbf{X}_{\mathcal{A}^\star}$ consists of the columns from $\mathbf{X}$ indexed by $\mathcal{A}^\star$. Fan, Xue and Zou (2014) proved that the LLA estimator has the strong oracle property, that is the LLA estimator equals the oracle estimator with high probability. Wang et al. (2013) proposed the ConCave Convex Procedure (CCCP) algorithm to solve the penalized linear squares with nonconvex penalty. The idea is based on the observation that nonconvex penalties such as the SCAD and the MCP can be written as the difference of two convex functions, or equivalently, the sum of one convex function and one concave function. Let $p_\lambda(|\beta|)$ be a nonconvex penalty and suppose it has the following decomposition,

$$p_\lambda(|\beta|) = J_\lambda(|\beta|) + \lambda|\beta|,$$

where $J_\lambda(|\beta|)$ is a differentiable concave function. For example, for the SCAD,

$$\begin{aligned} J_\lambda(|\beta|) = &- \frac{\beta^2 - 2\lambda|\beta| + \lambda^2}{2(a-1)} I(\lambda \le |\beta| \le a\lambda) \\ &+ \left(\frac{(a+1)^2\lambda^2}{2} - \lambda|\beta| \right) I(|\beta| > a\lambda). \end{aligned}$$

For the MCP,

$$J_\lambda(|\beta|) = \frac{\beta^2}{2a} I(0 \le |\beta| < a\lambda) + (a\lambda^2/2 - \lambda|\beta|) I(|\beta| \ge a\lambda).$$

Then the penalized least squares in (2.1) can be written as

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} J_\lambda(|\beta_j|) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.9}$$

The concave function $J_\lambda(|\beta|)$ can be approximated by its tight convex upper bound. Given a current estimator $\boldsymbol{\beta}^{(k)}$, the tight convex upper bound of (2.9) is given by

$$Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}, \lambda) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{p} \mathrm{sgn}(\beta_j^{(k)}) J'_\lambda(|\beta_j^{(k)}|)\beta_j + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.10}$$

where J'_λ is the derivative of J_λ. We update the current solution by

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}, \lambda).$$

The CCCP algorithm has the descent property, that is, the objective function decreases after each iteration. Starting with the initial value $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, the calibrated algorithm in Wang et al. (2013) consists of the following two steps:

Step 1. Let $\widehat{\boldsymbol{\beta}}^{(1)}(\lambda) = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(0)}, \tau\lambda)$ for some $\tau > 0$.

Step 2. Let $\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^{(1)}(\lambda), \tau\lambda)$.

In step 1, a smaller tuning parameter $\tau\lambda$ is adopted to increase the estimation accuracy. In practice, one can set $\tau = \lambda$ or $\tau = 1/\log n$. Assume design matrix $\mathbf{X}$ satisfies

$$\min_{\boldsymbol{\delta} \neq \mathbf{0}, \|\boldsymbol{\delta}_{\mathcal{A}^{\star c}}\|_1 \le 3\|\boldsymbol{\delta}_{\mathcal{A}^\star}\|_1} \frac{\|\mathbf{X}\boldsymbol{\delta}\|_2}{\sqrt{n}\|\boldsymbol{\delta}_{\mathcal{A}^\star}\|_2} = \kappa > 0,$$

$\tau = o(1)$, $\lambda = o(\min\{|\beta_j^\star| : j \in \mathcal{A}^\star\})$ and $\tau\kappa^{-2}s = o(1)$,then for all n sufficiently large

$$\mathrm{P}(\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^{(o)}) \ge 1 - 8p\exp\{-n\tau^2\lambda^2/(8\sigma^2)\}.$$

Further assume $n\tau^2\lambda^2 \to \infty$ and $\log p = o(n\tau^2\lambda^2)$, then

$$\mathrm{P}(\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^{(o)}) \to 1 \quad \text{as } n \to \infty.$$

To select the tuning parameter λ, Wang et al. (2013) extended the BIC criterion to the high-dimensional setting. For a given λ, the high-dimensional BIC criterion (HBIC) is defined by

$$\mathrm{HBIC}(\lambda) = \log(\widehat{\sigma}_\lambda^2) + \mathrm{df}_\lambda \frac{C_n \log p}{n},$$

where $\mathrm{df}_\lambda = \|\widehat{\boldsymbol{\beta}}(\lambda)\|_0$, the number of non-zero elements in $\widehat{\boldsymbol{\beta}}(\lambda)$ and $\widehat{\sigma}_\lambda^2 = n^{-1}\mathrm{SSE}_\lambda$ with $\mathrm{SSE}_\lambda = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)\|_2^2$, C_n is a sequence of numbers that diverges to ∞, e.g. $C_n = \log\log n$. The tuning parameter is selected by

$$\widehat{\lambda} = \arg\min_{\lambda} \mathrm{HBIC}(\lambda).$$

Wang et al. (2013) proved that the resulting estimator selected by HBIC can recover the true support with probability approaching to 1. Loh and Wainwright (2015) studied the statistical properties of regularized M-estimators in which they allow both the loss function and penalty to be nonconvex. In particular, the regularized M-estimator is of the following form

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{g(\boldsymbol{\beta}) \le R, \boldsymbol{\beta} \in \Omega} \ell_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.11}$$

where $g(\cdot)$ is a convex function satisfying the lower bound $g(\boldsymbol{\beta}) \ge \|\boldsymbol{\beta}\|_1$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$, $R > 0$ is a tuning parameter such that $\boldsymbol{\beta}^\star$ is a feasible point and Ω is some convex set containing $\boldsymbol{\beta}^\star$. In settings where such a constraint Ω is extraneous, one can simply set $\Omega = \mathbb{R}^p$. The penalty function $p_\lambda(\cdot)$ satisfies the following conditions:

(1) $p_\lambda(0) = 0$ and is symmetric around zero.

(2) $p_\lambda(t)$ is nondecreasing on the nonnegative real line.

(3) For $t > 0$, the function $p_\lambda(t)/t$ is nonincreasing in t.

(4) $p_\lambda(t)$ is differentiable for $t \neq 0$ and $\lim_{t \to 0^+} p'_\lambda(t) = \lambda L$.

(5) There exists $\mu > 0$ such that $p_{\lambda,\mu}(t) := p_\lambda(t) + \frac{\mu}{2}t^2$ is convex.

It is easy to see the standard L_1 penalty, the SCAD and the MCP satisfy all these conditions. Loh and Wainwright (2015) requires the loss function ℓ_n to be differentiable, but does not require it to be convex. Instead, they impose a weaker condition known as the restricted strong convexity (RSC), which involves a lower bound on the remainder in the first order expansion of ℓ_n. In particular, they assume

$$\langle \nabla \ell_n(\boldsymbol{\beta}^\star + \boldsymbol{\Delta}) - \nabla \ell_n(\boldsymbol{\beta}^\star), \boldsymbol{\Delta} \rangle \geq \begin{cases} \alpha_1 \|\boldsymbol{\Delta}\|_2^2 - \tau_1 \frac{\log p}{n} \|\boldsymbol{\Delta}\|_1^2, & \|\boldsymbol{\Delta}\|_2 \leq 1, \\ \alpha_2 \|\boldsymbol{\Delta}\|_2 - \tau_2 \sqrt{\frac{\log p}{n}} \|\boldsymbol{\Delta}\|_1, & \|\boldsymbol{\Delta}\|_2 \geq 1, \end{cases} \tag{2.12}$$

where the α_j's are strictly positive constants and the τ_j's are nonnegative constants. Suppose that $\widehat{\boldsymbol{\beta}}$ satisfies the first order necessary conditions to be a local minimum of the program (2.11)

$$\langle \nabla \ell_n(\widehat{\boldsymbol{\beta}}) + \nabla p_\lambda(\widehat{\boldsymbol{\beta}}), \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}} \rangle \geq 0, \quad \text{for all feasible } \boldsymbol{\beta} \in \mathbb{R}^p. \tag{2.13}$$

When $\widehat{\boldsymbol{\beta}}$ lies in the interior of the constraint set, this condition reduces to the usual zero-subgradient condition

$$\nabla \ell_n(\widehat{\boldsymbol{\beta}}) + \nabla p_\lambda(\widehat{\boldsymbol{\beta}}) = 0.$$

Suppose the loss function ℓ_n satisfies the RSC condition (2.12) with $\frac{3}{4}\mu < \alpha_1$ and consider any choice of λ such that

$$\frac{4}{L} \cdot \max\left\{ \|\nabla \ell_n(\boldsymbol{\beta}^\star)\|_\infty, \alpha_2 \sqrt{\frac{\log p}{n}} \right\} \leq \lambda \leq \frac{\alpha_2}{6RL},$$

and suppose $n \geq 16R^2 \max\{\tau_1^2, \tau_2^2\} \log p / \alpha_2^2$. Then any vector $\widehat{\boldsymbol{\beta}}$ satisfying the first order necessary conditions (2.13) has the following error bounds

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \leq \frac{6\lambda L \sqrt{s}}{4\alpha_1 - 3\mu} \quad \text{and} \quad \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \leq \frac{24\lambda L s}{4\alpha_1 - 3\mu}.$$

2.2 High-dimensional mean vector test

2.2.1 An overview

One-sample mean vector test or two-sample testing on the equality of two means is a fundamental problem in high-dimensional statistics. These tests are commonly encountered in genome-wide association studies. For instance, Chen and Qin (2010) performed a hypothesis testing to identify sets of genes which are significant with respect to certain treatments in a genetics research. Xu et al. (2016) applied various tests to the bipolar disorder dataset from a genome-wide association study collected by Consortium (2007) in which one would like to test whether there is any association between a disease and a large number of genetic variants. In these applications, the dimension of the data p is often much larger than the sample size n. Traditional methods such as Hotelling's T^2 test (Hotelling 1931) either cannot be directly applied or have low power against the alternative.

Consider a size n random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a p-dimensional population $\mathbf{x}$ with finite mean $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. Of interest is to test the following hypothesis

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \tag{2.14}$$

for some known vector $\boldsymbol{\mu}_0$. This problem is typically referred to as the one-sample hypothesis testing problem in multivariate analysis and has been extensively studied when $p < n$ and p is fixed. Without loss of generality, we assume $\boldsymbol{\mu}_0 = \mathbf{0}$ and the one-sample problem (2.14) becomes

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}. \tag{2.15}$$

In most of the cases, the test statistic constructed for one-sample problem can be easily extended to two-sample problem and the theoretical results hold as well. For this reason, we only focus on the one-sample problem (2.15) and assume $\boldsymbol{\mu}_0 = \mathbf{0}$ in this section.

Let $\bar{\mathbf{x}}$ and $\mathbf{S}$ be the sample mean vector and the sample covariance matrix

respectively,

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top. \tag{2.16}$$

The Hotelling's T^2 statistic for problem (2.14) is given by $T^2 = n\bar{\mathbf{x}}^\top\mathbf{S}^{-1}\bar{\mathbf{x}}$. Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are normally distributed, under H_0, we have

$$T^2 \sim \frac{(n-1)p}{n-p}F_{p,n-p},$$

where $F_{p,n-p}$ is the F distribution with degrees of freedom p and $n-p$. For the one sample problem (2.15), the Hotelling's T^2 test is equivalent to the likelihood ratio test. The Hotelling's T^2 requires that the sample covariance matrix $\mathbf{S}$ is invertible and cannot be directly used in high-dimensional setting where $p > n$. Despite the singularity of $\mathbf{S}$, it has been observed that the power of the Hotelling's T^2 test can be adversely affected even when $p < n$, if $\mathbf{S}$ is nearly singular; see Bai and Saranadasa (1996) and Pan and Zhou (2011).

Several one-sample tests for high-dimensional data have been proposed recently. These tests can be roughly categorized into three types. The first type is based on the sum-of-squares of the sample mean and can be regarded as modified versions of the Hotelling's T^2 test. Since the sample covariance matrix $\mathbf{S}$ is not invertible, these tests replace $\mathbf{S}$ by the identity matrix $\mathbf{I}$ or some other diagonal matrix leading to a sum-of-squares test statistic. The second type is based on the maximum of a sequence of tests. The third type is the projection test. The idea is to project the high-dimensional vector $\mathbf{x}_i$ onto a low-dimensional space and then we can apply the traditional methods such as Hotelling's T^2 to perform the test. These types of tests are powerful only against certain alternatives. For example, if the true mean $\boldsymbol{\mu}$ is dense in the sense that there is a large proportion of small to moderate nonzero components, then sum-of-squares type test is more powerful. In contrast, if the true mean $\boldsymbol{\mu}$ is sparse in the sense that there are only a few nonzero components in $\boldsymbol{\mu}$, then the maximum-type test is more powerful. In practice, since the true alternative hypothesis is unknown, it is unclear how to choose a powerful test. Furthermore, there are intermediate situations in which none of these tests is powerful (Xu et al. 2016). Some recent work showed that the sum-of-squares-type test statistic and the maximum-type test statistic are asymptoticly independent and hence combined the two test statistics to boost the power, see Li and Xue (2015) and Li et al. (2018).

2.2.2 Sum-of-squares-type Test

When $p > n$, the sample covariance matrix $\mathbf{S}$ is not invertible. To deal with the singularity issue, one simple way is ignoring the dependence among the p variables and replacing $\mathbf{S}$ by the identity matrix $\mathbf{I}$.

Without assuming that data comes from normal distribution, Bai and Saranadasa (1996) studied the hypothesis testing under a factor-like model structure. Let $\mathbf{x}$ be a random vector from the factor-like model with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then $\mathbf{x}$ can be written as $\mathbf{x} = \boldsymbol{\mu} + \mathbf{P}\mathbf{z}$, where $\mathbf{P}$ is a $p \times m$ matrix for some $m \geq p$ such that $\mathbf{P}\mathbf{P}^\top = \boldsymbol{\Sigma}$, and $\mathbf{z} = (Z_1, \ldots, Z_m)^\top$ consisting of m independent and identically distributed random variables satisfying $\mathrm{E}(Z_j) = 0$, $\mathrm{Var}(Z_j) = 1$, and $\mathrm{E}(Z_j^4) = m_4 < \infty$, for $j = 1, \ldots, m$. This factor-like model is also known as independent component model. Bai and Saranadasa (1996) proposed a test for the two-sample problem under the factor-like model and Srivastava (2009) studied its one-sample version. The test statistic for one-sample problem is defined as

$$T_{BS} = \bar{\mathbf{x}}^\top \bar{\mathbf{x}} - \mathrm{tr}\mathbf{S}/n.$$

The test statistic T_{BS} can be regarded as unscaled distance $\bar{\mathbf{x}}^\top \bar{\mathbf{x}}$ with offset $\mathrm{tr}\mathbf{S}/n$. If $\mathbf{z}$ satisfies that $\mathrm{E}(\prod_{j=1}^m Z_j^{v_j})$ equals 0 when there is at least one $v_k = 1$ and equals 1 when there are two v_j's equal 2, whenever $\sum_{j=1}^m v_j = 4$, $p/n \to c > 0$ and $\lambda_{\max}(\boldsymbol{\Sigma}) = o(\sqrt{\mathrm{tr}\boldsymbol{\Sigma}^2})$, then under the local alternative $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = o(\mathrm{tr}\boldsymbol{\Sigma}^2/n)$,

$$n(n-1)\mathrm{Var}(T_{BS}) \to 2\mathrm{tr}(\boldsymbol{\Sigma}^2).$$

Under H_0, T_{BS} has mean $\mathbf{0}$. Therefore, the asymptotic null distribution is

$$\frac{T_{BS}}{\sqrt{2\mathrm{tr}(\boldsymbol{\Sigma}^2)/(n(n-1))}} \to N(0,1).$$

The power function under local alternative that $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} = o(\mathrm{tr}\boldsymbol{\Sigma}^2/n)$ is

$$\beta_{T_{BS}}(\boldsymbol{\mu}) = \Phi\left(-z_\alpha + \frac{n\|\boldsymbol{\mu}\|_2^2}{\sqrt{2\mathrm{tr}\boldsymbol{\Sigma}^2}}\right),$$

where $\Phi(\cdot)$ is the cdf of standard normal distribution. Bai and Saranadasa (1996)

also gave a consistent estimator of $\mathrm{tr}(\mathbf{\Sigma}^2)$,

$$\widehat{\mathrm{tr}(\mathbf{\Sigma}^2)} = \frac{(n-1)^2}{(n-2)(n+1)}(\mathrm{tr}\mathbf{S}^2 - (\mathrm{tr}\mathbf{S})^2/n).$$

Inspired by the observation that both the term $\sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i$ in calculating $\bar{\mathbf{x}}^\top \bar{\mathbf{x}}$ and the term $\mathrm{tr}\mathbf{S}$ impose a restricted condition that p and n should be of the same order, Chen and Qin (2010) proposed the following test statistic in which the term $\mathbf{x}_i^\top \mathbf{x}_i$ is removed,

$$T_{CQ} = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} \mathbf{x}_i^\top \mathbf{x}_j.$$

Chen and Qin (2010) showed that if $\mathrm{tr}(\mathbf{\Sigma}^4) = o(\mathrm{tr}^2(\mathbf{\Sigma}^2))$ and $\mathrm{E}(\prod_{k=1}^q Z_{l_k}^{\alpha_k}) = \prod_{k=1}^q \mathrm{E}(Z_{l_k}^{\alpha_k})$ for a positive integer q such that $\sum_{l=1}^q \alpha_l \leq 8$ and $l_1 \neq, \ldots, \neq l_q$, then under the local alternative that $\boldsymbol{\mu}^\top \mathbf{\Sigma}^{-1} \boldsymbol{\mu} = o(\mathrm{tr}\mathbf{\Sigma}^2/n)$, we have $n(n-1)\mathrm{Var}(T_{CQ}) \to 2\mathrm{tr}(\mathbf{\Sigma}^2)$. Under H_0, T_{CQ} has mean $\mathbf{0}$. Therefore, the asymptotic null distribution is

$$\frac{T_{CQ}}{\sqrt{2\mathrm{tr}(\mathbf{\Sigma}^2)/(n(n-1))}} \to N(0,1),$$

and the power function is

$$\beta_{T_{CQ}}(\boldsymbol{\mu}) = \Phi\left(-z_\alpha + \frac{n\|\boldsymbol{\mu}\|_2^2}{\sqrt{2\mathrm{tr}\mathbf{\Sigma}^2}}\right).$$

Note that T_{BS} and T_{CQ} share the same asymptotic distribution and power function. In fact, we can show T_{BS} and T_{CQ} take exactly the same form

$$\begin{aligned}
T_{BS} =& \bar{\mathbf{x}}^\top \bar{\mathbf{x}} - \mathrm{tr}\mathbf{S}/n \\
=& \frac{n}{n-1}\bar{\mathbf{x}}^\top \bar{\mathbf{x}} - \frac{1}{n(n-1)} \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i \\
=& \frac{1}{n(n-1)} \left(\sum_{i \neq j} \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i \right) - \frac{1}{n(n-1)} \sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i \\
=& \frac{1}{n(n-1)} \sum_{i \neq j}^n \mathbf{x}_i^\top \mathbf{x}_j = T_{CQ}.
\end{aligned}$$

The order of n and p is not explicatly controlled in T_{CQ}. Instead, $\mathrm{tr}(\mathbf{\Sigma}^4) = o(\mathrm{tr}^2(\mathbf{\Sigma}^2))$

is used to control the growth of p. The estimator for $\mathrm{tr}\boldsymbol{\Sigma}^2$ is adapted from Bai and Saranadasa (1996) by excluding the term $\sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i$,

$$\widehat{\mathrm{tr}\boldsymbol{\Sigma}^2} = \frac{\mathrm{tr}\left(\sum_{j\neq k}^n (\mathbf{x}_j - \bar{\mathbf{x}}_{(j,k)})\mathbf{x}_j^\top (\mathbf{x}_k - \bar{\mathbf{x}}_{(j,k)})\mathbf{x}_k^\top\right)}{n(n-1)},$$

where $\bar{\mathbf{x}}_{(j,k)}$ is the sample mean after excluding $\mathbf{x}_j$ and $\mathbf{x}_k$. Though T_{BS} and T_{CQ} have exactly the same form, the estimators of variance for T_{BS} and T_{CQ} are slightly different when performing the test.

T_{BS} and T_{CQ} are not invariant under different scales. To get rid of the unit effect, Srivastava and Du (2008) and Srivastava (2009) constructed test statistics by replacing $\mathbf{S}$ with diagonal matrix $\mathbf{D}$, where $\mathbf{D} = \mathrm{diag}(\mathbf{S})$, a diagonal matrix with elements being the diagonal elements of $\mathbf{S}$. The test statistic in Srivastava and Du (2008) is defined as

$$T_{SD} = n\bar{\mathbf{x}}^\top \mathbf{D}^{-1}\bar{\mathbf{x}} - (n-1)p/(n-3).$$

Assume that $n = O(p^\zeta)$ for some $\frac{1}{2} \leq \zeta \leq 1$, $0 < \lim_{p\to\infty} \mathrm{tr}\mathbf{R}_0^i/p < \infty$ for $i = 1,2,3,4$ and $\lim_{p\to\infty} \max_{1\leq i\leq p} \lambda_i/\sqrt{p} = 0$ where $\mathbf{R}_0 = \mathbf{D}_{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}_{\boldsymbol{\Sigma}}^{-1/2}$ with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_p$, and $\mathbf{D}_{\boldsymbol{\Sigma}}$ is the diagonal matrix with diagonal elements from the covariance matrix $\boldsymbol{\Sigma}$. Under H_0 and normality assumption, we have

$$\frac{T_{SD}}{\sqrt{2(\mathrm{tr}\mathbf{R}^2 - p^2/(n-1))c_{n,p}}} \to N(0,1),$$

where $\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$ is the sample correlation matrix and $c_{n,p}$ is an adjustment coefficient that approaches 1 in probability as n and p tend to ∞. The adjustment coefficient $c_{n,p}$ is needed to improve the convergence of T_{SD}. The authors suggest using

$$c_{n,p} = 1 + \mathrm{tr}(\mathbf{R}^2)/p^{3/2}.$$

Under local alternative that $\boldsymbol{\mu} = (n(n-1))^{-1/2}\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a constant vector, if for any p, $\boldsymbol{\delta}^\top \mathbf{D}_{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\delta}$ is bounded by a constant that does not depend on p, then the asymptotic power function $\beta_{T_{SD}}$ is

$$\beta_{T_{SD}}(\boldsymbol{\mu}) = \Phi\left(-z_\alpha + \frac{n\boldsymbol{\mu}^\top \mathbf{D}_{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}}{\sqrt{2\mathrm{tr}\mathbf{R}_0^2}}\right).$$

Srivastava (2009) removed the the adjustment coefficient $c_{n,p}$ and normality assumption in Srivastava and Du (2008). Under the same conditions except that ζ is relaxed to $0 < \zeta \leq 1$, Srivastava (2009) showed that $\mathrm{Var}(T_{SD}) \to 2\mathrm{tr}\mathbf{R}_0^2$. Therefore, under H_0

$$\frac{T_{SD}}{\sqrt{2\mathrm{tr}\mathbf{R}_0^2}} \to N(0,1).$$

A consistent estimator of $\mathrm{tr}\mathbf{R}_0^2$ is $\mathrm{tr}\mathbf{R}^2 - p^2/(n-1)$. The condition $\frac{1}{2} \leq \zeta \leq 1$ in Srivastava and Du (2008) guarantees that the adjustment coefficient converge to 1. Since adjustment coefficient is removed in Srivastava (2009) and thus the condition can be relaxed to $0 < \zeta \leq 1$.

2.2.3 Maximum-type Test

If the null hypothesis $H_0 : \boldsymbol{\mu} \neq \mathbf{0}$ is rejected, then there is at least one element in $\boldsymbol{\mu}$ is not 0. Intuitively, one can construct a test for each of the p variables and hopefully at least one of the p tests is rejected if the alternative is true. Based on this observation, the maximum value of the p individual tests can be used as the test statistic. This type of tests is particularly powerful against sparse alternatives.

Cai et al. (2014) introduced a test that is based on a linear transformation of the data by the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ which incorporates the correlations among the variables. Given that the precision matrix $\boldsymbol{\Omega} = (\omega_{ij})_{p\times p}$ is known, the test statistic is defined as

$$T_{CLX} = n \max_{1\leq j\leq p} \frac{\bar{X}_j^2}{\omega_{jj}}. \tag{2.17}$$

Assume that $\boldsymbol{\Omega}$ is sparse, Cai et al. (2014) used the CLIME estimator (Cai et al. 2011) to estimate $\boldsymbol{\Omega}$. Let $\widehat{\boldsymbol{\Omega}}_1 = (\widehat{\omega}_{ij}^1)_{p\times p}$ be a solution to the following optimization problem

$$\min \|\boldsymbol{\Omega}\|_1 \text{ subjection to } \|\mathbf{S}\boldsymbol{\Omega} - \mathbf{I}\|_\infty < \lambda_n,$$

where $\|\boldsymbol{\Omega}\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\omega_{ij}|$, $\|\boldsymbol{\Omega}\|_\infty = \max_{1\leq i,j\leq p} |\omega_{ij}|$, and $\lambda_n = C\sqrt{\log p/n}$ for some sufficiently large constant C. In practice, λ_n can be chosen by cross validation, see Cai et al. (2011) for more details. To ensure that the resulting estimator of the precision matrix is symmetric, the final estimator of $\boldsymbol{\Omega}$ is defined to be $\widehat{\boldsymbol{\Omega}} = (\widehat{\omega}_{ij})_{p\times p}$ where

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

This estimator $\widehat{\mathbf{\Omega}}$ is called the CLIME estimator and can be implemented by linear programming. In practice, ω_{jj} in (2.17) is replaced by $\widehat{\omega}_{jj}$. Let $\mathbf{D}_\sigma = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$ and $\mathbf{D}_\omega = \text{diag}(\omega_{11}, \ldots, \omega_{pp})$, where σ_{jj} and ω_{jj} are the diagonal entries of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ respectively. The correlation matrix of $\mathbf{x}$ is $\mathbf{\Gamma} = (\gamma_{ij})_{p\times p} = \mathbf{D}_\sigma^{-1/2}\mathbf{\Sigma}\mathbf{D}_\sigma^{-1/2}$ and the correlation matrix of $\mathbf{\Omega}\mathbf{x}$ is $\mathbf{R} = (r_{ij})_{p\times p} = \mathbf{D}_\omega^{-1/2}\mathbf{\Omega}\mathbf{D}_\omega^{-1/2}$. To obtain the limiting distribution under H_0, Cai et al. (2014) imposed the following conditions

(C1) $c^{-1} \leq \lambda_{\min}(\mathbf{\Sigma}) < \lambda_{\max}(\mathbf{\Sigma}) \leq c$ for some constant c.

(C2) $\max_{1\leq i,j\leq p} |\gamma_{ij}| \leq r_2$ for some constant $0 < r_2 < 1$.

Suppose conditions (C1) and (C2) hold under H_0, Cai et al. (2014) showed that for every $x \in \mathbb{R}$,

$$P(T_{CLX} - 2\log p + 2\log\log p \leq x) \to \exp\left(-\frac{1}{\sqrt{\pi}}\exp(-\frac{x}{2})\right) \text{ as } p \to \infty. \quad (2.18)$$

The right hand side in (2.18) is known as the type I extreme value distribution or Gumbel distribution. Therefore, the null hypothesis is rejected at level α when

$$T_{CLX} > 2\log p - 2\log\log p + q_\alpha,$$

where q_α is the $1-\alpha$ quantile of the type I extreme value distribution, i.e.,

$$q_\alpha = -\log(\pi) - 2\log\log(1-\alpha)^{-1}.$$

Chen et al. (2014) proposed a test which removes components that are estimated to be zero via thresholding. The motivation is that zero components are expected to contribute little to the squared sample mean and those smaller than a given threshold can be ignored. The test statistic with index s is defined as

$$T_{CLZ}(s) = \sum_{j=1}^{p}\left\{\frac{n\bar{X}_j^2}{\sigma_{ii}} - 1\right\} I\left\{\frac{n\bar{X}_j^2}{\sigma_{ii}} > \lambda_p(s)\right\},$$

where the threshold level is set to be $\lambda_p(s) = 2s\log p$ for some $s \in (0,1)$. If

$\log p = o(n^{1/3})$, Chen et al. (2014) showed that for any $s \in (0,1)$,

$$\sigma_{CLZ}^{-1}(s)(T_{CLZ}(s) - \mu_{CLZ}(s)) \to N(0,1),$$

where $\mu_{CLZ}(s)$ and $\sigma_{CLZ}^2(s)$ are the expectation and variance of $T_{CLZ}(s)$. Therefore, an asymptotic level α test rejects H_0 if

$$T_{CLZ}(s) > z_\alpha \widehat{\sigma}_{CLZ,0}(s) + \widehat{\mu}_{CLZ,0}(s),$$

where $\widehat{\sigma}_{CLZ,0}(s)$ and $\widehat{\mu}_{CLZ,0}(s)$ are the estimators of $\sigma_{CLZ}(s)$ and $\mu_{CLZ}(s)$ under H_0, and z_α is the upper α quantile of $N(0,1)$. Chen et al. (2014) further proposed the multi-level thresholding statistic, which is defined as

$$T_{CLZ} = \max_{s \in (0,1-\eta)} \{T_{CLZ}(s) - \widehat{\mu}_{CLZ,0}(s)\}/\widehat{\sigma}_{CLZ,0}(s),$$

for some $\eta \in (0,1)$. The asymptotic null distribution of T_{CLZ} is the Gumbel distribution. Under H_0,

$$P\{a(\log p)T_{CLZ} - b(\log p, \eta) \le x\} \to \exp\{-e^{-x}\},$$

where $a(y) = (2\log y)^{1/2}$ and $b(y,\eta) = 2\log y + \log\log y/2 - \log(\sqrt{\pi}/(1-\eta))$. Therefore, the multi-level thresholding test of asymptotic level α rejects H_0 if

$$T_{CLZ} > (q_\alpha + b(\log y, \eta))/a(\log p),$$

where q_α is the upper α quantile of the Gumbel distribution. Due to the slow convergence to the asymptotic null distribution, Chen et al. (2014) proposed to use the parametric bootstrap to compute its p-value.

2.2.4 Projecton test

The idea of projection test is to project the high-dimensional vector $\mathbf{x}_i$ onto a space of low dimension and then traditional methods such as t-test or Hotelling's T^2 can be applied. Let $\mathbf{P}$ be a $p \times k$ matrix with $k \ll n$ and we can project the p-dimensional vector $\mathbf{x}_i$ to a k-dimensional space by left-multiplying the matrix $\mathbf{P}^\top$. More specifically, define $\mathbf{y}_i = \mathbf{P}^\top \mathbf{x}_i, i = 1, \ldots, n$, and thus $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^k$ are

independent and identically distributed with mean $\mathbf{P}^\top \boldsymbol{\mu}$ and covariance matrix $\mathbf{P}^\top \boldsymbol{\Sigma} \mathbf{P}$. The Hotelling's $T^2_{\mathbf{P}}$ after projection is defined to be

$$T^2_{\mathbf{P}} = n\bar{\mathbf{x}}^\top \mathbf{P}(\mathbf{P}^\top \widehat{\boldsymbol{\Sigma}} \mathbf{P})^{-1} \mathbf{P}^\top \bar{\mathbf{x}},$$

which is equivalent to the Hotelling's T^2 test based on $\mathbf{y}_1, \ldots, \mathbf{y}_n$.

Several methods have been proposed to determine the projection matrix $\mathbf{P}$. Lauter (1996) considered a test using the linear score $\mathbf{z} = (Z_1, \ldots, Z_n)^\top = \mathbf{X}\mathbf{d}$, where $\mathbf{d}$ is a $p \times 1$ projection vector depending on $\mathbf{X}$ only through $\mathbf{X}^\top \mathbf{X}$ and $\mathbf{d} \neq \mathbf{0}$ with probability 1. Then one can perform the one-sample t-test based on $Z_1, \ldots, Z_n$. Lauter (1996) also proposed two different ways to obtain the projection vector $\mathbf{d}$. For example, $\mathbf{d}$ can take the form of

$$\mathbf{d} = (\mathrm{diag}(\mathbf{X}^\top \mathbf{X}))^{-1/2},$$

or be the eigenvector corresponding to the largest eigenvalue $\lambda_{\max}$ for the following eigenvalue problem

$$(\mathbf{X}^\top \mathbf{X})\mathbf{d} = \mathrm{diag}(\mathbf{X}^\top \mathbf{X})\mathbf{d}\lambda_{\max}.$$

Lopes et al. (2011) proposed a random projection test where the entries in $\mathbf{P}$ are randomly drawn from the standard normal distribution. This random projection test is an exact test if $\mathbf{x}_i$'s are normally distributed. Instead of using random projection, Li et al. (2015) proposed a projection test using the optimal projection direction. Li et al. (2015) showed that the optimal choice of k in $\mathbf{P}$ is 1 and the optimal projection direction is $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ in the sense that the power of the test $T^2_{\mathbf{P}}$ is maximized. Let $y_i = \boldsymbol{\theta}^\top \mathbf{x}_i, i = 1, \ldots, n$. The projection test statistic is

$$T^2_{\boldsymbol{\theta}} = n\bar{\mathbf{x}}^\top \boldsymbol{\theta}(\boldsymbol{\theta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\theta})^{-1} \boldsymbol{\theta}^\top \bar{\mathbf{x}},$$

which follows $F_{1,n-1}$ distribution under H_0. It is equivalent to a one-sample t test based on $y_1, \ldots, y_n$. In order to control the type I error, Li et al. (2015) proposed a data-splitting strategy to estimate the optimal direction and obtained an exact t-test. They partition the random sample into two separate sets: $\mathcal{D}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}\}$ and $\mathcal{D}_2 = \{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n\}$. They use $\mathcal{D}_1$ to estimate the direction $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and use $\mathcal{D}_2$ to construct the test statistic $T^2_{\boldsymbol{\theta}}$. To estimate $\boldsymbol{\theta}$, they proposed a ridge-type estimator $\widehat{\boldsymbol{\theta}} = (\mathbf{S}_1 + \lambda \mathbf{D}_1)^{-1}\bar{\mathbf{x}}_1$, where $\bar{\mathbf{x}}_1$ and $\widehat{\boldsymbol{\Sigma}}_1$ are the sample mean vector and

the sample covariance matrix computed from $\mathcal{D}_1$ and $\mathbf{D}_1 = \text{diag}(\widehat{\boldsymbol{\Sigma}}_1)$. The test statistic $T^2_{\widehat{\boldsymbol{\theta}}}$ is constructed using $\widehat{\boldsymbol{\theta}}^\top \mathbf{x}_{n_1+1}, \ldots, \widehat{\boldsymbol{\theta}}^\top \mathbf{x}_n$. Li et al. (2015) also derived the asymptotic power function of the projection test $T^2_{\widehat{\boldsymbol{\theta}}}$ under the assumption that $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ in probability. However, there is no guarantee that the ridge-type estimator is consistent. In order to obtain a better estimation of $\boldsymbol{\theta}$, we assume the optimal projection direction is sparse. Under the sparsity assumption, we estimate $\boldsymbol{\theta}$ using regularized quadratic programming and it can be shown that the resulting estimator is consistent.

2.3 Feature screening of high-dimensional data

2.3.1 An overview

With the advent of modern technology for data collection, ultra-high dimensional datasets are widely encountered in machine learning, statistics, genomics, medicine, finance, marketing, etc. The ultra-high dimensionality causes challenges in both computation and methodology. Scalability is the major challenge to ultra-high dimensional data analysis. Other issues such as high collinearity, spurious correlation, and noise accumulation (Fan and Lv 2008, 2010) bring in additional challenges. Therefore, variable selection and feature screening have been a fundamental problem in the analysis of ultra-high dimensional data. Over the past two decades, a large amount of variable selection approaches based on regularized M-estimation have been developed. These approaches include the Lasso (Tibshirani 1996), the SCAD (Fan and Li 2001), the Dantzig selector (Candes and Tao 2007), and the MCP (Zhang 2010), among others. However, these regularization methods may not perform well for ultra-high dimensional data due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability (Fan et al. 2009). To improve the performance of regularized methods, a two stage approach can be applied. In the first stage, we reduce the number of features from a very large scale to a moderate size in a computationally fast way. In the second stage, we further implement refined variable selection algorithms such as regularized methods to the selected features from the first stage. Ideally, we select all the important features and may allow a few unimportant features entering model in the first stage. The first stage is referred to as the feature screening stage. We only focus on the

feature screening stage in this chapter.

Suppose we have p features $X_1, \ldots, X_p$ in the feature space and denote the index set of true important variables by $\mathcal{M}_\star$. The definition of $\mathcal{M}_\star$ may vary across different models. For example, in a parametric model associated with true parameters $\boldsymbol{\beta}^\star = (\beta_1^\star, \ldots, \beta_p^\star)^\top$, $\mathcal{M}_\star$ is typically defined to be

$$\mathcal{M}_\star = \{1 \leq j \leq p : \beta_j^\star \neq 0\}.$$

Our goal in the feature screening stage is to select a submodel $\widehat{\mathcal{M}} \subset \{1, \ldots, p\}$ such that $\mathcal{M}_\star \subset \widehat{\mathcal{M}}$ with high probability. This is referred to as the sure screening property. The sure screening property ensures that all the important features are included in the selected submodel with probability approaching to 1 as the sample size goes to infinity.

The most common feature screening method is the marginal screening, which uses the marginal utility of individual feature to rank the importance of all features. More specifically, the marginal feature screening procedure assigns an index, say $\widehat{\omega}_j$, to each of features X_j. This index $\widehat{\omega}_j$ measures the dependence between the jth feature and the response variable. Then we can rank all features according to the index and include the top important features in the submodel. For example, in the setting of linear regression, the index $\widehat{\omega}_j$ is chosen to be the absolute value of marginal Pearson correlation between the jth feature and the response (Fan and Lv 2008). Features with larger values of $\widehat{\omega}_j$ are more relevant to the response and thus have higher rankings. We rank all the features according to $\widehat{\omega}_j$ and include the top d_n features in the submodel,

$$\widehat{\mathcal{M}}_{d_n} = \{1 \leq j \leq p : \widehat{\omega}_j \text{ is among the top } d_n \text{ ones}\},$$

where d_n is some pre-specified threshold. Note that the marginal feature screening procedure only uses the information of jth feature and the response without looking at all other features and thus it can be carried out in a very efficient way. A large amount of literature have studied the the sure screening property of various marginal feature screening methods, see Fan and Lv (2008), Fan et al. (2009, 2011), Li et al. (2012) and Fan, Ma and Dai (2014).

As pointed out in Fan and Lv (2008), the marginal feature screening procedure may suffer from the following two issues:

1. Some unimportant features that are highly correlated with important features can have higher rankings than other important features that are relatively weakly related to the response.

2. An important feature that is marginally independent but jointly dependent on the response tends to have lower ranking.

The first issue says that the marginal feature screening has the chance to include some unimportant features in the submodel. This is not too bad from a feature screening's perspective. The second issue is a bigger issue, which says that the marginal feature screening may fail to include the important feature if it is marginally independent of the response. It is well known that absence of any important feature would lead to a biased estimation. To overcome the two aforementioned issues, one can use an iterative feature screening procedure, which iteratively carries out the marginal screening procedure. This iterative procedure was first introduced by Fan and Lv (2008) and can be viewed as an extension of the marginal feature screening. At the kth iteration, we apply marginal feature screening to the features survived from the previous step. Let $\widehat{\mathcal{M}}_k$ be the selected index set of important variables at the kth iteration and the final selected index set of important variables is given by $\widehat{\mathcal{M}} = \widehat{\mathcal{M}}_1 \cup \widehat{\mathcal{M}}_2 \cup \ldots$, the union of all selected index sets. For example, Fan and Lv (2008) used the residual as the new response and iteratively applied marginal feature screening based on Pearson correlation, where the residual is obtained from the linear regression with features selected from the previous step. The iterative feature screening can significantly improve the simple marginal screening, but it can also be much more computationally expensive. Another approach to improve the marginal screening is the sure joint screening, which utilizes all the features. The joint screening approach approximates the objective function by its Taylor's expansion (Xu and Chen 2014, Yang et al. 2016) such that the optimization problem can be solved in a fast manner. In many examples, one can obtain a closed form for each update.

2.3.2 Linear model and generalized linear model

Let us consider the linear regression model,

$$\mathbf{y} = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta} + \epsilon, \tag{2.19}$$

where β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional regression coefficient vector, and ϵ is the error term. In the ultra-high dimensional setting, the true regression coefficient vector $\boldsymbol{\beta}^\star = (\beta_1^\star, \ldots, \beta_p^\star)^\top$ is assumed to be sparse, meaning most of the coefficients $\beta_j^\star$ are 0. The index set of the true model is defined as

$$\mathcal{M}_\star = \{1 \leq j \leq p : \beta_j^\star \neq 0\}.$$

The features with indices in the support $\mathcal{M}_\star$ are called important features. To select the important features, Fan and Lv (2008) suggested ranking all features according to the marginal Pearson correlation coefficient between each feature and the response and select the top features which have strong correlation with the response. For a pre-specified value $\nu_n (0 < \nu_n < 1)$, the index set of selected features is given by

$$\widehat{\mathcal{M}}_{\nu_n} = \{1 \leq j \leq p : |\widehat{\text{corr}}(\mathbf{x}_{(j)}, \mathbf{y})| \text{ is among the top } \lfloor \nu_n n \rfloor \text{ largest ones}\},$$

where $\mathbf{x}_{(j)}$ is the jth column of $\mathbf{X}$, $\widehat{\text{corr}}$ denotes the sample Pearson correlation, and $\lfloor \nu_n n \rfloor$ is the integer part of $\nu_n n$. This procedure achieves the goal of feature screening since it reduces the ultra-high dimensionality down to a relatively moderate scale $\lfloor \nu_n n \rfloor$. This procedure is referred to as the sure independence screening (SIS). Then appropriate regularized methods such as the Lasso, the SCAD and the Dantzig selector can be further applied to the selected important features. This feature screening procedure is based on Pearson correlation and can be carried out in a extremely simple way at very low computational cost. In addition to the computational advantage, Fan and Lv (2008) also showed that under fairly general conditions, the SIS has the sure screening property. It can reduce from exponentially growing dimension p down to a relatively small scale $d_n = \lfloor \nu_n n \rfloor = O(n^{1-\theta}) < n$, while include all important features in the submodel with high probability. In practice, one can set $d_n = \lfloor n/\log n \rfloor$ or $n - 1$ as discussed in Fan and Lv (2008).

Since marginal Pearson correlation is employed to rank features, the SIS may suffer from the potential issues with marginal screening. On one hand, the SIS may fail to select an important feature when it is jointly correlated but marginally uncorrelated with the response. On the other hand, the SIS tends to select unimportant features which are jointly uncorrelated but highly marginally correlated with the response. To address these issues, Fan and Lv (2008) also introduced

an iterative SIS procedure (ISIS) by iteratively replacing the response with the residuals obtained from the regression on the selected features survived the previous step.

A natural extension of the SIS is to apply the feature screening procedure to the generalized linear model. Assume that the response Y is from an exponential family with the following canonical form

$$f_Y(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\},$$

for some known functions $b(\cdot)$, $c(\cdot)$ and unknown parameter θ. Consider the following generalized linear model

$$\mathrm{E}(Y|\mathbf{x}) = g^{-1}(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}), \tag{2.20}$$

where $g(\cdot)$ is the link function, β_0 is a unknown scalar, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a p-dimensional unknown vector. The linear regression model (2.19) is a special case of (2.20) by taking $g(\mu) = \mu$. Without loss of generality, we assume that all the features are standardized to have mean zero and standard deviation one. Fan and Song (2010) proposed a feature screening procedure for (2.20) by ranking the maximum marginal likelihood estimator (MMLE). For each $1 \le j \le p$, the MMLE $\widehat{\boldsymbol{\beta}}_j^M$ is defined as

$$\widehat{\boldsymbol{\beta}}_j^M = (\widehat{\beta}_{j0}^M, \widehat{\beta}_{j1}^M)^\top = \underset{\beta_{j0},\beta_{j1}}{\arg\min} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \beta_{j0} + \beta_{j1}X_{ij}), \tag{2.21}$$

where $\ell(y,\theta) = -y\theta + b(\theta) - c(y)$ is the negative log-likelihood function. The minimization problem (2.21) can be rapidly computed and its implementation is robust since it only involves two parameters. Such a feature screening procedure ranks features according to their magnitude of marginal regression coefficients. The set of important features is defined as

$$\widehat{\mathcal{M}}_{\nu_n} = \{1 \le j \le p : |\widehat{\beta}_{j1}^M| > \nu_n\},$$

where ν_n is some pre-specified threshold. As a result, we dramatically decrease the dimension from p to a moderate size by choosing ν_n properly. In the linear regression setting, the MMLE ranking is equivalent to the marginal correlation

ranking. However, the MMLE screening does not rely on the normality assumption and can be more easily to applied to other models (Fan and Song 2010).

Fan et al. (2009) extended SIS and ISIS to a general pseudo-likelihood framework in which the aim is to find the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ that minimizes an objective function of the form

$$Q(\beta_0, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i),$$

This framework includes a lot of important applications such as

1. **Generalized linear model**: All generalized linear models, including the logistic regression and the Poisson log-linear model, fit very naturally into the framework.

2. **Classification**: Some common approaches to classification assume the response takes values in $\{-1, 1\}$ also fit the framework. For instance, support vector machine (Vapnik 2013) uses the hinge loss function $\ell(Y_i, \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) = (1 - Y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}))_+$, while the boosting algorithm AdaBoost (Freund and Schapire 1997) uses $\ell(Y_i, \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) = \exp\{-Y_i(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta})\}$.

3. **Robust fitting**: Instead of the conventional least squares loss function, one may prefer a robust loss function such as the ℓ_1 loss $\ell(Y_i, \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) = |Y_i - \beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta}|$ or the Huber loss (Huber 1964), which also fits into the framework.

Fan et al. (2009) suggested using the marginal utility to rank the features. The marginal utility of the jth feature X_j is quantified by

$$L_j = \min_{\beta_0, \beta_j} n^{-1} \sum_{i=1}^{n} \ell(Y_i, \beta_0 + X_{ij}\beta_j).$$

The idea is to compute the vector of marginal utilities $\mathbf{L} = (L_1, \ldots, L_p)^\top$ and rank the features according to the marginal utilities: the smaller L_j is, the more important X_j is. Note that in order to compute L_j , we only need to fit a model with two parameters, β_0 and β_j, so computing the vector $\mathbf{L}$ can be done very quickly and stably, even for an ultra-high dimensional problem. The feature X_j is

selected if the corresponding utility L_j is among the d_n smallest components of $\mathbf{L}$. Typically, we may take $d_n = \lfloor n/\log n \rfloor$. When d_n is large enough, all important features would be selected with high probability. Fan et al. (2009) also proposed an iterative feature screening procedure, which consists of the following steps.

Step 1. Compute the vector of marginal utilities $\mathbf{L} = (L_1, \ldots, L_p)^\top$ and select the set $\widehat{\mathcal{A}}_1 = \{1 \le j \le p : L_j \text{ is among the first } k_1 \text{ smallest ones}\}$. Then apply a penalized (pseudo)-likelihood, such as the Lasso and the SCAD, to select a subset $\widehat{\mathcal{M}}$.

Step 2. For each $j \in \{1, \ldots, p\}/\widehat{\mathcal{M}}$, compute

$$L_j^{(2)} = \min_{\beta_0, \beta_j, \boldsymbol{\beta}_{\widehat{\mathcal{M}}}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}_{i,\widehat{\mathcal{M}}}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}} + X_{ij}\beta_j), \tag{2.22}$$

where $\mathbf{x}_{i,\widehat{\mathcal{M}}}$ denotes the sub-vector of $\mathbf{x}_i$ consisting of those elements in $\widehat{\mathcal{M}}$. Then select the set

$$\widehat{\mathcal{A}}_2 = \{j \in \{1, \ldots, p\}/\widehat{\mathcal{M}} : L_j^{(2)} \text{ is among the fist } k_2 \text{ smallest ones}\}.$$

Step 3. Use penalized likelihood to the features in set $\widehat{\mathcal{M}} \cup \widehat{\mathcal{A}}_2$,

$$\widehat{\boldsymbol{\beta}}_2 = \arg\min_{\beta_0, \boldsymbol{\beta}_{\widehat{\mathcal{A}}_2}, \boldsymbol{\beta}_{\widehat{\mathcal{M}}}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, \beta_0 + \mathbf{x}_{i,\widehat{\mathcal{M}}}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}} + \mathbf{x}_{i,\widehat{\mathcal{A}}_2}^\top \boldsymbol{\beta}_{\widehat{\mathcal{A}}_2}) + \sum_{j \in \widehat{\mathcal{M}} \cup \widehat{\mathcal{A}}_2} p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is some penalty function. The indices of $\widehat{\boldsymbol{\beta}}_2$ that are non-zero yield a new estimated set $\widehat{\mathcal{M}}$.

Step 4. Repeat Step 2 and Step 3 until $|\widehat{\mathcal{M}}| = d_n$.

Note that $L_j^{(2)}$ can be interpreted as the additional contribution of feature X_j given the existence of features in $\widehat{\mathcal{M}}$. The optimization problem in Step 2 is a low-dimensional problem which can be solved easily. An alternative approach of Step 2 is to substitute the fitted value $\widehat{\boldsymbol{\beta}}_{\widehat{\mathcal{M}}_1}$ from the Step 1 into (2.22). Then the optimization in (2.22) only involves two parameters and is exactly an extension of Fan and Lv (2008). To this end, let $r_i = Y_i - \mathbf{x}_{i,\widehat{\mathcal{M}}}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}}$ denote the residual from the

previous step and we choose the square loss function, then

$$\ell(Y_i, \beta_0 + \mathbf{x}_{i,\widehat{\mathcal{M}}}^\top \boldsymbol{\beta}_{\widehat{\mathcal{M}}} + X_{ij}\beta_j) = (r_i - \beta_0 - \beta_j X_{ij})^2.$$

Without explicit definition of residuals, the idea of additional contribution can be applied to a much more general statistical framework.

2.3.3 Nonparametric regression model

Fan et al. (2011) proposed a nonparametric independence screening (NIS) for ultra-high dimensional additive model of the following form,

$$Y = \sum_{j=1}^{p} m_j(X_j) + \epsilon, \tag{2.23}$$

where $m_j(X_j)$ is assumed to have zero mean for identifiability. The index set of the true important features is defined as

$$\mathcal{M}_\star = \{1 \le j \le p : \mathrm{E}m_j^2(X_j) > 0\}.$$

To identify the important features in model (2.23), Fan et al. (2011) considered the following p marginal nonparametric regression problems

$$\min_{f_j \in L_2(P)} \mathrm{E}(Y - f_j(X_j))^2, \tag{2.24}$$

where P denotes the joint distribution of $(\mathbf{x}, Y)$ and $L_2(P)$ is the family of square integrable functions under the measure P. The minimizer of (2.24) is $f_j = \mathrm{E}(Y|X_j)$, which can be used as a population level marginal utility. With a random sample $\{(\mathbf{x}_i, Y_i)\}, i = 1, \ldots, n,\ f_j(x)$ can be estimated by a set of B-spline basis. Let $\mathbf{B}(x) = (B_1(x), \ldots, B_L(x))^\top$ be the B-spline basis and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jL})^\top$ be the corresponding coefficients. Consider the following least squares,

$$\widehat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j} n^{-1} \sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}_j^\top \mathbf{B}(X_{ij}))^2.$$

Thus $f_j(x)$ can be estimated by $\widehat{f}_j(x) = \widehat{\boldsymbol{\beta}}_j^\top \mathbf{B}(x)$. The index set of selected submodel

is given by

$$\widehat{\mathcal{M}}_{\nu_n} = \{1 \le j \le p : \|\widehat{f}_j\|_n^2 \ge \nu_n\},$$

where $\|\widehat{f}_j\|_n^2 = n^{-1}\sum_{i=1}^n \widehat{f}_j(X_{ij})^2$ and ν_n is some prespecified threshold. This NIS procedure enjoys the sure screening property. The larger the minimum signal level or the smaller the number of basis functions, the higher the dimensionality that the NIS can handle. However, the number of basis functions cannot be too small since the approximation error would be too large if we only use a small number of basis functions.

Varying coefficient model is another important nonparametric statistical model that allows us to examine how the effects of features vary with some exposure variable. Consider the following varying coefficient model,

$$Y = \sum_{j=1}^{p} \beta_j(U) X_j + \epsilon,$$

where U is some observable exposure variable and the coefficient $\beta_j(\cdot)$ is a smooth function of variable U. The index set of true important features is defined as

$$\mathcal{M}_\star = \{1 \le j \le p : \mathrm{E}(\beta_j^2(U)) > 0\}$$

with model size $s = |\mathcal{M}_\star|$. For each feature $X_j, j = 1, \ldots, p$, Fan, Ma and Dai (2014) considered the following marginal regression

$$\min_{a_j, b_j} \mathrm{E}[(Y - a_j - b_j X_j)^2 | U]. \tag{2.25}$$

Let $a_j(U)$ and $b_j(U)$ be the solution to (2.25). The marginal contribution of X_j for the response can be characterized by

$$\omega_j = \|a_j(U) + b_j(U) X_j)\|^2 - \|a_0(U)\|^2, \tag{2.26}$$

where $a_0(U) = E[Y|U]$ and $\|f\|^2 = \mathrm{E} f^2$. By some algebra, it can be seen that

$$\omega_j = \mathrm{E}\left[\frac{(\mathrm{Cov}[X_j, Y|U])^2}{\mathrm{Var}[X_j|U]}\right].$$

The marginal utility $\omega_j = 0$ if and only if $\mathrm{Cov}[X_j, Y|U] = 0$. Given a random

sample $\{(\mathbf{x}_i, Y_i, U_i)\}, i = 1, \ldots, n$, we can estimate $a_j(U), b_j(U)$ and $a_0(U)$ using B-spline. Let $\mathbf{B}(U) = (B_1(U), \ldots, B_L(U))^\top$ be the B-spline basis and consider the marginal regression problem

$$(\widehat{\boldsymbol{\eta}}_j, \widehat{\boldsymbol{\theta}}_j) = \min_{\boldsymbol{\eta}_j, \boldsymbol{\theta}_j} n^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{B}(U_i)^\top \boldsymbol{\eta}_j - \mathbf{B}(U_i)^\top \boldsymbol{\theta}_j X_{ij})^2,$$

$$\widehat{\boldsymbol{\eta}}_0 = \min_{\boldsymbol{\eta}_0} n^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{B}(U_i)^\top \boldsymbol{\eta}_0)^2,$$

where $\boldsymbol{\eta}_0 = (\eta_{01}, \ldots, \eta_{0L})^\top$, $\boldsymbol{\eta}_j = (\eta_{j1}, \ldots, \eta_{jL})^\top$, and $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jL})^\top$. Thus $\widehat{a}_j(U), \widehat{b}_j(U)$ and $\widehat{a}_0(U)$ can be estimated by

$$\widehat{a}_j(U) = \mathbf{B}(U)^\top \widehat{\boldsymbol{\eta}}_j, \widehat{b}_j(U) = \mathbf{B}(U)^\top \widehat{\boldsymbol{\theta}}_j, \text{ and } \widehat{a}_0(U) = \mathbf{B}(U)^\top \widehat{\boldsymbol{\eta}}_0.$$

The sample marginal utility for screening is

$$\widehat{\omega}_j = \|\widehat{a}_j(U) + \widehat{b}_j(U)\|_n^2 - \|\widehat{a}_0(U)\|_n^2.$$

The submodel is selected by $\widehat{\mathcal{M}}_{\nu_n} = \{1 \leq j \leq p : \widehat{\omega}_j \geq \nu_n\}$. Instead of using the marginal utility ω_j in (2.26) to rank the features, Liu et al. (2014) proposed a screening procedure based on conditional correlation for varying-coefficient model. Conditioning on U, the conditional correlation between X_j and Y is defined as the conditional Pearson correlation

$$\rho(X_j, Y|U) = \frac{\text{cov}(X_j, Y|U)}{\sqrt{\text{cov}(X_j, X_j|U)\text{cov}(Y, Y|U)}}.$$

$\text{E}[\rho^2(X_j, Y|U)]$ can be used as a population level marginal utility to evaluate the importance of X_j and it can be estimated by the kernel regression (Liu et al. 2014). The features with high conditional correlations will be included in the selected submodel.

2.3.4 Model free feature screening

In previous sections, we have discussed model-based feature screening procedures for ultra-high dimensional data, which require us to specify the underlying true model structure. However, it is quite challenging to correctly specify the model

structure on the regression function in high-dimensional modeling. In practice, one may do not know what model to use until the dimensionality of feature space is reduced to a moderate size. Therefore, model free feature screening is necessary for high-dimensional modeling. In this section, we review several model free feature screening procedures.

Recall that in parametric modeling, the index set of true important features $\mathcal{M}_\star$ is defined as the indices of nonzero elements in $\boldsymbol{\beta}^\star$. Since we do not assume any underlying model, there is no such true parameter $\boldsymbol{\beta}^\star$ and thus we need to redefine the true index set of important features $\mathcal{M}_\star$. Let Y be the response variable, which can be univariate or multivariate and let $\mathbf{x} = (X_1, \ldots, X_p)^\top$ be the p-dimensional covariate vector. Define the index set of important features as

$$\mathcal{M}_\star = \{1 \leq j \leq p : F(y|\mathbf{x}) \text{ functionally depends on } X_j \text{ for any } y \in \Psi_y\},$$

where $F(y|\mathbf{x}) = \Pr(Y < y|\mathbf{x})$ is the conditional distribution function of Y given $\mathbf{x}$ and Ψ_y is the support of Y. This indicates that conditioning on $\mathbf{x}_{\mathcal{M}_\star}$, Y is statistically independent of $\mathbf{x}_{\mathcal{M}_\star^c}$, where $\mathbf{x}_{\mathcal{M}_\star}$ is a s-dimensional vector consisting of all X_j with $j \in \mathcal{M}_\star$.

Zhu et al. (2011) considered a general model framework under which $F(y|\mathbf{x})$ depends on $\mathbf{x}$ only through $\mathbf{B}^\top \mathbf{x}_{\mathcal{M}_\star}$, where $\mathbf{B}$ is a $s \times K$ parameter matrix. In other words, we assume $F(y|\mathbf{x}) = F(y|\mathbf{B}^\top \mathbf{x}_{\mathcal{M}_\star})$. Note that $\mathbf{B}$ may not be identifiable. What is identifiable is the space spanned by the columns of $\mathbf{B}$. However, the identifiability of $\mathbf{B}$ is of no concern here because our primary goal is to identify important features rather than estimating $\mathbf{B}$. This general framework covers a wide range of existing models including the linear regression model, the generalized linear models, the partially linear model (Hardle et al. 2012), the single-index model (Hardle et al. 1993), and the partially linear single-index model (Carroll et al. 1997), etc. It also includes the transformation regression model with a general transformation $h(Y)$.

Zhu et al. (2011) proposed a unified screening procedure for this general framework. Without loss of generality, assume $\mathrm{E}(X_j) = 0$ and $\mathrm{Var}(X_j) = 1$. Define $\boldsymbol{\Omega}(y) = \mathrm{E}[\mathbf{x}F(y|\mathbf{x})]$. It then follows by the law of iterated expectations that $\boldsymbol{\Omega}(y) = \mathrm{E}[\mathbf{x}\mathrm{E}(\mathbf{1}(Y < y|\mathbf{x}))|\mathbf{x}] = \mathrm{cov}(\mathbf{x}, \mathbf{1}(Y < y))$. Let $\Omega_j(y)$ be the jth element of

$\mathbf{\Omega}(y)$ and define

$$\omega_j(y) = \mathrm{E}(\Omega_j^2(y)),\ j = 1, \ldots, p.$$

ω_j can be regarded as the marginal utility which measures the dependence between feature X_j and the response Y. If X_j and Y are independent, so are X_j and $\mathbf{1}(Y < y)$. Consequently $\Omega_j(y) = 0$ for all $y \in \Psi_y$ and $\omega_j = 0$. On the other hand, if X_j and Y are dependent, then there exists some $y \in \Psi_y$ such that $\Omega_j(y) \neq 0$, and hence ω_j must be positive. Based on this observation, one can employ the sample estimate of ω_j to rank the features. Given a random sample $\{(\mathbf{x}_i, Y_i)\}, i = 1, \ldots, n$, and assume the features are standardized in the sense that $n^{-1}\sum_{i=1}^n X_{ij} = 0$ and $n^{-1}\sum_{i=1}^n X_{ij}^2 = 1$ for all j. A natural estimator for ω_j is

$$\tilde{\omega}_j = \frac{1}{n}\sum_{k=1}^n \left\{ \frac{1}{n}\sum_{i=1}^n X_{ij}\mathbf{1}(Y_i < Y_k) \right\}^2.$$

It is easy to see that $\widehat{\omega}_j = n^2/(n-1)(n-2)\tilde{\omega}_j$ is the corresponding U-statistic of $\tilde{\omega}_j$ and we can use $\widehat{\omega}_j$ to select important features. The selected submodel is given by

$$\widehat{\mathcal{M}}_{\nu_n} = \{1 \leq j \leq p : \widehat{\omega}_j > \nu_n\}.$$

Zhu et al. (2011) established the consistency in ranking (CIR) property of their procedure, which means that $\widehat{\omega}_j$ always ranks important features above unimportant ones with high probability. This procedure is referred to as the sure independent ranking screening (SIRS). Provided an ideal cutoff is available, this property would lead to consistency in selection in the ultra-high dimensional setup. In practice, one can choose the cutoff value by introducing extra artificial auxiliary variables to the dataset (Zhu et al. 2011). Lin et al. (2013) proposed an improved version of the SIRS for a setting where the relationship between the response and individual feature is symmetric.

Li et al. (2012) proposed a model free feature screening procedure based on the distance correlation (Székely et al. 2007). Let $\mathbf{u} \in \mathbb{R}^{d_u}$ and $\mathbf{v} \in \mathbb{R}^{d_v}$ be two random vectors. The squared distance covariance is defined as

$$\mathrm{dcov}^2(\mathbf{u}, \mathbf{v}) = S_1 + S_2 - 2S_3,$$

where $S_j, j = 1, 2, 3$ are defined as

$$
\begin{aligned}
S_1 &= \mathrm{E}\{\|\mathbf{u} - \tilde{\mathbf{u}}\|_{d_u}\|\mathbf{v} - \tilde{\mathbf{v}}\|_{d_v}\}, \\
S_2 &= \mathrm{E}\{\|\mathbf{u} - \tilde{\mathbf{u}}\|_{d_u}\}\mathrm{E}\{\|\mathbf{v} - \tilde{\mathbf{v}}\|_{d_v}\}, \\
S_3 &= \mathrm{E}\{\mathrm{E}(\|\mathbf{u} - \tilde{\mathbf{u}}\|_{d_u}|\mathbf{u})\mathrm{E}(\|\mathbf{v} - \tilde{\mathbf{v}}\|_{d_v}|\mathbf{v})\},
\end{aligned}
$$

where $(\tilde{\mathbf{u}}, \tilde{\mathbf{v}})$ is an independent copy $(\mathbf{u}, \mathbf{v})$. The distance correlation (DC) between $\mathbf{u}$ and $\mathbf{v}$ is defined as

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

The distance correlation has many appealing properties. For two univariate normal random variables U and V, the distance correlation $\mathrm{dcorr}(U, V)$ is strictly increasing in $|\rho|$, where ρ is the Pearson correlation between U and V. This property implies that the DC-based marginal feature screening procedure is equivalent to the SIS in Fan and Lv (2008) for linear regression if features and errors are normally distributed. The second appealing property is that $\mathrm{dcorr}(\mathbf{u}, \mathbf{v}) = 0$ if and only if $\mathbf{u}$ and $\mathbf{v}$ are independent (Székely et al. 2007). Note that two univariate random variables U and V are independent if and only if U and $T(V)$, a strictly monotone transformation of V, are independent. This implies that a DC-based feature screening procedure can be more effective than the Pearson correlation based procedure since DC can capture the nonlinear relationship between U and V. In addition, DC is well-defined for any random vectors, thus DC-based screening procedure can be directly used for grouped predictors. These remarkable properties make distance correlation a good candidate for feature screening.

Given a random sample $\{(\mathbf{u}_i, \mathbf{v}_i)\}, i = 1, \ldots, n$ from $(\mathbf{u}, \mathbf{v})$, the squared distance covariance between $\mathbf{u}$ and $\mathbf{v}$ is estimated by $\widehat{\mathrm{dcov}}^2(\mathbf{u}, \mathbf{v}) = \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3$, where

$$
\begin{aligned}
\widehat{S}_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i - \mathbf{u}_j\|_{d_u}\|\mathbf{v}_i - \mathbf{v}_j\|_{d_v}, \\
\widehat{S}_2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{u}_i - \mathbf{u}_j\|_{d_u}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{v}_i - \mathbf{v}_j\|_{d_v}, \\
\widehat{S}_3 &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\|\mathbf{u}_i - \mathbf{u}_k\|_{d_u}\|\mathbf{v}_j - \mathbf{v}_k\|_{d_v},
\end{aligned}
$$

and $\|\mathbf{a}\|_d$ stands for the Euclidean norm of $\mathbf{a} \in \mathbb{R}^d$. Similarly, we can define the sample distance covariances $\widehat{\mathrm{dcov}}(\mathbf{u}, \mathbf{u})$ and $\widehat{\mathrm{dcov}}(\mathbf{v}, \mathbf{v})$. Accordingly, the sample distance correlation between $\mathbf{u}$ and $\mathbf{v}$ is defined by

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

Let $\mathbf{y} = (Y_1, \ldots, Y_q)^\top$ be the response vector with support Ψ_y, and $\mathbf{x} = (X_1, \ldots, X_p)^\top$ be the covariate vector. Here we allow the response to be multivariate and assume q is a fixed number. For each $j = 1, \ldots, p$, we can calculate the sample distance correlation $\widehat{\mathrm{dcorr}}(X_j, \mathbf{y})$. Based on the fact that $\mathrm{dcorr}(X_j, \mathbf{y}) = 0$ if and only if X_j and $\mathbf{y}$ are independent, $\widehat{\mathrm{dcorr}}(X_j, \mathbf{y})$ can be regarded as a marginal utility to measure the importance of X_j. Therefore, the set of important variables is defined as

$$\widehat{\mathcal{M}}_{\nu_n} = \{1 \leq j \leq p : \widehat{\mathrm{dcorr}}(X_j, \mathbf{y}) > \nu_n\}$$

with $\nu_n = cn^{-\kappa}$ for some pre-specified constants c and κ. This model free feature screening procedure is known as DC-SIS, which allows for arbitrary regression relationship of Y onto $\mathbf{x}$, regardless of whether it is linear or nonlinear. It also permits univariate and multivariate responses, regardless of whether it is continuous, discrete, or categorical. This DC-SIS is completely model free and it does not require a model assumption on the relationship between features and the response. Since distance correlation is well defined for any random vectors, it can be directly utilized for screening grouped variables and multivariate responses. An iterative procedure for DC-SIS was proposed by Zhong and Zhu (2015) to address the issues of marginal independent learning.

Chapter 3 Regularized Quadratic Programming and its Applications

In this chapter, we consider the regularized quadratic programming with nonconvex penalty and linear constraint. This regularized quadratic programming has many statistical applications including penalized linear regression, linear discriminant analysis, estimation of precision matrix, etc. We study the theoretical properties of the resulting estimator of a general regularized quadratic programming. Under the assumption that the quadratic form satisfies the so-called restricted strong convexity (RSC) condition, we establish the deterministic L_1 and L_2 error bounds for any stationary point. Under slightly stronger condition, we also show that the stationary point is unique and establish the support recovery and L_∞ error bound using the primal-dual witness (PDW) technique. Furthermore, we consider two applications of the regularized quadratic programming: (1) estimation of linear functional, (2) F-type test for regression coefficients and (3) sparse linear discriminant analysis. To solve the regularized quadratic programming, we propose to use the ADMM algorithm with local linear approximation. We also propose a BIC-type criteria to select the regularization parameter.

3.1 Motivation

The work of regularized quadratic programming is motivated by the estimation of functional of the form $\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is a non-singular matrix and $\boldsymbol{\eta} \in \mathbb{R}^p$ is a p-dimensional vector. For example, $\boldsymbol{\Sigma}$ can be the covariance matrix and $\boldsymbol{\eta}$ can be the mean vector or the difference of two mean vectors. Consider the following

quadratic programming

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta} - \boldsymbol{\eta}^\top \boldsymbol{\beta}. \tag{3.1}$$

Clearly, the solution to (3.1) is $\boldsymbol{\beta}^\star = \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$. Alternatively, we can reformulate (3.1) as a quadratic programming with linear equality constraint,

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta} \text{ subject to } \boldsymbol{\eta}^\top \boldsymbol{\beta} = 1. \tag{3.2}$$

The solution to (3.2) is $\boldsymbol{\beta}^\star = \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}/(\boldsymbol{\eta}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta})$. Note that the reformulation (3.2) rules out $\boldsymbol{\eta} = \mathbf{0}$ since $\boldsymbol{\beta}^\star$ is not well-defined when $\boldsymbol{\eta} = \mathbf{0}$. With the equality constraint, the solution $\boldsymbol{\beta}^\star$ is not exactly the same as $\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ but proportional to it. In many statistical applications, such as projection test for mean vector and linear discriminant analysis, the magnitude of the linear functional is not important, it is the direction of the functional that matters.

In high-dimensional data setting where $p > n$, it is well known that consistent estimators for $\boldsymbol{\beta}$ cannot be achieved unless additional structural assumptions are imposed on the model. Following the standard assumption in literature of high-dimensional statistics, we assume that the linear functional $\boldsymbol{\beta}^\star = \boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}$ is sparse, i.e., most of the elements in $\boldsymbol{\beta}^\star$ are 0. To obtain a sparse solution, we consider the following regularized quadratic programming without linear constraint,

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta} - \boldsymbol{\eta}^\top \boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \tag{3.3}$$

or with linear constraint

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}) \text{ subject to } \boldsymbol{\eta}^\top \boldsymbol{\beta} = 1, \tag{3.4}$$

where $P_\lambda(\cdot)$ is some penalty function we will discuss later. In practice, $\boldsymbol{\Sigma}$ and $\boldsymbol{\eta}$ are usually unknown and need to be estimated from data. Replacing $\boldsymbol{\Sigma}$ and $\boldsymbol{\eta}$ by their sample counterparts $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\eta}}$, (3.3) and (3.4) become

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \widehat{\boldsymbol{\eta}}^\top \boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \tag{3.5}$$

and

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}) \text{ subject to } \widehat{\boldsymbol{\eta}}^\top \boldsymbol{\beta} = 1. \tag{3.6}$$

Both (3.5) and (3.6) are special cases of the following regularized quadratic programming with linear constraints,

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top \boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \\ &\text{subject to } \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}, \end{aligned} \tag{3.7}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the p-dimensional unknown parameter of interest, $\mathbf{W} \in \mathbb{R}^{p\times p}$ is a symmetric matrix, $\mathbf{q} \in \mathbb{R}^p$ is a p-dimensional vector, $\mathbf{C} \in \mathbb{R}^{r\times p}$ is a matrix and $\mathbf{b} \in \mathbb{R}^r$ is a r-dimensional vector. Typically, $\mathbf{W}, \mathbf{q}$ and $\mathbf{C}$ are unknown and can be estimated from data and $\mathbf{b}$ is typically a known constant vector. Note that (3.7) also includes the cases where no constraint is imposed. Simply set $\mathbf{C} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$, (3.7) is reduced to the case without constraint. Besides the estimation of the functional, the regularized quadratic programming also includes other applications. We list a few examples here.

Example 3.1: Consider the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is the vector of response and $\mathbf{X}$ is the design matrix. The penalized least squares estimator is given by

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + P_\lambda(\boldsymbol{\beta}). \tag{3.8}$$

We can reformulate the penalized least squares in the form of (3.7) by setting $\mathbf{W} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, $\mathbf{q} = \frac{1}{n}\mathbf{X}^\top\mathbf{y}$, $\mathbf{C} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$. Then the penalized least squares (3.8) becomes

$$\frac{1}{2n}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \frac{1}{n}(\mathbf{X}^\top\mathbf{y})^\top\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}).$$

Example 3.2: Consider the Gaussian graphical model. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n independent copies of $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Since the conditional independence among $\mathbf{x}$ is characterized by the pattern of the precision matrix $\boldsymbol{\Sigma}^{-1}$, inferring the component of the unknown precision matrix $\boldsymbol{\Sigma}^{-1}$ is often of interest. Denote the kth column of $\boldsymbol{\Sigma}^{-1}$ as $\boldsymbol{\beta}_k$, and the estimator of $\boldsymbol{\beta}_k$ based on the following column-wise loss function is proposed by Liu and Luo (2012),

$$\widehat{\boldsymbol{\beta}}_k = \min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \mathbf{e}_k^\top\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \tag{3.9}$$

where $\widehat{\mathbf{\Sigma}}$ is the sample covariance matrix and $\mathbf{e}_k$ is a vector with its k-th element being 1 and all other elements being 0. The objective function in (3.9) is a special case of (3.7) with choices of $\mathbf{W} = \widehat{\mathbf{\Sigma}}$, $\mathbf{q} = \mathbf{e}_k$, $\mathbf{C} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$.

3.2 Theoretical results

3.2.1 Notations and assumptions on penalty function

In this chapter, we study the theoretical results for a regularized quadratic programming with nonconvex penalty and linear constraint. Under the assumption that the quadratic form satisfies the restricted strong convexity (RSC) condition, we establish the deterministic L_1 and L_2 error bounds for our estimator. Our theory applies to any stationary point that satisfies the first order necessary condition to be a local minimum. Under the strict dual feasibility assumption, we also show that the stationary point is actually unique with the help of the primal-dual witness technique and establish the support recovery and L_∞ error bound. In addition, we consider three applications of the regularized quadratic programming: (1) estimation of linear functional, (2) F-type test for regression coefficients and (3) sparse linear discriminant analysis. We establish the convergence rate of our estimator under the sub-Gaussian assumption. Direct computation of the global solution to the nonconvex optimization problem is quite challenging when p is large. We propose an ADMM algorithm with local linear approximation (LLA). The nonconvex penalty function is approximated by it first order expansion and thus becomes convex. It is guaranteed that the solution converges to a local minimum and thus the theoretical results hold for the numerical solution. The ADMM algorithm can naturally handle the liner constraint. A BIC-type criteria is developed to choose the tuning parameter in the penalty function.

We first introduce some notations. For a vector $\mathbf{v} = (v_1, \ldots, v_p)^\top$, its absolute value is defined to be $|\mathbf{v}| = (|v_1|, \ldots, |v_p|)^\top$. We use $\|\mathbf{v}\|_p$ $(p \geq 1)$ to denote L_p norm of the vector $\mathbf{v}$, i.e., $\|\mathbf{v}\|_p = (|v_1|^p +, \ldots, +|v_p|^p)^{\frac{1}{p}}$; we use $\|\mathbf{v}\|_\infty = \max\{|v_1|, \ldots, |v_p|\}$ to denote its infinity norm and $\|\mathbf{v}\|_0 = \#\{j : v_j \neq 0\}$ to denote its L_0 norm, where $\#S$ denotes the cardinality of the set S. For a $p \times p$ matrix $\mathbf{A}$, let $\|\mathbf{A}\|_\infty = \max\{|a_{ij}|, 1 \leq i, j \leq p|\}$ and let $\|\mathbf{A}\|_{L_\infty} = \sup \|\mathbf{A}\mathbf{v}\|_\infty / \|\mathbf{v}\|_\infty$ be

the corresponding infinity norm induced from the infinity norm for a vector. As a result, $\|\mathbf{A}\|_{L_\infty} = \max_i \sum_{j=1}^p |a_{ij}|$ the maximum absolute row sum of the matrix. Let $a \wedge b$ denote smaller one of a and b and let $a \vee b$ denote the larger one of a and b. Let $\boldsymbol{\beta}^\star$ denote the true parameter of interest and $s = \#\{\beta_j^\star : \beta_j^\star \neq 0\}$ be the number of nonzero elements in $\boldsymbol{\beta}^\star$.

A random variable X is sub-exponential if there exists some constant $K > 0$ such that $\mathrm{P}(|X| > t) \leq 2\exp(-t/K)$ for all $t > 0$. A random variable X is sub-Gaussian if there exists some constant $K > 0$ such that $\mathrm{P}(|X| > t) \leq 2\exp(-t^2/K^2)$ for all $t > 0$. If X is sub-exponential, its sub-exponential norm is defined as $\|X\|_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathrm{E}|X|^p)^{1/p}$. If X is sub-Gaussian, its sub-Gaussian norm is defined as $\|X\|_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathrm{E}|X|^p)^{1/p}$. Then X is sub-exponential if and only if $\|X\|_{\psi_1} < \infty$ and X is sub-Gaussian if and only if $\|X\|_{\psi_2} < \infty$. A random vector $\mathbf{x} = (X_1, \ldots, X_p)^\top$ is sub-Gaussian if $\sup_{\|\mathbf{v}\|_2=1} \|\mathbf{v}^\top \mathbf{x}\|_{\psi_2} = K < \infty$ and its norm is defined to be $\|\mathbf{x}\|_{\psi_2} = K$, which implies each component X_j is also sub-Gaussian with sub-Gaussian norm at most K.

Following Loh and Wainwright (2015), we assume $P_\lambda(\cdot)$ satisfies the following conditions.

(i) $P_\lambda(0) = 0$ and $P_\lambda(t)$ is symmetric around 0.

(ii) $P_\lambda(t)$ is differentiable for $t \neq 0$ and $\lim_{t\to 0^+} P'_\lambda(t) = \lambda$.

(iii) $P_\lambda(t)$ is a non-decreasing function on $t \in [0, \infty)$.

(iv) $P_\lambda(t)/t$ is a non-increasing function on $t \in [0, \infty)$.

(v) There exists $\mu > 0$ such that $P_\lambda(t) + \frac{\mu}{2}t^2$ is convex.

(vi) There exists $a > 0$ such that $P'_\lambda(t) = 0$ for all $t \in [a\lambda, \infty)$.

Such conditions on $P_\lambda(t)$ are relatively mild and are satisfied for a wide variety of regularizers. Examples include the SCAD (Fan and Li 2001) and the MCP (Zhang 2010). Fan, Xue and Zou (2014) imposed similar conditions on $P_\lambda(t)$ and such penalties are known as folded concave penalty functions. More specifically, the first derivative of the SCAD is defined by

$$P'_\lambda(|t|) = \lambda \left\{ I(|t| \leq \lambda) + \frac{(a\lambda - |t|)_+}{(a-1)\lambda} I(|t| > \lambda) \right\},$$

for some $a > 2$, where $I(\cdot)$ is the indicator function and b_+ stands for the positive part of b, that is $b_+ = bI(b > 0)$. It is recommended to use $a = 3.7$ by Fan and Li (2001). The first derivative of the MCP is defined by

$$P'_\lambda(|t|) = a^{-1}(a\lambda - |t|)_+ \quad a > 0.$$

The standard L_1 penalty is not in this family of regularizers since it only satisfies conditions (i)-(v). The $L_q(0 < q < 1)$ bridge penalty (Frank and Friedman 1993) is excluded by condition (ii) since its derivative at 0 is unbounded. The capped-L_1 penalty is also excluded because it has points of non-differentiability on the positive real line. Condition (v) is known as weak convexity and is a type of curvature constraint that controls the level of nonconvexity. Although this condition is satisfied by many regularizers of interest, it is again not satisfied by the capped-L_1 penalty for any $\mu > 0$. Appendix A provides more details on the properties of the regularizers.

3.2.2 Main results

We consider the following penalized quadratic optimization problem with linear constraint,

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top \boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}) \\ &\text{s.t. } \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}, \end{aligned} \tag{3.10}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown parameter, $\mathbf{W} \in \mathbb{R}^{p\times p}$ is a symmetric matrix, $\mathbf{q} \in \mathbb{R}^p$, $\mathbf{C} \in \mathbb{R}^{r\times p}$, $\mathbf{b} \in \mathbb{R}^r$ and $P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^p P_\lambda(\beta_j)$ is the penalty function. We allow $\mathbf{C} = \mathbf{0}$, that is there is no constraint on $\boldsymbol{\beta}$.

We impose the following restricted strong convexity (RSC) condition on the matrix $\mathbf{W}$,

$$\boldsymbol{\delta}^\top \mathbf{W}\boldsymbol{\delta} \geq \alpha\|\boldsymbol{\delta}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1 \text{ for all } \|\boldsymbol{\delta}\|_1 \geq 1, \tag{3.11}$$

where $\alpha > 0$ is a strictly positive constant and $\tau \geq 0$ is a nonnegative constant. If $\mathbf{W}$ is positive definite, then the RSC condition in (3.11) naturally holds with $\alpha = \lambda_{\min}(\mathbf{W})$ and $\tau = 0$, where $\lambda_{\min}(\mathbf{W})$ denotes the minimum eigenvalue of $\mathbf{W}$. In the high-dimensional setting where $p > n$, the matrix $\mathbf{W}$ in general is not positive definite or not even semi-positive definite, the RSC condition can still hold with

strictly positive α and τ. In fact, if $\mathbf{W}$ is semi-positive definite (but not positive definite), then $\boldsymbol{\delta}^\top \mathbf{W}\boldsymbol{\delta} \geq 0$ for any $\boldsymbol{\delta} \in \mathbb{R}^p$, thus the RSC condition holds trivially for $\{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1/\|\boldsymbol{\delta}\|_2^2 > c\}$, where $c = \frac{\alpha}{\tau}\sqrt{\frac{n}{\log p}}$. As a result, we only require the RSC condition holds in the set $\{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1/\|\boldsymbol{\delta}\|_2^2 \leq c, \|\boldsymbol{\delta}\|_1 \geq 1\}$. Note that the RSC condition in (3.11) is slightly different from the RSC conditions proposed in Loh and Wainwright (2015), in which they stated two separate RSC inequalities for different ranges of $\|\boldsymbol{\delta}\|_2$. Though we impose the RSC condition on $\mathbf{W}$ only for $\|\boldsymbol{\delta}\|_1 \leq 1$, the following lemma shows that the inequality (3.11) holds for all $\boldsymbol{\delta} \in \mathbb{R}^p$.

Lemma 3.1. *If the RSC condition (3.11) holds, then*

(i) For all $\boldsymbol{\delta} \in \mathbb{R}^p$, $\boldsymbol{\delta}^\top \mathbf{W}\boldsymbol{\delta} \geq \alpha\|\boldsymbol{\delta}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1$.

(ii) For all $\boldsymbol{\delta} \in \mathbb{R}^p$, $\boldsymbol{\delta}^\top \mathbf{W}\boldsymbol{\delta} \geq \alpha\|\boldsymbol{\delta}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1^2$.

Proof. We first prove part (ii). For any $\boldsymbol{\delta} \in \mathbb{R}^p$, the L_1 norm of $\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_1$ is 1 and hence satisfies the RSC condition. We have

$$\begin{aligned}
\frac{\boldsymbol{\delta}^\top}{\|\boldsymbol{\delta}\|_1}\mathbf{W}\frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_1} &\geq \alpha\frac{\|\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}\|_1^2} - \tau\sqrt{\frac{\log p}{n}}\frac{\|\boldsymbol{\delta}\|_1}{\|\boldsymbol{\delta}\|_1} \\
\frac{\boldsymbol{\delta}^\top}{\|\boldsymbol{\delta}\|_1}\mathbf{W}\frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_1} &\geq \alpha\frac{\|\boldsymbol{\delta}\|_2^2}{\|\boldsymbol{\delta}\|_1^2} - \tau\sqrt{\frac{\log p}{n}}\frac{\|\boldsymbol{\delta}\|_1^2}{\|\boldsymbol{\delta}\|_1^2} \\
\boldsymbol{\delta}^\top \mathbf{W}\boldsymbol{\delta} &\geq \alpha\|\boldsymbol{\delta}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1^2.
\end{aligned}$$

If $\|\boldsymbol{\delta}\|_1 < 1$, then $\|\boldsymbol{\delta}\|_1^2 \leq \|\boldsymbol{\delta}\|_1$, implying

$$\boldsymbol{\delta}^\top \mathbf{W}\boldsymbol{\delta} \geq \alpha\|\boldsymbol{\delta}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1,$$

which completes the proof for part (i). □

We establish the deterministic error bounds for any $\widehat{\boldsymbol{\beta}}$ that satisfies the first-order necessary condition to be a local minimum of program (3.10)

$$\langle \mathbf{W}\widehat{\boldsymbol{\beta}} - \mathbf{q} + \nabla P_\lambda(\widehat{\boldsymbol{\beta}}), \boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\rangle \geq 0, \text{ for all feasible } \boldsymbol{\beta} \in \mathbb{R}^p. \tag{3.12}$$

This condition (3.12) is even weaker than the first order KKT condition to be a local minimum of program (3.10). When $\widehat{\boldsymbol{\beta}}$ lies in the interior of the constraint

set $\{\boldsymbol{\beta} : \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}\}$, this condition reduces to the usual zero-subgradient condition $\mathbf{W}\widehat{\boldsymbol{\beta}} - \mathbf{q} + \nabla P_\lambda(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$. We first state a theorem that provides guarantees on the error bounds $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star$ as measured in L_1 and L_2 norms, where $\boldsymbol{\beta}^\star$ is the true parameter.

Theorem 3.1. *Assume that* $\mathbf{W}$ *satisfies the RSC condition in (3.11) with* $\frac{3}{4}\mu < \alpha$ *and* $\boldsymbol{\beta}^\star$ *satisfies the equality constraint* $\mathbf{C}\boldsymbol{\beta}^\star = \mathbf{b}$. *Let* $\widehat{\boldsymbol{\beta}}$ *be any vector that satisfies the first-order necessary condition (3.12) with the choice of* λ *satisfying* $\lambda \geq 4\max\{\min_{\boldsymbol{\xi}\geq\mathbf{0}} \|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q} + \mathbf{C}^\top\boldsymbol{\xi}\|, \tau\sqrt{\log p/n}\}$. *Then we have*

(i) $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \leq \frac{24\lambda s}{4\alpha - 3\mu}$ *and* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \leq \frac{6\lambda\sqrt{s}}{4\alpha - 3\mu}$.

(ii) $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)^\top \mathbf{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \leq \lambda^2 s \left(\frac{9}{4\alpha - 3\mu} + \frac{27\mu}{(4\alpha - 3\mu)^2}\right)$.

Remark. Loh and Wainwright (2015) established similar error bounds in L_1 and L_2 norms for M-estimators under the following general framework,

$$\widehat{\boldsymbol{\beta}} \in \underset{g(\boldsymbol{\beta})\leq R, \boldsymbol{\beta}\in\Omega}{\arg\min}\ \ell_n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}), \tag{3.13}$$

where $\ell_n(\boldsymbol{\beta})$ is some differentiable loss function and is not necessarily convex, $g(\boldsymbol{\beta})$ is some convex function satisfying the lower bound $g(\boldsymbol{\beta}) \geq \|\boldsymbol{\beta}\|_1$, and R is a tuning parameter that needs to be chosen carefully to make sure $\boldsymbol{\beta}^\star$ is in the feasible set. In our case, $\ell_n(\boldsymbol{\beta})$ is a quadratic form

$$\ell_n(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top\boldsymbol{\beta}.$$

The constraint $g(\boldsymbol{\beta}) \leq R$ guarantees the existence of global minimum and the L_1 norm of any stationary point is bounded by R. The error bounds established in Loh and Wainwright (2015) relies on the fact that $\|\widehat{\boldsymbol{\beta}}\|_1 \leq R$ and the choice of tuning parameter λ also depends on R. However, it is not clear how to choose R in practice. On one hand, to ensure $\boldsymbol{\beta}^\star$ is in the feasible set, one needs to choose a relatively large R such that $\|\boldsymbol{\beta}^\star\|_1 \leq R$. On the other hand, they require the tuning parameter λ satisfies $4\|\nabla\ell_n(\boldsymbol{\beta}^\star)\|_\infty \leq \lambda \leq \alpha/6R$ and the sample size satisfies $n \geq 16R^2\tau^2 \log p/\alpha$. If R is too large, then a relatively large sample size n is needed and it is possible that no such λ exists. We modify the RSC condition in Loh and Wainwright (2015) such that the constraint $g(\boldsymbol{\beta}) \leq R$ is no longer needed.

Consequently, we do not require a lower bound on the sample size. In addition, we take advantage of the linear constraint such that a possible smaller lower bound can be found for λ since $\min_{\boldsymbol{\xi}\geq\mathbf{0}} \|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q} + \mathbf{C}^\top\boldsymbol{\xi}\|_\infty \leq \|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q}\|_\infty = \|\nabla\ell_n(\boldsymbol{\beta}^\star)\|_\infty$. A smaller λ leads to narrower L_1 and L_2 error bounds. Similar error bounds can be established under the so called restricted eigenvalue (RE) condition, more details can be found in Appendix C.

Next, we establish the support recovery and L_∞ error bound of $\widehat{\boldsymbol{\beta}}$. We borrow the primal-dual witness (PDW) technique introduced in Loh and Wainwright (2017). This technique is based on the construction of $\widetilde{\boldsymbol{\beta}}$ provided the true support $\mathcal{A} = \{j : \beta_j^\star \neq 0\}$ is known in advance. In particular, we consider the restricted optimization program

$$\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} \in \underset{\boldsymbol{\beta}_{\mathcal{A}}\in\mathbb{R}^{\mathcal{A}},\boldsymbol{\beta}_{\mathcal{A}^c}=\mathbf{0},\mathbf{C}\boldsymbol{\beta}\leq\mathbf{b}}{\arg\min} \frac{1}{2}\boldsymbol{\beta}^\top\mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \tag{3.14}$$

where $\mathbb{R}^{\mathcal{A}} = \{\boldsymbol{\beta}_{\mathcal{A}} : \boldsymbol{\beta} \in \mathbb{R}^p\}$ and $\boldsymbol{\beta}_{\mathcal{A}}$ is a subvector of $\boldsymbol{\beta}$ consisting of elements in set $\mathcal{A}$. We restrict the solution $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}$ in the subspace $\mathbb{R}^{\mathcal{A}}$ such that $\mathrm{supp}(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}) \subset \mathcal{A}$. We partition $\mathbf{W}, \mathbf{q}, \mathbf{C}$ in the following way

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{\mathcal{A}\mathcal{A}} & \mathbf{W}_{\mathcal{A}\mathcal{A}^c} \\ \mathbf{W}_{\mathcal{A}^c\mathcal{A}} & \mathbf{W}_{\mathcal{A}^c\mathcal{A}^c} \end{pmatrix}, \mathbf{q} = \begin{pmatrix} \mathbf{q}_{\mathcal{A}} \\ \mathbf{q}_{\mathcal{A}^c} \end{pmatrix}, \mathbf{C} = \begin{pmatrix} \mathbf{C}_{\mathcal{A}} & \mathbf{C}_{\mathcal{A}^c} \end{pmatrix}.$$

By construction, $\widetilde{\boldsymbol{\beta}}$ satisfies the first order KKT condition of (3.14), that is, there exists $\boldsymbol{\gamma} \in \mathbb{R}^r \geq \mathbf{0}$ satisfying $\gamma_j(\mathbf{b} - \mathbf{C}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}})_j = 0$ for all $1 \leq j \leq r$ such that

$$\mathbf{W}_{\mathcal{A}\mathcal{A}}\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} - \mathbf{q}_{\mathcal{A}} + \nabla P_\lambda(\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}) + \mathbf{C}_{\mathcal{A}}^\top\boldsymbol{\gamma} = \mathbf{0}. \tag{3.15}$$

Define $J_\lambda(t) = \lambda|t| - P_\lambda(t)$. Then zero-gradient condition (3.15) can be rewritten as

$$(\mathbf{W}\widetilde{\boldsymbol{\beta}})_{\mathcal{A}} - \mathbf{q}_{\mathcal{A}} - (\nabla J_\lambda(\widetilde{\boldsymbol{\beta}}))_{\mathcal{A}} + \lambda\widetilde{\mathbf{z}}_{\mathcal{A}} + \mathbf{C}_{\mathcal{A}}^\top\boldsymbol{\gamma} = \mathbf{0},$$

where $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}, \mathbf{0}_{\mathcal{A}^c})$ and $\widetilde{\mathbf{z}}_{\mathcal{A}} \in \partial\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}\|_1$. Then we can find a vector $\widetilde{\mathbf{z}}_{\mathcal{A}^c}$ such that the zero-gradient condition holds in $\mathbb{R}^p$,

$$\mathbf{W}\widetilde{\boldsymbol{\beta}} - \mathbf{q} - \nabla J_\lambda(\widetilde{\boldsymbol{\beta}}) + \lambda\widetilde{\mathbf{z}} + \mathbf{C}^\top\boldsymbol{\gamma} = \mathbf{0}, \tag{3.16}$$

where $\widetilde{\mathbf{z}} = (\widetilde{\mathbf{z}}_{\mathcal{A}}, \widetilde{\mathbf{z}}_{\mathcal{A}^c})$. The next theorem shows that if $\|\widetilde{\mathbf{z}}_{\mathcal{A}^c}\|_\infty < 1$, then all stationary points $\widehat{\boldsymbol{\beta}}$ satisfying condition (3.12) are supported on $\mathcal{A}$ when λ is properly chosen and n is larger than some lower bound. If the submatrix $\mathbf{W}_{\mathcal{A}\mathcal{A}}$ is invertible and $\lambda_{\min}(\mathbf{W}_{\mathcal{A}\mathcal{A}}) \geq \frac{\mu}{2}$, then the restricted program (3.14) is actually convex and the constructed estimator $\widetilde{\boldsymbol{\beta}}$ is unique. Then we can show the stationary point $\widehat{\boldsymbol{\beta}}$ is also unique and agrees with constructed estimator $\widetilde{\boldsymbol{\beta}}$ and the oracle estimator $\widehat{\boldsymbol{\beta}}^{(o)} = (\widehat{\boldsymbol{\beta}}^{(o)}_{\mathcal{A}}, \mathbf{0}_{\mathcal{A}^c})$, where $\widehat{\boldsymbol{\beta}}^{(o)}_{\mathcal{A}}$ is the solution to the unpenalized the quadratic programming restricted on the true support $\mathcal{A}$,

$$\widehat{\boldsymbol{\beta}}^{(o)}_{\mathcal{A}} \in \underset{\boldsymbol{\beta}_{\mathcal{A}} \in \mathbb{R}^{\mathcal{A}}, \boldsymbol{\beta}_{\mathcal{A}^c} = \mathbf{0}, \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}}{\arg\min} \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top \boldsymbol{\beta}.$$

Theorem 3.2. *Assume that* $\mathbf{W}$ *satisfies the RSC condition in (3.11). If* $\|\widetilde{\mathbf{z}}_{\mathcal{A}^c}\|_\infty \leq 1 - \nu$ *for some* $\nu \in (0, 1]$, $\lambda \geq \frac{2\tau}{\nu}\sqrt{\frac{\log p}{n}}$ *and* $n \geq \left(\frac{\tau s}{\alpha - \mu}\right)^2 \left(\frac{4}{\nu} + 2\right)^4 \log p$, *then*

(i) For any $\widehat{\boldsymbol{\beta}}$ *satisfying condition (3.12), we have* $\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c} = \mathbf{0}$.

(ii) If $\lambda_{\min}(\mathbf{W}_{\mathcal{A}\mathcal{A}}) > \frac{\mu}{2}$, *then the stationary point* $\widehat{\boldsymbol{\beta}}$ *is unique and*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_\infty \leq \|\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{W}^{-1}_{\mathcal{A}\mathcal{A}}(\mathbf{C}^\top_{\mathcal{A}}\boldsymbol{\gamma} - \mathbf{q}_{\mathcal{A}})\|_\infty + \|\mathbf{W}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty.$$

(iii) Let $\beta^\star_{\min} = \min\{|\beta^\star_j|, j \in \mathcal{A}\}$ *and further assume*

$$\beta^\star_{\min} \geq \lambda(a + \|\mathbf{W}^{-1}_{\mathcal{A}\mathcal{A}}\|_\infty) + \|\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{W}^{-1}_{\mathcal{A}\mathcal{A}}(\mathbf{C}^\top_{\mathcal{A}}\boldsymbol{\gamma} - \mathbf{q}_{\mathcal{A}})\|_\infty,$$

then $\widehat{\boldsymbol{\beta}}$ *agrees withe the oracle estimator* $\widehat{\boldsymbol{\beta}}^{(o)}$ *and*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_\infty \leq \|\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{W}^{-1}_{\mathcal{A}\mathcal{A}}(\mathbf{C}^\top_{\mathcal{A}}\boldsymbol{\gamma} - \mathbf{q}_{\mathcal{A}})\|_\infty.$$

3.2.3 Application 1: estimation of linear functional

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent and identically distributed random vectors with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Of interest is to estimate the estimate the the linear functional of the form $\boldsymbol{\beta}^\star = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/(\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$. Let $\bar{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}}$ be the

sample mean vector and sample covariance matrix,

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

We can apply the regularized quadratic programming (3.10) to estimate the the linear functional $\boldsymbol{\beta}^\star$. Simply set $\mathbf{W} = \widehat{\boldsymbol{\Sigma}}$, $\mathbf{q} = \mathbf{0}$, $\mathbf{C} = \bar{\mathbf{x}}^\top$, $\mathbf{b} = 1$ and take the equality linear constraint, then (3.10) becomes

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \\ &\text{s.t. } \bar{\mathbf{x}}^\top\boldsymbol{\beta} = 1. \end{aligned} \tag{3.17}$$

Many problems can be formulated in the form of (3.17), including one-sample projection test for mean vector and Markowitz portfolio allocation problem. To apply the results in Theorem 3.1, the true parameter $\boldsymbol{\beta}^\star$ needs to satisfy the equality constraint. However in our case, the true linear functional $\boldsymbol{\beta}^\star$ doest not necessarily satisfy the linear equality constraint $\bar{\mathbf{x}}^\top\boldsymbol{\beta} = 1$. To this end, we define $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/\bar{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, which is proportional to the true linear functional $\boldsymbol{\beta}^\star$ but also satisfies the linear constraint $\bar{\mathbf{x}}^\top\widetilde{\boldsymbol{\beta}} = 1$. In many applications, the magnitude of the linear functional does not matter, it is the direction of the linear functional that plays the key role. In other words, the two linear functionals $\boldsymbol{\beta}^\star$ and $\widetilde{\boldsymbol{\beta}}$ has exactly the same performance in these applications. Define $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. The next theorem states the L_1 and L_2 error bounds for $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star$.

Theorem 3.3. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are identically and independently distributed sub-Gaussian vectors with finite sub-Gaussian norm K and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ satisfies the RSC condition in (3.11). Let $\widehat{\boldsymbol{\beta}}$ be a stationary point of the program (3.17) with $\lambda = sM\sqrt{\log p/n}$ for some large constant M. Assume that there exists $C_1, C_2 > 0$ and $0 < \epsilon < 1$ such that $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \geq C_1$, $\|\boldsymbol{\theta}\|_\infty \leq C_2$ and $C_2\lambda \leq 1 - \epsilon$. Then with probability at least $1 - cp^{-1}$ for some absolute constant c, we have*

(i). $\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_1 = O(s\lambda)$ *and* $\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2 = O(\sqrt{s}\lambda)$.

(ii). $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 = O(s\lambda)$ *and* $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O(\sqrt{s}\lambda)$.

The error bounds between $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ can be regarded as a direct application of Theorem 3.10 since $\widetilde{\boldsymbol{\beta}}$ satisfies the equality constraint. The error bounds depend on the the sparsity level s and the choice of λ. Note that $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/\bar{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. The conditions $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \geq C_1$ and $C_2\lambda \leq 1-\epsilon$ are needed to ensure that $\bar{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is close enough to $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and thus $\bar{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is bounded away from 0. The error bounds between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^\star$ is based on the triangle inequality,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_k \leq \|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_k + \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_k, \text{ for } k = 1, 2.$$

The conditions $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \geq C_1$ and $C_2\lambda \leq 1-\epsilon$ also ensure that $\widetilde{\boldsymbol{\beta}}$ is close enough to $\boldsymbol{\beta}^\star$.

We may also consider the unconstrained version of (3.17)

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \bar{\mathbf{x}}^\top\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}). \tag{3.18}$$

Similarly, we can also establish the error bounds for $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ with $\widehat{\boldsymbol{\theta}}$ being any stationary point of the program (3.18).

Theorem 3.4. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are identically and independently distributed sub-Gaussian vectors with finite sub-Gaussian norm K and the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ satisfies the RSC condition in (3.11). Let $\widehat{\boldsymbol{\theta}}$ be a stationary point of the program (3.18) with $\lambda = sM\sqrt{\log p/n}$ for some large constant M. Assume that $\|\boldsymbol{\theta}\|_\infty \leq C_2$ for some $C_2 > 0$. Then we have*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_1 = O(s\lambda) \text{ and } \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_2 = O(\sqrt{s}\lambda).$$

with probability at least $1 - cp^{-1}$ with some absolute constant c.

The proof of Theorem 3.4 is very similar to that of Theorem 3.3 and is omitted here. Theorem 3.4 characterizes the L_1 and L_2 error bounds for $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ under weaker conditions. Since $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, we do not need to worry about the scale $\gamma = 1/\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\widehat{\gamma} = 1/\bar{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and thus we can get rid of the assumptions $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \geq C_1$ and $C_2\lambda \leq 1-\epsilon$ in Theorem 3.3. As a result, Theorem 3.4 also holds even when $\boldsymbol{\mu} = \mathbf{0}$. However, based on our empirical study, solving program (3.17) is less time consuming than solving program (3.18).

The conditions in Theorem 3.3 and Theorem 3.4 are relatively mild. To bet-

ter understand the conditions, we examine two examples with commonly used correlation structures.

Example 3.3: In this example, we consider the compound symmetry correlation structure $\mathbf{\Sigma}_1 = (1-\rho)\mathbf{I} + \rho\mathbf{E}$, where $\mathbf{I} \in \mathbb{R}^{p\times p}$ is the identity matrix and $\mathbf{E} \in \mathbb{R}^{p\times p}$ is a matrix with all its elements being 1. It is inverse can be written as

$$\mathbf{\Sigma}_1^{-1} = \frac{1}{1-\rho}\left(\mathbf{I} - \frac{1}{1/\rho + p - 1}\mathbf{E}\right).$$

In order to have an approximately sparse projection direction, we assume $\boldsymbol{\mu}$ is also sparse $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_s, 0, \ldots, 0)^\top$ with $s \ll p$. For example, with the choice $\mu_j = a$ for all $j = 1, \ldots, s$, one can see that $\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}$ is dominated by its first s components and the rest $(p-s)$ components converge to 0 as $p \to \infty$. As $p \to \infty$, $\frac{1}{1/\rho+p-1}\mathbf{E}$ degenerates to $\mathbf{0}$ and thus $\mathbf{\Sigma}_1^{-1}$ is dominated by $\frac{1}{1-\rho}\mathbf{I}$. As a result $\boldsymbol{\mu}^\top\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu} \approx \frac{1}{\rho}\|\boldsymbol{\mu}\|_2^2$. The condition $\boldsymbol{\mu}^\top\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu} \geq C_1$ becomes $\frac{1}{\rho}\|\boldsymbol{\mu}\|_2^2 \geq C_1$, which imposes the minimal signal condition on the mean vector. The condition $\|\boldsymbol{\theta}\|_\infty \leq C_2$ is roughly $\|\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}\|_\infty \approx \frac{1}{1-\rho}\|\boldsymbol{\mu}\|_\infty \leq C_2$, which says that the largest magnitude in $\boldsymbol{\mu}$ should be bounded.

Example 3.4: In this example, we consider the autocorrelation structure $\mathbf{\Sigma}_2 = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = \rho^{|i-j|}$. Its inverse is a banded matrix and can be written as

$$\mathbf{\Sigma}_2^{-1} = (1+\rho^2)\mathbf{I} - \rho\mathbf{F} - \rho^2\mathbf{G},$$

where $\mathbf{F} = (f_{ij})_{p\times p}$ with $f_{ij} = 1$ if $|i-j| = 1$ otherwise 0 and $\mathbf{G} = \mathrm{diag}\{1, 0, \ldots, 0, 1\}$.

$$\begin{aligned}
\boldsymbol{\mu}^\top\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu} &= (1+\rho^2)\sum_{j1}^{p}\mu_j^2 - \rho\sum_{|i-j|=1}\mu_i\mu_j - \rho^2(\mu_1^2 + \mu_p^2) \\
&= (1+\rho^2-2\rho)\sum_{j=1}^{p}\mu_j^2 + \rho\left(\mu_1^2 + \mu_p^2 + 2\sum_{j=2}^{p-1}\mu_j^2 - \sum_{|i-j|=1}\mu_i\mu_j\right) \\
&+ (\rho-\rho^2)(\mu_1^2+\mu_p^2) \\
&\geq (1-\rho)^2\sum_{j=1}^{p}\mu_j^2 + \rho\sum_{i-j=1}(\mu_i-\mu_j)^2 \\
&\geq (1-\rho)^2\|\boldsymbol{\mu}\|_2^2.
\end{aligned}$$

Therefore, $\boldsymbol{\mu}^{-1}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu} \geq C_1$ is satisfied if $\|\boldsymbol{\mu}\|_2^2 \geq C_2$. We can also verify that $\|\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}\|_\infty$ is bounded if the largest magnitude in $\boldsymbol{\mu}$ is bounded.

Since the magnitude of the projection direction does not matter, we can measure the distance between the estimator and true linear functional in terms of cosine similarity. The cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$ is defined by

$$\cos\langle \mathbf{u}, \mathbf{v}\rangle = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}.$$

If $\mathbf{u}$ and $\mathbf{v}$ are of the same direction, i.e., $\mathbf{u} = a\mathbf{v}$ for some $a > 0$, then the cosine similarity is 1. In other words, if the cosine similarity is close to 1, then the two vectors are close to each other. If we can obtain good estimators $\boldsymbol{\beta}^\star$ or $\boldsymbol{\theta}$, we can show that the cosine similarity converges to 1. We state the results in Corollary 3.1 and Corollary 3.2.

Corollary 3.1. *Let $\widehat{\boldsymbol{\beta}}$ be a stationary point of the program (3.17) with $\lambda = sM\sqrt{\log p/n}$ for some large constant M. Suppose the conditions in Theorem 3.3 hold and further assume $\sqrt{s}\lambda/\|\boldsymbol{\beta}^\star\|_2 = o(1)$, then we have $\cos\langle \widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star\rangle \to 1$.*

Corollary 3.2. *Let $\widehat{\boldsymbol{\theta}}$ be a stationary point of the program (3.18) with $\lambda = sM\sqrt{\log p/n}$ for some large constant M. Suppose the conditions in Theorem 3.4 hold and further assume $\sqrt{s}\lambda/\|\boldsymbol{\theta}\|_2 = o(1)$, then we have $\cos\langle \widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}\rangle \to 1$.*

These results can be easily extended to the two-sample problem. Let $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n_1}^{(1)}$ be identically and independently distributed random vectors with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}$ and $\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)}$ be identically and independently distributed random vectors with mean vector $\boldsymbol{\mu}_2$ and common covariance matrix $\boldsymbol{\Sigma}$. Of interest is to estimate the functional $\boldsymbol{\theta}_d = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$ or its scaled version $\boldsymbol{\beta}_d^\star = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d/(\boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d)$, where $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is the difference of the two population mean vectors. Let $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ be the sample means of the two populations, $\bar{\mathbf{x}}_d$ be the difference of the sample means and $\widehat{\boldsymbol{\Sigma}}$ be the pooled sample covariance matrix,

$$\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbf{x}_i^{(k)},\ \widehat{\boldsymbol{\Sigma}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} (\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}_k)(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}_k)^\top, k = 1, 2,$$
$$\bar{\mathbf{x}}_d = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2,\ \widehat{\boldsymbol{\Sigma}} = (n_1\widehat{\boldsymbol{\Sigma}}_1 + n_2\widehat{\boldsymbol{\Sigma}}_2)/(n_1 + n_2).$$

Consider the constrained quadratic programming with $\mathbf{W} = \widehat{\boldsymbol{\Sigma}}$, $\mathbf{q} = \mathbf{0}$, $\mathbf{C} = \bar{\mathbf{x}}_d^\top$ and $\mathbf{b} = 1$, (3.17) becomes

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \\ &\text{s.t. } \bar{\mathbf{x}}_d^\top \boldsymbol{\beta} = 1. \end{aligned} \tag{3.19}$$

Or its unconstrained version

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} - \bar{\mathbf{x}}_d^\top \boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}). \tag{3.20}$$

The resulting estimator from problem (3.19) or (3.20) can be applied to projection test for two-sample mean vector test, linear discriminant analysis, etc. Let $s_d = \|\boldsymbol{\theta}_d\|_0 = \|\boldsymbol{\beta}_d^\star\|_0$ and $n' = \min\{n_1, n_2\}$. We simply list the results in the following theorems.

Theorem 3.5. *Suppose $\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{n_k}^{(k)}$ are identically and independently distributed sub-Gaussian vectors with finite sub-Gaussian norm K for $k = 1, 2$ and the pooled sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ satisfies the RSC condition in (3.11). Let $\widehat{\boldsymbol{\beta}}_d$ be a stationary point of the program (3.19) with $\lambda = s_d M\sqrt{\log p/n'}$ for some large constant M. Assume that there exists $C_1, C_2 > 0$ and $0 < \epsilon < 1$ such that $\boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d \geq C_1$, $\|\boldsymbol{\theta}_d\|_\infty \leq C_2$ and $C_2\lambda \leq 1 - \epsilon$. Then with probability at least $1 - cp^{-1}$ for some absolute constant c, we have*

(i). $\|\widehat{\boldsymbol{\beta}}_d - \widetilde{\boldsymbol{\beta}}_d\|_1 = O(s\lambda)$ and $\|\widehat{\boldsymbol{\beta}}_d - \widetilde{\boldsymbol{\beta}}_d\|_2 = O(\sqrt{s}\lambda)$.

(ii). $\|\widehat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d^\star\|_1 = O(s\lambda)$ and $\|\widehat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d^\star\|_2 = O(\sqrt{s}\lambda)$.

Theorem 3.6. *Suppose $\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{n_k}^{(k)}$ are identically and independently distributed sub-Gaussian vectors with finite sub-Gaussian norm K for $k = 1, 2$ and the pooled sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ satisfies the RSC condition in (3.11). Let $\widehat{\boldsymbol{\beta}}_d$ be a stationary point of the program (3.20) with $\lambda = sM\sqrt{\log p/n'}$ for some large constant M. Assume that $\|\boldsymbol{\theta}_d\|_\infty \leq C_2$ for some $C_2 > 0$. Then we have*

$$\|\widehat{\boldsymbol{\theta}}_d - \boldsymbol{\theta}_d\|_1 = O(s\lambda) \text{ and } \|\widehat{\boldsymbol{\theta}}_d - \boldsymbol{\theta}_d\|_2 = O(\sqrt{s}\lambda).$$

with probability at least $1 - cp^{-1}$ with some absolute constant c.

3.2.4 Application 2: F-type test for $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{b}$

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.21}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times p}$ is the matrix of covariates, $\mathbf{y} \in \mathbb{R}^n$ is the response vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is the error term with each ε_i having mean 0 and variance σ^2 and is independent of $\mathbf{X}$. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the vector of unknown regression coefficients. Of interest is to test if the linear combination of the coefficient is equal to a known vector $\mathbf{b}$,

$$H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{b} \quad \text{verses} \quad H_1 : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{b}, \tag{3.22}$$

where $\mathbf{A} \in \mathbb{R}^{m\times p}$ is a constant matrix and $\mathbf{b} \in \mathbb{R}^{m\times 1}$ is a known vector. A special case of (3.22) is to test whether a subset of $\boldsymbol{\beta}$ is $\mathbf{0}$ or not,

$$H_0 : \boldsymbol{\beta}_{\mathcal{A}} = \mathbf{0} \quad \text{verses} \quad H_1 : \boldsymbol{\beta}_{\mathcal{A}} \neq \mathbf{0},$$

where $\mathcal{A}$ is a subset of $\{1, \ldots, p\}$. A F-type test can be constructed to test whether $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$. Let $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_1$ be the estimators for the coefficient under H_0 and H_1, respectively. The F-type test statistic is defined as

$$F = \frac{\mathrm{RSS}(\widehat{\boldsymbol{\beta}}_0) - \mathrm{RSS}(\widehat{\boldsymbol{\beta}}_1)}{\mathrm{RSS}(\widehat{\boldsymbol{\beta}}_1)},$$

where $\mathrm{RSS}(\widehat{\boldsymbol{\beta}}_0)$ and $\mathrm{RSS}(\widehat{\boldsymbol{\beta}}_1)$ are the residual sum squares under H_0 and H_1 respectively. Under the high-dimensional setting where $p > n$, we assume the true regression coefficient $\boldsymbol{\beta}^\star$ is sparse. Under H_1, we consider the following penalized least squares

$$\widehat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + P_\lambda(\boldsymbol{\beta}), \tag{3.23}$$

where $P_\lambda(\cdot)$ is some nonconvex penalty function. The penalized least squares in (3.23) can be rewritten as

$$\widehat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \frac{1}{n}(\mathbf{X}^\top\mathbf{y})^\top\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}),$$

which is in the form of (3.7) with $\mathbf{W} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, $\mathbf{q} = \frac{1}{n}\mathbf{X}^\top\mathbf{y}$, $\mathbf{C} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$. Similarly, under H_0, $\boldsymbol{\beta}$ can be estimated by

$$\widehat{\boldsymbol{\beta}}_0 = \arg\min_{\mathbf{A}\boldsymbol{\beta}=\mathbf{b}} \frac{1}{2n}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} - \frac{1}{n}(\mathbf{X}^\top\mathbf{y})^\top\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \tag{3.24}$$

which is in the form of (3.7) with $\mathbf{W} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$, $\mathbf{q} = \frac{1}{n}\mathbf{X}^\top\mathbf{y}$ and linear constraint $\mathbf{A}\boldsymbol{\beta} = \mathbf{b}$.

Suppose that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent and identically distributed sub-Gaussian random variables with $\|\varepsilon_i\|_{\psi_2} = K_1 < \infty$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent and identically distributed sub-Gaussian random vectors with $\|\mathbf{x}_i\|_{\psi_2} = K_2 < \infty$. The following theorem states the L_1 and L_2 error bounds for the estimators under null and alternative.

Theorem 3.7. *Let $\widehat{\boldsymbol{\beta}}_1$ be a stationary point of (3.23) and $\widehat{\boldsymbol{\beta}}_0$ be a stationary point of (3.24) with the choice of $\lambda = M\sqrt{\log p/n}$ for some large constant M. If $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ satisfies the RSC condition and $\log p < n$, then we have*

(1) $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^\star\|_1 = O(\lambda s)$, $\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^\star\|_2 = O(\lambda\sqrt{s})$.

(2) Under $H_0 : \mathbf{A}\boldsymbol{\beta}^\star = \mathbf{b}$, $\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^\star\|_1 = O(\lambda s)$, $\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^\star\|_2 = O(\lambda\sqrt{s})$.

with probability at least $1 - cp^{-1}$, where c is an absolute constant and $\kappa = 4\alpha - 3\mu$.

Remark. In the above theorem, we assume the covariate $\mathbf{x}_i$ and error ε_i are both sub-Gaussian random variables. In some settings, people assume the design matrix $\mathbf{X}$ is fixed. Instead of the sub-Gaussian assumption on $\mathbf{X}$, we assume $\mathbf{X}$ is normalized in the sense that $\|\mathbf{x}_{(j)}\|_2/\sqrt{n} \leq 1$, where $\mathbf{x}_{(j)}$ is the j-th column of $\mathbf{X}$. Terror bounds in Theorem 3.7 still hold, see Negahban et al. (2012).

3.2.5 Application 3: sparse linear discriminant analysis

In this section, we consider the linear discriminant analysis with a large number of features. Let $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n_1}^{(1)}$ be identically and independently distributed random samples from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ (class 1) and $\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)}$ be identically and independently distributed random samples from $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ (class 2) with the same covariance matrix $\boldsymbol{\Sigma}$, respectively. Given a new observation $\mathbf{z}$, we are interested in classifying the new observation to one of the two classes. Let $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ be the difference

of the two population means and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Fisher's linear discriminant rule classifies the new observation $\mathbf{z}$ into class 1 if and only if $\delta_F(\mathbf{z}) = 1$ where

$$\delta_F(\mathbf{z}) = I\left\{(\mathbf{z} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d \geq \log \frac{\pi_2}{\pi_1}\right\},$$

π_1 and π_2 are the prior probabilities for class 1 and class 2, and $I(\cdot)$ denotes the indicator function. If we assume that the two classes have the same prior probability $\pi_1 = \pi_2 = \frac{1}{2}$, then the Fisher's discriminant rule becomes

$$\delta_F(\mathbf{z}) = I\{(\mathbf{z} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d \geq 0\}.$$

However, this rule is not directly applicable in practice since the parameters in the rule is typically unknown and needs to be estimated from samples. For any linear discriminant rule associated with the classification direction $\boldsymbol{\beta}$

$$\delta_{\boldsymbol{\beta}}(\mathbf{z}) = I\{\boldsymbol{\beta}^\top (\mathbf{z} - \boldsymbol{\mu}) > 0\},$$

the theoretical misclassification rate of the classifier $\delta_{\boldsymbol{\beta}}$ is

$$W(\delta_{\boldsymbol{\beta}}) = 1 - \Phi\left(\frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\mu}_d / (\boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta})^{1/2}\right),$$

and the empirical misclassification rate is

$$W_n(\delta_{\boldsymbol{\beta}}) = 1 - \Phi\left(\frac{1}{2}\boldsymbol{\beta}^\top \bar{\mathbf{x}}_d / (\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta})^{1/2}\right),$$

where $\bar{\mathbf{x}}_d = \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}$ is the difference of the two sample means and $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}(n_1 \widehat{\boldsymbol{\Sigma}}^{(1)} + n_2 \widehat{\boldsymbol{\Sigma}}^{(2)})$ is the pooled sample covariance matrix,

$$n = n_1 + n_2, \ \widehat{\boldsymbol{\Sigma}}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} (\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})(\mathbf{x}_i^{(k)} - \bar{\mathbf{x}}^{(k)})^\top, k = 1, 2.$$

The theoretical misclassification rate of the Fisher's discriminant rule is $1 - \Phi(\frac{1}{2}\Delta_p^{1/2})$, where $\Delta_p = \boldsymbol{\mu}_d^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d$. However, for high-dimensional data, it is almost impossible to achieve such a good performance empirically. In addition, when $p > n$, the estimated sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is typically ill-conditioned or not invertible.

One possible solution is simply ignoring the dependence among variables. For example, Bickel and Levina (2004) proposed an independence rule, which uses a diagonal matrix with diagonal elements from the sample covariance matrix to replace $\widehat{\boldsymbol{\Sigma}}$. To improve the performance of the independence rule, Fan and Fan (2008) proposed the Features Annealed Independence Rule (FAIR), which consists of two steps: (1) select important features by two-sample t-test and (2) apply the independence rule to the selected features. Fan and Fan (2008) proved that using all features may increase the misclassification rate due to the noise accumulation and therefore feature selection is important for high-dimensional classification problem. Another class of methods focus on estimating $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$ directly under the sparsity assumption. Cai and Liu (2011*b*) used the Dantzig-type selector to estimate the classification direction. Mai et al. (2012) proposed a penalized least squares estimator by introducing some dummy variables as the response. Fan et al. (2012) proposed the ROAD estimator which solves quadratic programming with linear constraint and L_1 penalty.

This motivates us to apply the regularized method to estimate linear functional $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$ directly. Set $\mathbf{W} = \widehat{\boldsymbol{\Sigma}}$, $\mathbf{q} = \mathbf{0}$, $\mathbf{C} = \bar{\mathbf{x}}_d$ and $\mathbf{b} = 1$ in (3.7), we have

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}\, \boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}) \\ &\text{s.t. } \bar{\mathbf{x}}_d^\top \boldsymbol{\beta} = 1. \end{aligned} \tag{3.25}$$

Remark. The ROAD estimator in Fan et al. (2012) is given by

$$\widehat{\boldsymbol{\beta}}_{\text{ROAD}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta} \quad \text{s.t. } \bar{\mathbf{x}}_d^\top \boldsymbol{\beta} = 1 \text{ and } \|\boldsymbol{\beta}\|_1 \le c. \tag{3.26}$$

It is equivalent to

$$\widehat{\boldsymbol{\beta}}_{\text{ROAD}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{\beta} + \lambda \|\boldsymbol{\beta}\|_1 \quad \text{s.t. } \bar{\mathbf{x}}_d^\top \boldsymbol{\beta} = 1,$$

where λ depends on the parameter c in (3.26). Therefore, the ROAD estimator is equivalent to the estimator resulting from the first step of the two-step LLA estimator, more details about two-step LLA estimator can be found in Section 3.3.

Let $\widehat{\boldsymbol{\beta}}$ be the LLA estimator for (3.25), then

$$\widehat{\boldsymbol{\beta}} = \underset{\bar{\mathbf{x}}_d^\top \boldsymbol{\beta}=1}{\arg\min} \ \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + \sum_{j=1}^{p} \lambda_j |\beta_j|,$$

where $(\lambda_1, \ldots, \lambda_p)^\top = (P'_\lambda(|\beta_1^{(1)}|), \ldots, P'_\lambda(|\beta_p^{(1)}|))^\top$ and $\widehat{\boldsymbol{\beta}}^{(1)}$ is the estimator resulting from the first step. It is equivalent to the following constraint version

$$\widehat{\boldsymbol{\beta}} = \underset{\sum_{j=1}^{p} \lambda_j |\beta_j| \le c_\lambda, \ \bar{\mathbf{x}}_d^\top \boldsymbol{\beta}=1}{\arg\min} \ \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}.$$

The constant c_λ depends on λ_j's. We consider the constraint version with population parameters $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}_d$ using the same c_λ,

$$\widetilde{\boldsymbol{\beta}} = \underset{\sum_{j=1}^{p} \lambda_j |\beta_j| \le c_\lambda, \ \boldsymbol{\mu}_d^\top \boldsymbol{\beta}=1}{\arg\min} \ \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}. \tag{3.27}$$

Note that $1 = \boldsymbol{\mu}_d^\top \boldsymbol{\beta} \le \|\boldsymbol{\beta}\|_1 \|\boldsymbol{\mu}_d\|_\infty$, thus $\lambda_- \le \lambda_- \|\boldsymbol{\beta}\|_1 \|\boldsymbol{\mu}_d\|_\infty \le c_\lambda \|\boldsymbol{\mu}_d\|_\infty$, where $\lambda_- = \min\{|\lambda_j|, \lambda_j \neq 0\}$. The existence of a feasible solution for (3.27) dictates that $c_\lambda \ge \lambda_- / \|\boldsymbol{\mu}_d\|_\infty$. When $c_\lambda \ge |\boldsymbol{\lambda}|^\top |\boldsymbol{\beta}^\star|$, the constraint $\sum_{j=1}^{p} \lambda_j |\beta_j| \le c_\lambda$ becomes redundant and it reduces to the Fisher's discriminant rule. When c_λ is small, we obtain a sparse solution and achieve feature selection by using covariance information. To study the theoretical property of the LLA classifier based on $\widehat{\boldsymbol{\beta}}$. We introduce an intermediate optimization problem for convenience:

$$\bar{\boldsymbol{\beta}} = \underset{\sum_{j=1}^{p} \lambda_j |\beta_j| \le c_\lambda, \ \bar{\mathbf{x}}^\top \boldsymbol{\beta}=1}{\arg\min} \ \frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Sigma}\boldsymbol{\beta}.$$

Theorem 3.8. *Let $\tilde{s} = \|\tilde{\boldsymbol{\beta}}\|_0$, $\bar{s} = \|\bar{\boldsymbol{\beta}}\|_0$, $\widehat{s} = \|\widehat{\boldsymbol{\beta}}\|_0$ and $\nu = (\|\tilde{\boldsymbol{\beta}}\|_1 \vee \|\bar{\boldsymbol{\beta}}\|_1 \vee \|\widehat{\boldsymbol{\beta}}\|_1)$. Assume that $\lambda_{\min}(\boldsymbol{\Sigma}) \ge \sigma_0^2 > 0$, $\|\bar{\mathbf{x}}_d - \boldsymbol{\mu}_d\|_\infty = O(a_n)$ and $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty = O(b_n)$. If $c_\lambda > (\boldsymbol{\lambda}^\top |\tilde{\boldsymbol{\beta}}| / \tilde{\boldsymbol{\beta}}^\top \bar{\mathbf{x}}_d \vee \boldsymbol{\lambda}^\top |\bar{\boldsymbol{\beta}}| / \bar{\boldsymbol{\beta}}^\top \boldsymbol{\mu}_d)$, then we have*

(i) $W(\widehat{\boldsymbol{\beta}}) - W(\tilde{\boldsymbol{\beta}}) = O(d_n)$ and

(ii) $W_n(\widehat{\boldsymbol{\beta}}) - W(\widehat{\boldsymbol{\beta}}) = O(c_n)$,

where $c_n = (\nu^2 b_n \vee a_n \sqrt{\tilde{s} \vee \bar{s}})$ and $d_n = c_n \vee a_n \sqrt{\widehat{s}}$.

Remark. Under the assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d sub-Gaussian, we can take $a_n = b_n = \sqrt{\frac{\log p}{n}}$.

3.3 ADMM algorithm with local linear approximation

3.3.1 Local linear approximation

The nonconvex penalty function in (3.7) brings difficulty in solving the regularized quadratic programming. Direct computation of the global solution to the nonconvex optimization (3.7) is quite challenging, especially in high-dimensional settings. Liu et al. (2016) developed a mixed integer linear programming that finds a provably global optimal solution to a certain class of nonconvex learning problems. In terms of solution quality, this mixed integer linear programming outperforms other out-of-art algorithms such as gradient-based methods, which typically find a local minimum solution. However, this mixed integer linear programming can be much more computationally expensive than gradient methods. For practical data analysis, it is critical to find an efficient procedure which can find a local solution with satisfactory theoretical properties. Zou and Li (2008) proposed the local linear approximation (LLA) algorithm to deal with nonconvex penalty in the setting of penalized least squares. The idea of LLA is to approximate the nonconvex penalty function by its first order expansion. According to Theorem 3.1, the error bounds we derive hold for any stationary point, including all the local minimum. We propose an ADMM algorithm with local linear approximation, which converges to a local minimum of the regularized linear programming.

Suppose β_{j0} is close to β_j, ignoring the constant that does not involve β_j, the penalty function can be approximated by

$$P_\lambda(|\beta_j|) \approx P'_\lambda(|\beta_{j0}|)(|\beta_j| - |\beta_{j0}|).$$

Given the current solution $\boldsymbol{\beta}^{(k)}$, program (3.7) can be approximated by

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top \boldsymbol{\beta} + \sum_{j=1}^{p} \omega_j^{(k)} |\beta_j|, \\ &\text{s.t. } \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}, \end{aligned} \tag{3.28}$$

where $\omega_j^{(k)} = P'_\lambda(|\beta_j^{(k)}|)$. We summarize the details in Algorithm 1.

Algorithm 1 Local linear approximation (LLA) algorithm.

Initialize $\widehat{\boldsymbol{\beta}}^{(0)}$ and compute the adaptive weights

$$\boldsymbol{\omega}^{(0)} = (\omega_1^{(0)}, \ldots, \omega_p^{(0)})^\top = (P'_\lambda(|\widehat{\beta}_1^{(0)}|), \ldots, P'_\lambda(|\widehat{\beta}_p^{(0)}|))^\top.$$

For $k = 1, 2, \ldots$ repeat the following steps till convergence:

1. Update $\widehat{\boldsymbol{\beta}}$ by solving the following optimization problem

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \arg\min_{\mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}} \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top \boldsymbol{\beta} + \sum_{j=1}^{p} \omega_j^{(k)} |\beta_j|.$$

2. Update the adaptive weight by setting $\omega_j^{(k+1)} = P'_\lambda(|\widehat{\beta}_j^{(k+1)}|)$.

Fan, Xue and Zou (2014) showed that the LLA estimator finds the oracle estimator after one iteration provided that the initial estimator $\widehat{\boldsymbol{\beta}}^{(0)}$ is close to the true parameter. Let

$$Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}, \lambda) = \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top \boldsymbol{\beta} + \sum_{j=1}^{p} P'_\lambda(|\beta_j^{(k)}|)|\beta_j|.$$

Starting with initial value $\mathbf{0}$, we propose a two-step LLA estimator consisting of the following two steps:

Step 1 :$\widehat{\boldsymbol{\beta}}^{(1)} = \arg\min_{\mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}} Q(\boldsymbol{\beta}|\mathbf{0}, \tau\lambda)$,

Step 2 :$\widehat{\boldsymbol{\beta}} = \arg\min_{\mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}} Q(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^{(1)}, \lambda)$.

Remark. Note that in step 1, the tuning parameter we use is $\tau\lambda$. Starting with the initial value $\mathbf{0}$ in Step 1, we have $P'_{\tau\lambda}(|0|) = \tau\lambda$. Therefore, the first step is

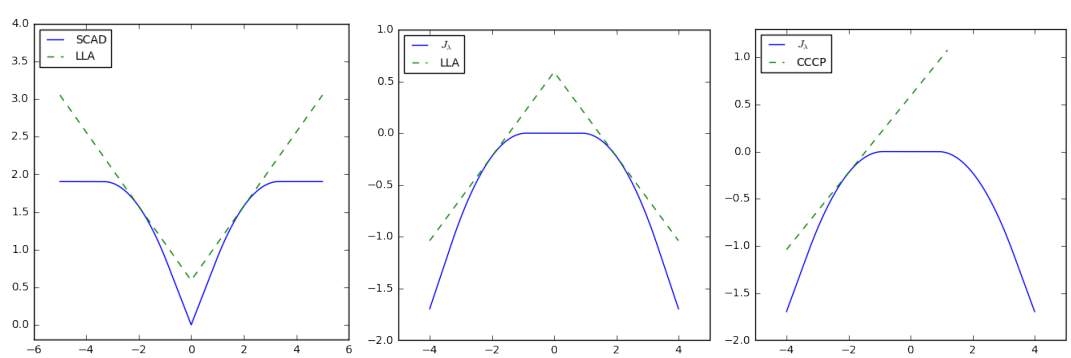
essentially a regularized quadratic programming with L_1 penalty. A smaller tuning parameter $\tau\lambda$ is adopted to increase the estimation accuracy such that we obtain a good initial estimator for Step 2. Typically, we choose τ to be small number such as $\tau = 1/\log n$ or $\tau = \lambda$.

Wang et al. (2013) proposed the two-step SCAD-CCCP (CCCP stands for ConCave Convex procedure) algorithm for high-dimensional linear regression model with the SCAD penalty. This algorithm is based on the observation that the SCAD penalty can be decomposed as the difference of two convex functions, or equivalently, the sum of a convex function and a concave function. More specifically, the SCAD penalty has the following decomposition,

$$P_\lambda(|\beta_j|) = J_\lambda(|\beta_j|) + \lambda|\beta_j|,$$

where $J_\lambda(|\beta_j|)$ is a differentiable concave function of the following form,

$$\begin{aligned} J_\lambda(|\beta_j|) = &- \frac{\beta_j^2 - 2\lambda|\beta_j| + \lambda^2}{2(a-1)} I(\lambda \le |\beta_j| \le a\lambda) \\ &+ \left(\frac{(a+1)^2\lambda^2}{2} - \lambda|\beta_j| \right) I(|\beta_j| > a\lambda). \end{aligned}$$



(a) (b) (c)

Figure 3.1: The solid line in (a) is the SCAD penalty function with $\lambda = 0.9$ and $a = 3.7$ and the dashed line is its local linear approximation at $t = 2$. In plot (b), the solid line is the J_λ function and the dashed line is its local linear approximation; In plot (c), the solid line is the J_λ function and the dashed line is its tight convex upper bound.

Given that β_{j0} is close to β_j, the concave function J_λ can be approximated by its

tight convex upper bound $\text{sign}(\beta_{j0})J'_\lambda(|\beta_{j0}|)\beta_j$. Thus the (3.7) can be approximated by

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} + \sum_{j=1}^{p} \text{sign}(\beta_{j0})J'_\lambda(|\beta_{j0}|)\beta_j + \lambda\|\boldsymbol{\beta}\|_1 \\ &\text{s.t. } \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}, \end{aligned} \tag{3.29}$$

which is a convex optimization problem. In fact, the LLA algorithm and the CCCP algorithm are closely connected to each other. The LLA algorithm essentially approximates the nonconvex penalty function by a clipped linear function (see Figure 3.1(a)). It is easy to see that it is equivalent to approximate the function J_λ by

$$J_\lambda(|\beta_j|) \approx J_\lambda(|\beta_{j0}|) + J'_\lambda(|\beta_{j0}|)(|\beta_j| - |\beta_{j0}|), \tag{3.30}$$

that is, a clipped linear function is used to approximate J_λ as shown in Figure 3.1(b). Instead of approximating the nonconvex function J_λ by a clipped linear function, the CCCP algorithm approximates J_λ by its tight convex upper bound as shown in Figure 3.1(c), which is a linear function.

3.3.2 ADMM algorithm for regularized quadratic programming with linear constraint

Recently, the alternating direction method of multipliers (ADMM) has received intensive attention from a broad spectrum of areas (Boyd et al. 2011, Fang et al. 2015). This algorithm can be applied to solve the regularized quadratic programming after local linear approximation. Given the current solution $\boldsymbol{\beta}^{(k)}$, (3.7) can be approximated by

$$\begin{aligned} &\operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top\boldsymbol{\beta} + \mathbf{w}^\top|\boldsymbol{\beta}| \\ &\text{subject to } \mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}, \end{aligned} \tag{3.31}$$

where $\mathbf{w} = (w_1, \ldots, w_p)^\top = (P'_\lambda(|\beta_1^{(k)}|), \ldots, P'_\lambda(|\beta_p^{(k)}|))^\top$. ADMM algorithm naturally deals with equality constraint, thus we replace inequality constraint in (3.31) with equality constraint by introducing an indicator function,

$$\begin{aligned} &\operatorname*{argmin}_{\boldsymbol{\beta},\mathbf{y}} \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top\boldsymbol{\beta} + \mathbf{w}^\top|\boldsymbol{\beta}| + I(\mathbf{y}) \\ &\text{subject to } \mathbf{C}\boldsymbol{\beta} + \mathbf{y} = \mathbf{b}, \end{aligned} \tag{3.32}$$

where $\mathbf{y} \in \mathbb{R}^r$ and $I(\mathbf{y}) = \sum_{i=1}^r I(y_i)$ with

$$I(y_i) = \begin{cases} 0 & \text{if } y_i \geq 0, \\ +\infty & \text{if } y_i < 0. \end{cases}$$

In ADMM form, (3.32) can be written as

$$\begin{aligned} &\underset{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y}}{\operatorname{argmin}} \ \frac{1}{2}\boldsymbol{\beta}_1^\top \mathbf{W}\boldsymbol{\beta}_1 - \mathbf{q}^\top \boldsymbol{\beta}_1 + \mathbf{w}^\top |\boldsymbol{\beta}_2| + I(\mathbf{y}) \\ &\text{subject to } \mathbf{C}\boldsymbol{\beta}_1 + \mathbf{y} = \mathbf{b}, \ \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2. \end{aligned} \tag{3.33}$$

The augmented Lagrangian for (3.33) is

$$\begin{aligned} L_\rho(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y}, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2) =& \frac{1}{2}\boldsymbol{\beta}_1^\top \mathbf{W}\boldsymbol{\beta}_1 - \mathbf{q}^\top \boldsymbol{\beta}_1 + \mathbf{w}^\top |\boldsymbol{\beta}_2| + I(\mathbf{y}) + \boldsymbol{\gamma}_1^\top (\mathbf{C}\boldsymbol{\beta}_1 + \mathbf{y} - \mathbf{b}) + \\ & \boldsymbol{\gamma}_2^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) + \frac{\rho}{2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 + \frac{\rho}{2}\|\mathbf{C}\boldsymbol{\beta}_1 + \mathbf{y} - \mathbf{b}\|_2^2. \end{aligned}$$

Let $\mathbf{u}_1 = \boldsymbol{\gamma}_1/\rho$, $\mathbf{u}_2 = \boldsymbol{\gamma}_2/\rho$ be the scaled dual variables, we can express the augmented Lagrangian as

$$\begin{aligned} L_\rho(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{y}, \mathbf{u}_1, \mathbf{u}_2) =& \frac{1}{2}\boldsymbol{\beta}_1^\top \mathbf{W}\boldsymbol{\beta}_1 - \mathbf{q}^\top \boldsymbol{\beta}_1 + \mathbf{w}^\top |\boldsymbol{\beta}_2| + \frac{\rho}{2}\|\mathbf{C}\boldsymbol{\beta}_1 + \mathbf{y} - \mathbf{b} + \mathbf{u}_1\|_2^2 + \\ & \frac{\rho}{2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 + \mathbf{u}_2\|_2^2 - \frac{\rho}{2}\|\mathbf{u}_1\|_2^2 - \frac{\rho}{2}\|\mathbf{u}_2\|_2^2 + I(\mathbf{y}). \end{aligned} \tag{3.34}$$

Given the kth iteration $(\boldsymbol{\beta}_1^{(k)}, \boldsymbol{\beta}_2^{(k)}, \mathbf{y}^{(k)}, \mathbf{u}_1^{(k)}, \mathbf{u}_2^{(k)})$, the ADMM algorithm for (3.34) has the following procedures,

$$\begin{aligned} \boldsymbol{\beta}_1^{(k+1)} &= \underset{\boldsymbol{\beta}_1}{\operatorname{argmin}} \ \frac{1}{2}\boldsymbol{\beta}_1^\top \mathbf{W}\boldsymbol{\beta}_1 - \mathbf{q}^\top \boldsymbol{\beta}_1 + \frac{\rho}{2}\|\mathbf{C}\boldsymbol{\beta}_1 + \mathbf{y}^{(k)} - \mathbf{b} + \mathbf{u}_1^{(k)}\|_2^2 + \\ &\qquad \frac{\rho}{2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2^{(k)} + \mathbf{u}_2^{(k)}\|_2^2, \\ \boldsymbol{\beta}_2^{(k+1)} &= \underset{\boldsymbol{\beta}_2}{\operatorname{argmin}} \ \mathbf{w}^\top |\boldsymbol{\beta}_2| + \frac{\rho}{2}\|\boldsymbol{\beta}_1^{(k+1)} - \boldsymbol{\beta}_2 + \mathbf{u}_2^{(k)}\|_2^2, \\ \mathbf{y}^{(k+1)} &= \underset{\mathbf{y}}{\operatorname{argmin}} \ \frac{\rho}{2}\|\mathbf{C}\boldsymbol{\beta}_1^{(k+1)} + \mathbf{y} - \mathbf{b} + \mathbf{u}_1^{(k)}\|_2^2 + I(\mathbf{y}), \\ \mathbf{u}_1^{(k+1)} &= \mathbf{u}_1^{(k)} + \mathbf{C}\boldsymbol{\beta}_1^{(k+1)} + \mathbf{y}^{(k+1)} - \mathbf{b}, \\ \mathbf{u}_2^{(k+1)} &= \mathbf{u}_2^{(k)} + \boldsymbol{\beta}_1^{(k+1)} - \boldsymbol{\beta}_2^{(k+1)}. \end{aligned} \tag{3.35}$$

Fortunately, we can derive the closed forms for the updates of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and $\mathbf{y}$. $\boldsymbol{\beta}_1$-update is similar to a ridge-type regression with the following closed form

$$\boldsymbol{\beta}_1^{(k+1)} = \rho \left(\mathbf{W} + \rho \mathbf{C}^\top \mathbf{C} + \rho \mathbf{I}\right)^{-1} \left(\mathbf{q}/\rho + \mathbf{C}^\top (\mathbf{b} - \mathbf{y}^{(k)} - \mathbf{u}_1^{(k)}) + (\boldsymbol{\beta}_2^{(k)} - \mathbf{u}_2^{(k)})\right).$$

For $\boldsymbol{\beta}_2$-update, soft thresholding can be applied for each entry in $\boldsymbol{\beta}_2$. More specifically, each entry in $\boldsymbol{\beta}_2$ can be updated as follows,

$$\beta_{2j}^{(k+1)} = S_{w_j/\rho}(\beta_{1j}^{(k+1)} + u_{2j}^{(k)}), \quad j = 1, 2, \ldots, p.$$

where S_λ is the soft thresholding and defined to be

$$S_\lambda(a) = (a - \lambda)_+ - (-a - \lambda)_+.$$

The introduced variable $\mathbf{y}$ is updated as follows,

$$\mathbf{y}^{(k+1)} = \max\{\mathbf{0}, \mathbf{C}\boldsymbol{\beta}_1^{(k+1)} + \mathbf{b} - \mathbf{u}_1^{(k)}\},$$

where the max operator is taken element-wisely. This ADMM algorithm for regularized quadratic programming with linear inequality constraint is summarized in Algorithm 2. In many applications, the inequality constraint $\mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}$ in replaced by equality constraint $\mathbf{C}\boldsymbol{\beta} = \mathbf{b}$. In that case, there is no need to introduce the variable $\mathbf{y}$. The ADMM procedures in (3.35) are reduced to

Algorithm 2 ADMM for regularized quadratic programming with linear inequality constraint.

input: $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}, \mathbf{y}^{(0)}, \mathbf{u}_1^{(0)}, \mathbf{u}_2^{(0)}$ and $\rho > 0$.
while not convergent **do**
1. $\boldsymbol{\beta}_1$-update: $\boldsymbol{\beta}_1^{(k+1)} = \mathbf{W}_\rho^{-1}(\mathbf{q}/\rho + \mathbf{C}^\top(\mathbf{b} - \mathbf{y}^{(k)} - \mathbf{u}_1^{(k)}) + (\boldsymbol{\beta}_2^{(k)} - \mathbf{u}_2^{(k)}))$, where $\mathbf{W}_\rho = \rho(\mathbf{W} + \rho\mathbf{C}^\top\mathbf{C} + \rho\mathbf{I})$.
2. $\boldsymbol{\beta}_2$-update: $\beta_{2j}^{(k+1)} = S_{w_j/\rho}(\beta_{1j}^{(k+1)} + u_{2j}^{(k)})$, $j = 1, 2, \ldots, p$.
3. $\mathbf{y}$-update: $\mathbf{y}^{(k+1)} = \max\{\mathbf{0}, \mathbf{C}\boldsymbol{\beta}_1^{(k+1)} + \mathbf{b} - \mathbf{u}_1^{(k)}\}$.
4. $\mathbf{u}_1$-update: $\mathbf{u}_1^{(k+1)} = \mathbf{u}_1^{(k)} + \mathbf{C}\boldsymbol{\beta}_1^{(k+1)} + \mathbf{y}^{(k+1)} - \mathbf{b}$.
5. $\mathbf{u}_2$-update: $\mathbf{u}_2^{(k+1)} = \mathbf{u}_2^{(k)} + \boldsymbol{\beta}_1^{(k+1)} - \boldsymbol{\beta}_2^{(k+1)}$.

end while

$$
\begin{aligned}
\boldsymbol{\beta}_1^{(k+1)} &= \underset{\boldsymbol{\beta}_1}{\operatorname{argmin}} \ \frac{1}{2}\boldsymbol{\beta}_1^\top \mathbf{W}\boldsymbol{\beta}_1 - \mathbf{q}^\top \boldsymbol{\beta}_1 + \frac{\rho}{2}\|\mathbf{C}\boldsymbol{\beta}_1 - \mathbf{b} + \mathbf{u}_1^{(k)}\|_2^2 + \\
&\qquad \frac{\rho}{2}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2^{(k)} + \mathbf{u}_2^{(k)}\|_2^2, \\
\boldsymbol{\beta}_2^{(k+1)} &= \underset{\boldsymbol{\beta}_2}{\operatorname{argmin}} \ \mathbf{w}^\top |\boldsymbol{\beta}_2| + \frac{\rho}{2}\|\boldsymbol{\beta}_1^{(k+1)} - \boldsymbol{\beta}_2 + \mathbf{u}_2^{(k)}\|_2^2, \\
\mathbf{u}_1^{(k+1)} &= \mathbf{u}_1^{(k)} + \mathbf{C}\boldsymbol{\beta}_1^{(k+1)} - \mathbf{b}, \\
\mathbf{u}_2^{(k+1)} &= \mathbf{u}_2^{(k)} + \boldsymbol{\beta}_1^{(k+1)} - \boldsymbol{\beta}_2^{(k+1)}.
\end{aligned}
\tag{3.36}
$$

The ADMM algorithm for regularized quadratic programming with linear equality constraint is summarized in Algorithm 3.

Algorithm 3 ADMM for regularized quadratic programming with linear inequality constraint.

input: $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}, \mathbf{u}_1^{(0)}, \mathbf{u}_2^{(0)}$ and $\rho > 0$.
while not convergent **do**
1. $\boldsymbol{\beta}_1$-update: $\boldsymbol{\beta}_1^{(k+1)} = \mathbf{W}_\rho^{-1}(\mathbf{q}/\rho + \mathbf{C}^\top(\mathbf{b} - \mathbf{u}_1^{(k)}) + (\boldsymbol{\beta}_2^{(k)} - \mathbf{u}_2^{(k)}))$, where $\mathbf{W}_\rho = \rho(\mathbf{W} + \rho\mathbf{C}^\top\mathbf{C} + \rho\mathbf{I})$.
2. $\boldsymbol{\beta}_2$-update: $\beta_{2j}^{(k+1)} = S_{w_j/\rho}(\beta_{1j}^{(k+1)} + u_{2j}^{(k)})$, $j = 1, 2, \ldots, p$.
3. $\mathbf{u}_1$-update: $\mathbf{u}_1^{(k+1)} = \mathbf{u}_1^{(k)} + \mathbf{C}\boldsymbol{\beta}_1^{(k+1)} - \mathbf{b}$.
4. $\mathbf{u}_2$-update: $\mathbf{u}_2^{(k+1)} = \mathbf{u}_2^{(k)} + \boldsymbol{\beta}_1^{(k+1)} - \boldsymbol{\beta}_2^{(k+1)}$.

end while

In chapter 4, we will study an application of the regularized quadratic programming: the estimation of the linear functional, which can be formulated as follows,

$$
\frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}) \text{ subject to } \bar{\mathbf{x}}^\top \boldsymbol{\beta} = 1,
$$

where $\bar{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}}$ are the sample mean and sample covariance matrix respectively. In this application, we have $\mathbf{W} = \widehat{\boldsymbol{\Sigma}}$, $\mathbf{q} = \mathbf{0}$, $\mathbf{C} = \bar{\mathbf{x}}^\top$ and the corresponding dual variable u_1 is a scalar. We summarize the ADMM algorithm for the estimation of linear functional in Algorithm 4.

3.3.3 Choice of tuning parameter λ

The performance of resulting estimator depends on the choice of tuning parameter λ and the optimal choice of λ relies on some unknown parameter. In practice, we

Algorithm 4 ADMM for estimation of linear functional.

input: $\boldsymbol{\beta}_1^{(0)}, \boldsymbol{\beta}_2^{(0)}, u_1^{(0)}, \mathbf{u}_2^{(0)}$ and $\rho > 0$.
while not convergent **do**
1. $\boldsymbol{\beta}_1$-update: $\boldsymbol{\beta}_1^{(k+1)} = \widehat{\boldsymbol{\Sigma}}_\rho^{-1}((1 - u_1^{(k)})\bar{\mathbf{x}} + (\boldsymbol{\beta}_2^{(k)} - \mathbf{u}_2^{(k)}))$,
where $\widehat{\boldsymbol{\Sigma}}_\rho = \rho(\widehat{\boldsymbol{\Sigma}} + \rho\bar{\mathbf{x}}\bar{\mathbf{x}}^\top + \rho\mathbf{I})$.
2. $\boldsymbol{\beta}_2$-update: $\beta_{2j}^{(k+1)} = S_{w_j/\rho}(\beta_{1j}^{(k+1)} + u_{2j}^{(k)})$, $j = 1, 2, \ldots, p$.
3. u_1-update: $u_1^{(k+1)} = u_1^{(k)} + \bar{\mathbf{x}}^\top \boldsymbol{\beta}_1^{(k+1)} - 1$.
4. $\mathbf{u}_2$-update: $\mathbf{u}_2^{(k+1)} = \mathbf{u}_2^{(k)} + \boldsymbol{\beta}_1^{(k+1)} - \boldsymbol{\beta}_2^{(k+1)}$.

end while

consider a sequence of tuning parameters and select the one optimizes some criteria. In the high-dimensional linear regression setting, Wang et al. (2013) proposed the following high-dimensional BIC to choose the tuning parameter,

$$\mathrm{HBIC}(\lambda) = \log(\widehat{\sigma}_\lambda^2) + \|\widehat{\boldsymbol{\beta}}_\lambda\|_0 \frac{C_n \log p}{n}, \tag{3.37}$$

where $\widehat{\sigma}_\lambda^2 = n^{-1}\mathrm{SSE}_\lambda$ with $\mathrm{SSE}_\lambda = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda\|_2^2$, and C_n is a sequence of positive scalars that diverges. The optimal tuning parameter is chosen by minimizing the HBIC criteria

$$\widehat{\lambda} = \operatorname*{argmin}_\lambda \mathrm{HBIC}(\lambda).$$

Motivated by this, we propose BIC-type criterion the estimation of linear functional (3.17) and (3.19). To this end, we replace sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ by $\widehat{\boldsymbol{\Sigma}}_\phi = \widehat{\boldsymbol{\Sigma}} + \phi\mathbf{I}_p$ with a small positive number $\phi = \sqrt{\log p/n}$. Such a perturbation does not noticeably affect the computational accuracy of the final solution and all the theoretical results still hold when $\phi \leq \sqrt{\log p/n}$. For the one-sample problem (3.17), we define

$$\mathrm{SSE}_{\lambda,\phi}^1 = \left\| \frac{\bar{\mathbf{x}}}{\bar{\mathbf{x}}^\top \widehat{\boldsymbol{\Sigma}}_\phi^{-1} \bar{\mathbf{x}}} - \widehat{\boldsymbol{\Sigma}}_\phi \widehat{\boldsymbol{\beta}}_\lambda \right\|_2^2.$$

We propose the following BIC-type criterion for one-sample problem

$$\mathrm{HBIC}_1(\lambda) = \mathrm{SSE}_{\lambda,\phi}^1 + \|\widehat{\boldsymbol{\beta}}_\lambda\|_0 \frac{C_n \log p}{n}.$$

The optimal choice of λ is the one that minimizes $\mathrm{HBIC}_1(\lambda)$. Similarly, for the two-sample problem (3.19), we define

$$\mathrm{SSE}^2_{\lambda,\phi} = \left\| \frac{\bar{\mathbf{x}}_d}{\bar{\mathbf{x}}_d^\top \widehat{\boldsymbol{\Sigma}}_\phi^{-1} \bar{\mathbf{x}}_d} - \widehat{\boldsymbol{\Sigma}}_\phi \widehat{\boldsymbol{\beta}}_\lambda \right\|_2^2,$$

where $\bar{\mathbf{x}}_d = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$. The BIC-type criterion for two-sample problem is given by

$$\mathrm{HBIC}_2(\lambda) = \mathrm{SSE}^2_{\lambda,\phi} + \|\widehat{\boldsymbol{\beta}}_\lambda\|_0 \frac{C_n \log p}{n}.$$

3.4 Simulation studies

3.4.1 Comparison with ridge-type estimator

In this section, we compare the proposed LLA estimator for the linear functional $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ with the ridge-type estimator. We generate a random sample of size n from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = c \cdot (\mathbf{1}_s^\top, \mathbf{0}_{p-s}^\top)^\top$ and $s = 10$. We set c to different values to represent different signal strengths of $\boldsymbol{\mu}$. For $\rho \in (0, 1)$, we consider the following three covariance structures:

(1) Compound symmetry with $\boldsymbol{\Sigma}_1 = (1-\rho)\mathbf{I}_p + \rho \mathbf{1}_p \mathbf{1}_p^\top$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix;

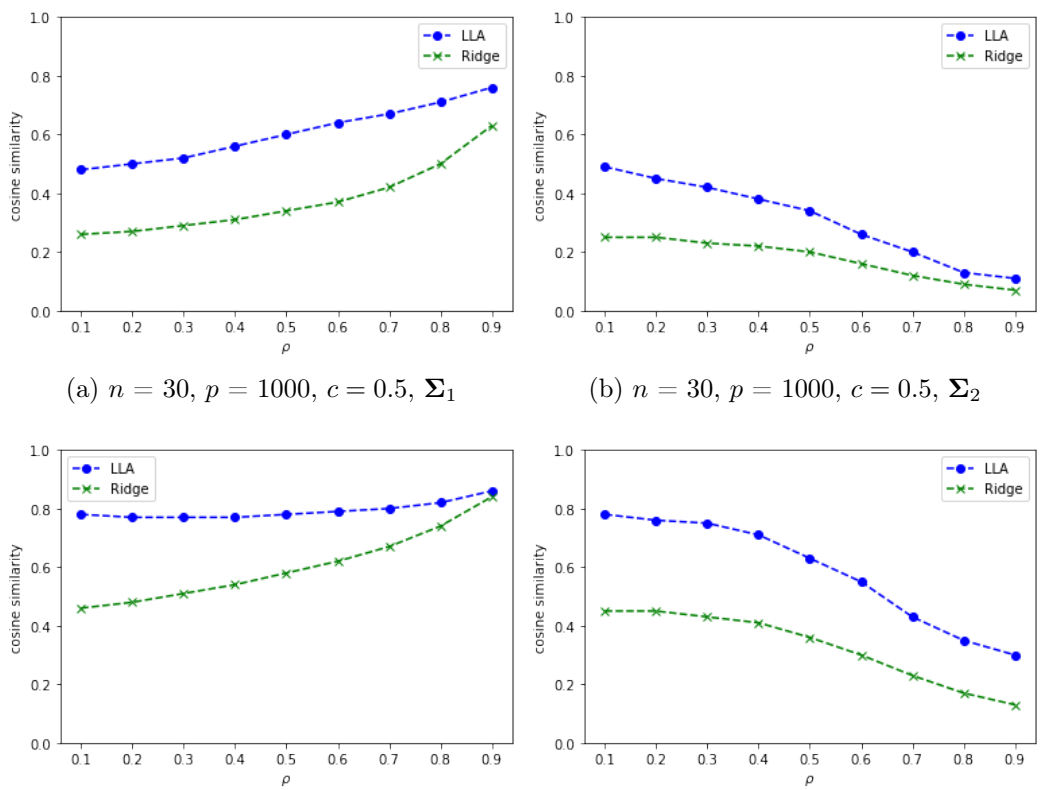
(2) Autocorrelation with $\boldsymbol{\Sigma}_2 = (\rho^{|i-j|})_{i,j}$.

Compound symmetry covariance structure $\boldsymbol{\Sigma}_1$ indicates that any pair of variables $(X_i, X_j), i \neq j$, has equal correlation ρ. It turns out that $\boldsymbol{\Sigma}_1^{-1}$ is an approximately sparse matrix and is diagonally dominant with the off-diagonal entries of order p^{-1}. As a result, the linear functional $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}$ is also approximately sparse in the sense that its first s entries dominate the rest entries. For the autocorrelation covariance matrix, $\boldsymbol{\Sigma}_2$ can be well approximated by a sparse matrix and its inverse $\boldsymbol{\Sigma}_2^{-1}$ is a 3-sparse matrix, and thus the linear functional $\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}$ is sparse too.

We compare the cosine similarity for the LLA estimator and the ridge estimator proposed in Li et al. (2015). The cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$ is defined by

$$\cos\langle \mathbf{u}, \mathbf{v} \rangle = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}.$$

A large value of cosine similarity indicates that the two vectors share similar direction. We set sample size $n = 30$, dimension $p = 1000$, $c = 0.5$ and 1. We let ρ vary from 0.1 to 0.9 with increasement 0.1. For each combination of (n, p, c, ρ), we compute the cosine similarity under both compound symmetry structure and autocorrelation structure. Figure 3.2 shows that the LLA estimator is closer to the true direction than the ridge-type estimator in all the situations. For the compound symmetry structure, Figures 3.2.(a) and 3.2.(c) show that we obtain more accurate estimation as ρ increases. For the autocorrelation structure, Figures 3.2.(b) and 3.2.(d) show that the cosine similarity decreases as ρ increases.



(a) $n = 30$, $p = 1000$, $c = 0.5$, $\mathbf{\Sigma}_1$

(b) $n = 30$, $p = 1000$, $c = 0.5$, $\mathbf{\Sigma}_2$

(c) $n = 30$, $p = 1000$, $c = 1$, $\mathbf{\Sigma}_1$

(d) $n = 30$, $p = 1000$, $c = 1$, $\mathbf{\Sigma}_2$

Figure 3.2: Comparison of cosine similarity for LLA estimator and ridge-type estimator.

3.4.2 Sparse discriminant analysis

In this section, we study the numerical performance of the proposed LLA estimator when it is applied to high-dimensional linear discriminant analysis. Suppose we have two classes with different mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and same covariance matrix $\boldsymbol{\Sigma}$. Without loss of generality, we set the mean vector of the first class $\boldsymbol{\mu}_1$ to be $\mathbf{0}$. Let n_1 and n_2 be the sample size for class 1 and class 2, respectively. The setup of the simulation study is as follows. In all simulations, the number of variables is $p = 1000$, and the sample size of the training and test data is $n_1 = n_2 = 200$ for each class. Each simulation is repeated 100 times to test the stability of the proposed method. We set $\boldsymbol{\mu}_2 = (\mathbf{1}_s^\top, \mathbf{0}_{p-s}^\top)^\top$, where $s = 10$. We consider two difference covariance matrix structure: $\boldsymbol{\Sigma}_1$(compound symmetry) and $\boldsymbol{\Sigma}_2$ (autocorrelation). We use two different ways to choose the penalty parameter λ, represented by LLA-BIC and LLA-CV. We use the HBIC discussed in Section 3.3.3 to choose λ for LLA-BIC and use 5-fold cross-validation to choose λ for LLA-CV. In particular, we compare our LLA classifier with the ROAD classifier (Fan et al. 2012), the FAIR (Fan and Fan 2008), the NSC classifier (Tibshirani et al. 2003) and the oracle classifier. The oracle classifier is defined to be the discriminant rule using true parameter $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

The simulation results for compound symmetry covariance structure with pairwise correlations ranging from 0.1 to 0.9 are shown in Table 3.1. Among all the classifiers, the LLA-BIC has the best performance and its performance is very close to the oracle classifier. The HBIC criterion works slightly better than cross-validation. The LLA-CV and the ROAD have very similar performance. We can see from Table 3.1 that the oracle misclassification rate decreases as ρ increases. The LLA-BIC, the LLA-CV and the ROAD successfully captures the pattern while the classifiers based on independence rule fail to capture this pattern. The misclassification rates of the FAIR and the NSC increase as ρ increases. This huge discrepancy demonstrates that employing the dependence among the variables will boost the classification power. Table 3.2 shows the results for the autocorrelation covariance structure, where the correlation among variables is not as strong as that in compound symmetry structure. When the correlation is not very strong ($\rho \leq 0.5$), the LLA-BIC, the LLA-CV, the ROAD and the FAIR have very similar performance and performs slightly better than the NSC. When ρ is small, $\boldsymbol{\Sigma}_2^{-1}$ can

be well approximated by identity matrix, which explains why those independence rule based classifiers also perform well when ρ is small. As ρ increases, the LLA-CV and the ROAD outperform other classifiers. For the autocorrelation structure, the cross validation criterion performs better than the BIC-type criterion.

Table 3.1: Average of the percentage of misclassification rates for different classifiers under compound symmetry covariance structure.

$n_1 = n_2 = 200, p = 1000$						
ρ	LLA-BIC	LLA-CV	ROAD	FAIR	NSC	Oracle
0.1	6.9 (1.5)	7.3 (1.6)	7.3 (1.4)	12.9 (1.6)	19.4 (9.6)	5.1(1.1)
0.2	5.2 (1.1)	5.5 (1.3)	5.5 (1.2)	17.2 (1.8)	26.1 (11.9)	3.8(1.1)
0.3	4.4 (1.0)	4.6 (0.9)	4.7 (1.1)	20.6 (2.1)	30.5 (12.2)	3.1(0.8)
0.4	3.0 (0.8)	3.1 (0.9)	3.3 (0.9)	23.0 (2.1)	33.2 (12.7)	2.1(0.6)
0.5	2.0 (0.6)	2.1 (0.8)	2.3 (0.8)	25.0 (2.2)	35.1 (12.8)	1.3(0.6)
0.6	1.0 (0.5)	1.1 (0.6)	1.3 (0.7)	26.6 (2.1)	36.5 (12.9)	0.6(0.4)
0.7	0.3 (0.3)	0.4 (0.4)	0.5 (0.4)	27.8 (2.4)	37.8 (12.3)	0.2(0.2)
0.8	0.0 (0.1)	0.2 (0.3)	0.2 (0.3)	29.1 (2.4)	38.6 (12.0)	0.0(0.1)
0.9	0.0 (0.0)	0.2 (0.3)	0.2 (0.3)	30.1 (2.3)	39.5 (11.9)	0.0(0.0)

Table 3.2: Average of the percentage of misclassification rates for different classifiers under autocorrelation covariance structure.

$n_1 = n_2 = 200, p = 1000$						
ρ	LLA-BIC	LLA-CV	ROAD	FAIR	NSC	Oracle
0.1	7.9 (1.4)	8.0 (1.4)	7.9 (1.3)	7.4 (1.4)	10.3 (1.5)	7.2 (1.3)
0.2	10.2 (1.6)	10.2 (1.4)	10.1 (1.4)	9.6 (1.4)	12.2 (1.8)	9.4 (1.4)
0.3	13.0 (1.9)	12.5 (1.7)	12.5 (1.8)	11.9 (1.8)	14.6 (1.9)	11.5 (1.6)
0.4	15.3 (1.9)	14.6 (1.6)	14.7 (1.7)	14.1 (1.7)	16.7 (2.0)	13.1 (1.7)
0.5	18.6 (2.4)	16.9 (1.7)	16.9 (1.7)	16.6 (1.7)	19.3 (1.9)	15.0 (1.7)
0.6	21.8 (3.8)	18.6 (2.3)	18.7 (2.3)	18.9 (2.0)	22.0 (2.5)	16.4 (1.9)
0.7	27.8 (4.3)	20.3 (2.3)	20.4 (2.2)	21.7 (2.2)	24.7 (2.2)	17.3 (1.9)
0.8	32.0 (3.2)	19.5 (2.6)	20.5 (2.6)	24.9 (2.1)	27.7 (2.4)	16.6 (1.9)
0.9	30.2 (2.6)	14.0 (2.0)	17.0 (2.3)	28.2 (2.3)	31.9 (2.8)	11.7 (1.5)

3.5 Proofs

Proof of Theorem 3.1.

Proof. Let $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star$. Since $\boldsymbol{\beta}^\star$ is in the feasible set, the first order necessary condition (3.12) implies that

$$-\widehat{\boldsymbol{\delta}}^\top \mathbf{W}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\delta}}^\top \mathbf{q} - \langle \nabla P_\lambda(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\delta}} \rangle \geq 0. \tag{3.38}$$

By the RSC condition (3.11), we have

$$\widehat{\boldsymbol{\delta}}^\top \mathbf{W}\widehat{\boldsymbol{\delta}} \geq \alpha \|\widehat{\boldsymbol{\delta}}\|_2^2 - \tau \sqrt{\frac{\log p}{n}} \|\widehat{\boldsymbol{\delta}}\|_1. \tag{3.39}$$

Add (3.38) to (3.39), we have

$$-\widehat{\boldsymbol{\delta}}^\top \mathbf{W}\boldsymbol{\beta}^\star + \widehat{\boldsymbol{\delta}}^\top \mathbf{q} - \langle \nabla P_\lambda(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\delta}} \rangle \geq \alpha \|\widehat{\boldsymbol{\delta}}\|_2^2 - \tau \sqrt{\frac{\log p}{n}} \|\widehat{\boldsymbol{\delta}}\|_1. \tag{3.40}$$

Lemma A.1 shows that $P_{\lambda,\mu}(\boldsymbol{\beta}) = P_\lambda(\boldsymbol{\beta}) + \frac{\mu}{2}\|\boldsymbol{\beta}\|_2^2$ is convex, hence

$$P_{\lambda,\mu}(\boldsymbol{\beta}^\star) - P_{\lambda,\mu}(\widehat{\boldsymbol{\beta}}) \geq \langle \nabla P_\lambda(\widehat{\boldsymbol{\beta}}) + \mu\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star - \widehat{\boldsymbol{\beta}} \rangle,$$

which implies

$$-\langle \nabla P_\lambda(\widehat{\boldsymbol{\beta}}), \widehat{\boldsymbol{\delta}} \rangle \leq P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) + \frac{\mu}{2}\|\widehat{\boldsymbol{\delta}}\|_2^2. \tag{3.41}$$

Combining (3.40) and (3.41), we have

$$\alpha \|\widehat{\boldsymbol{\delta}}\|^2 - \tau \sqrt{\frac{\log p}{n}} \|\widehat{\boldsymbol{\delta}}\|_1 \leq -\widehat{\boldsymbol{\delta}}^\top \mathbf{W}\boldsymbol{\beta}^\star + \widehat{\boldsymbol{\delta}}^\top \mathbf{q} + P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) + \frac{\mu}{2}\|\widehat{\boldsymbol{\delta}}\|_2^2.$$

Note that we assume $\boldsymbol{\beta}^\star$ satisfies the equality of the linear constraint, hence $\mathbf{C}\widehat{\boldsymbol{\delta}} = \mathbf{C}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \leq \mathbf{0}$. Let $\boldsymbol{\xi}^\star = \arg\min_{\boldsymbol{\xi} \geq \mathbf{0}} \|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q} + \mathbf{C}^\top \boldsymbol{\xi}\|_\infty$ and we have

$\widehat{\boldsymbol{\delta}}^\top \mathbf{C}^\top \boldsymbol{\xi}^\star \leq 0$. Then we have

$$\alpha\|\widehat{\boldsymbol{\delta}}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\delta}}\|_1 \leq -\widehat{\boldsymbol{\delta}}^\top \mathbf{W}\boldsymbol{\beta}^\star + \widehat{\boldsymbol{\delta}}^\top \mathbf{q} - \widehat{\boldsymbol{\delta}}^\top \mathbf{C}^\top \boldsymbol{\xi}^\star + P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) + \frac{\mu}{2}\|\widehat{\boldsymbol{\delta}}\|_2^2$$

$$(\alpha - \mu/2)\|\widehat{\boldsymbol{\delta}}\|_2^2 \leq P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) + \|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q} + \mathbf{C}^\top \boldsymbol{\xi}^\star\|_\infty \|\widehat{\boldsymbol{\delta}}\|_1 + \tau\sqrt{\frac{\log p}{n}}\|\widehat{\boldsymbol{\delta}}\|_1$$

$$(\alpha - \mu/2)\|\widehat{\boldsymbol{\delta}}\|_2^2 \leq P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) + \left(\|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q} + \mathbf{C}^\top \boldsymbol{\xi}^\star\|_\infty + \tau\sqrt{\frac{\log p}{n}}\right)\|\widehat{\boldsymbol{\delta}}\|_1.$$

By the assumptions of the theorem, we know $\|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q} + \mathbf{C}^\top \boldsymbol{\xi}^\star\|_\infty + \tau\sqrt{\frac{\log p}{n}} \leq \lambda/2$. Hence

$$\begin{aligned}(\alpha - \mu/2)\|\widehat{\boldsymbol{\delta}}\|_2^2 &\leq P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) + \frac{\lambda}{2}\|\widehat{\boldsymbol{\delta}}\|_1 \\ &\leq P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) + \frac{1}{2}P_\lambda(\widehat{\boldsymbol{\delta}}) + \frac{\mu}{4}\|\widehat{\boldsymbol{\delta}}\|_2^2,\end{aligned}$$

where the second inequality is due to the inequality $\frac{\lambda}{2}\|\widehat{\boldsymbol{\delta}}\|_1 \leq \frac{1}{2}P_\lambda(\widehat{\boldsymbol{\delta}}) + \frac{\mu}{4}\|\widehat{\boldsymbol{\delta}}\|_2^2$ by Lemma A.1. By the subadditivity of P_λ, we have $P_\lambda(\widehat{\boldsymbol{\delta}}) = P_\lambda(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \leq P_\lambda(\widehat{\boldsymbol{\beta}}) + P_\lambda(\boldsymbol{\beta}^\star)$. Then

$$\begin{aligned}(\alpha - \mu/2)\,\|\widehat{\boldsymbol{\delta}}\|_2^2 &\leq P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) + \frac{1}{2}P_\lambda(\widehat{\boldsymbol{\beta}}) + \frac{1}{2}P_\lambda(\boldsymbol{\beta}^\star) + \frac{\mu}{4}\|\widehat{\boldsymbol{\delta}}\|_2^2 \\ (\alpha - 3\mu/4)\,\|\widehat{\boldsymbol{\delta}}\|_2^2 &\leq \frac{3}{2}P_\lambda(\boldsymbol{\beta}^\star) - \frac{1}{2}P_\lambda(\widehat{\boldsymbol{\beta}}) \\ (2\alpha - 3\mu/2)\,\|\widehat{\boldsymbol{\delta}}\|_2^2 &\leq 3P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}).\end{aligned}$$

By (iii) in Lemma A.1, we have $3\lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_1 - \lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}^c}\|_1 \geq 3P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) \geq 0$, where $\mathcal{I}$ denotes the index set of the s largest elements of $\widehat{\boldsymbol{\delta}}$ in magnitude. Since $\alpha \geq 3\mu/4$, we have

$$0 \leq (2\alpha - 3\mu/2)\,\|\widehat{\boldsymbol{\delta}}\|_2^2 \leq 3\lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_1 - \lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}^c}\|_1. \tag{3.42}$$

As a result, we have $\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}^c}\|_1 \leq 3\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_1$ and

$$\left(2\alpha - \frac{3}{2}\mu\right)\|\widehat{\boldsymbol{\delta}}\|_2^2 \leq 3\lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_1 - \lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}^c}\|_1 \leq 3\lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_1 \leq 3\lambda\sqrt{s}\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_2,$$

from which we can conclude that

$$\|\widehat{\boldsymbol{\delta}}\|_2 \leq \frac{6\lambda\sqrt{s}}{4\alpha - 3\mu}.$$

The L_1 norm bound follows immediately from the L_2 norm bound

$$\|\widehat{\boldsymbol{\delta}}\|_1 = \|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_1 + \|\widehat{\boldsymbol{\delta}}_{\mathcal{I}^c}\|_1 \leq 4\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_1 \leq 4\sqrt{s}\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_2 \leq \frac{24\lambda s}{4\alpha_1 - 3\mu},$$

which completes the proof of part (i).

For part (ii), note that

$$\widehat{\boldsymbol{\delta}}^\top \mathbf{W}\widehat{\boldsymbol{\delta}} = \langle \mathbf{W}\widehat{\boldsymbol{\beta}} - \mathbf{q} + \mathbf{C}^\top \boldsymbol{\xi}^\star, \widehat{\boldsymbol{\delta}}\rangle - \langle \mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q} + \mathbf{C}^\top \boldsymbol{\xi}^\star, \widehat{\boldsymbol{\delta}}\rangle.$$

Combine the first order condition (3.12), (3.41) and the fact $\widehat{\boldsymbol{\delta}}^\top \mathbf{C}^\top \boldsymbol{\xi}^\star \leq 0$, we have

$$\langle \mathbf{W}\widehat{\boldsymbol{\beta}} - \mathbf{q} + \mathbf{C}^\top \boldsymbol{\xi}^\star, \widehat{\boldsymbol{\delta}}\rangle \leq \langle \mathbf{W}\widehat{\boldsymbol{\beta}} - \mathbf{q}, \widehat{\boldsymbol{\delta}}\rangle \leq P_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\widehat{\boldsymbol{\beta}}) + \frac{\mu}{2}\|\widehat{\boldsymbol{\delta}}\|_2^2. \tag{3.43}$$

Furthermore, Lemma A.1 implies that

$$\begin{aligned}
\langle \mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q} + \mathbf{C}^\top \boldsymbol{\xi}^\star, \widehat{\boldsymbol{\delta}}\rangle &\leq \|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q} + \mathbf{C}^\top \boldsymbol{\xi}^\star\| \cdot \|\widehat{\boldsymbol{\delta}}\|_1 \\
&\leq \frac{1}{2}\left(P_\lambda(\widehat{\boldsymbol{\delta}}) + \frac{\mu}{2}\|\widehat{\boldsymbol{\delta}}\|_2^2\right) \\
&\leq \frac{1}{2}\left(P_\lambda(\widehat{\boldsymbol{\beta}}) + P_\lambda(\boldsymbol{\beta}^\star)\right) + \frac{\mu}{4}\|\widehat{\boldsymbol{\delta}}\|_2^2.
\end{aligned} \tag{3.44}$$

The last inequality is because of the additivity of $P_\lambda(\cdot)$. Combine (3.43) and (3.44), we have

$$\begin{aligned}
\widehat{\boldsymbol{\delta}}^\top \mathbf{W}\widehat{\boldsymbol{\delta}} &\leq \frac{3}{2}P_\lambda(\boldsymbol{\beta}^\star) - \frac{1}{2}P_\lambda(\widehat{\boldsymbol{\beta}}) + \frac{3}{4}\mu\|\widehat{\boldsymbol{\delta}}\|_2^2 \\
&\leq \frac{3}{2}\lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}}\|_1 - \frac{1}{2}\lambda\|\widehat{\boldsymbol{\delta}}_{\mathcal{I}^c}\|_1 + \frac{3}{4}\mu\|\widehat{\boldsymbol{\delta}}\|_2^2 \\
&\leq \frac{3}{2}\sqrt{s}\lambda\|\widehat{\boldsymbol{\delta}}\|_2 + \frac{3}{4}\mu\|\widehat{\boldsymbol{\delta}}\|_2^2 \\
&\leq \lambda^2 s\left(\frac{9}{4\alpha - 3\mu} + \frac{27\mu}{(4\alpha - 3\mu)^2}\right).
\end{aligned}$$

□

Proof of Theorem 3.2.

Proof. (i) We first show the support recovery. Define $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$, $J_\lambda(t) = \lambda|t| - P_\lambda(t)$, $\ell_n(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top \boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta})$, and $\bar{\ell}_n(\boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}) - J_\lambda(\boldsymbol{\beta})$. By Lemma 3.1, we know

$$\langle \nabla \ell_n(\widehat{\boldsymbol{\beta}}) - \nabla \ell_n(\widetilde{\boldsymbol{\beta}}), \boldsymbol{\delta}\rangle \geq \alpha\|\boldsymbol{\delta}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1^k, \ k = 1, 2. \tag{3.45}$$

Lemma A.2 shows that $J_\lambda(t) - \frac{\mu}{2}t^2$ is a differential concave function and thus its derivative $\nabla J_\lambda(t) - \mu t$ is a decreasing function. Therefore,

$$\langle \nabla J_\lambda(\widehat{\boldsymbol{\beta}}) - \mu\widehat{\boldsymbol{\beta}} - (\nabla J_\lambda(\widetilde{\boldsymbol{\beta}}) - \mu\widetilde{\boldsymbol{\beta}}), \boldsymbol{\delta}\rangle \leq 0.$$

Rearranging (3.5), we have

$$\langle -\nabla J_\lambda(\widehat{\boldsymbol{\beta}}) + \nabla J_\lambda(\widetilde{\boldsymbol{\beta}}), \boldsymbol{\delta}\rangle \geq -\mu\|\boldsymbol{\delta}\|_2^2. \tag{3.46}$$

Adding (3.46) to (3.45), we have

$$\langle \nabla \bar{\ell}_n(\widehat{\boldsymbol{\beta}}) - \nabla \bar{\ell}_n(\widetilde{\boldsymbol{\beta}}), \boldsymbol{\delta}\rangle \geq (\alpha - \mu)\|\boldsymbol{\delta}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1^k. \tag{3.47}$$

$\widehat{\boldsymbol{\beta}}$ is a stationary point satisfying the first order condition (3.12) and $\widetilde{\boldsymbol{\beta}}$ is a feasible point by construction, hence we have $\langle \nabla \ell_n(\widehat{\boldsymbol{\beta}}) + \nabla P_\lambda(\widehat{\boldsymbol{\beta}}), \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle \geq 0$. Define $\widehat{\mathbf{z}} \in \partial\|\widehat{\boldsymbol{\beta}}\|_1$ and we have

$$\langle \nabla \bar{\ell}_n(\widehat{\boldsymbol{\beta}}), \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle + \lambda\langle \widehat{\mathbf{z}}, \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle \geq 0, \tag{3.48}$$

since $\nabla \ell_n(\widehat{\boldsymbol{\beta}}) + \nabla P_\lambda(\widehat{\boldsymbol{\beta}}) = \nabla \bar{\ell}_n(\widehat{\boldsymbol{\beta}}) + \lambda\widehat{\mathbf{z}}$. The zero gradient condition (3.16) implies that

$$\langle \bar{\ell}_n(\widetilde{\boldsymbol{\beta}}) + \lambda\widetilde{\mathbf{z}} + \mathbf{C}^\top\boldsymbol{\gamma}, \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle = 0. \tag{3.49}$$

Subtracting (3.49) from (3.48), we have

$$\begin{aligned}\langle \bar{\ell}_n(\widehat{\boldsymbol{\beta}}) - \bar{\ell}_n(\widetilde{\boldsymbol{\beta}}), \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle + \lambda\langle \widehat{\mathbf{z}}, \widetilde{\boldsymbol{\beta}}\rangle - \lambda\|\widehat{\boldsymbol{\beta}}\|_1 + \lambda\langle \widetilde{\mathbf{z}}, \widehat{\boldsymbol{\beta}}\rangle - \lambda\|\widetilde{\boldsymbol{\beta}}\|_1 - \langle \mathbf{C}^\top\boldsymbol{\gamma}, \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle \geq 0 \\ \langle \bar{\ell}_n(\widehat{\boldsymbol{\beta}}) - \bar{\ell}_n(\widetilde{\boldsymbol{\beta}}), \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle + \lambda\langle \widehat{\mathbf{z}}, \widetilde{\boldsymbol{\beta}}\rangle - \lambda\|\widehat{\boldsymbol{\beta}}\|_1 + \lambda\langle \widetilde{\mathbf{z}}, \widehat{\boldsymbol{\beta}}\rangle - \lambda\|\widetilde{\boldsymbol{\beta}}\|_1 \geq 0.\end{aligned} \tag{3.50}$$

The second inequality in (3.50) is because

$$\langle \mathbf{C}^\top\boldsymbol{\gamma}, \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle = \boldsymbol{\gamma}^\top\mathbf{C}\widetilde{\boldsymbol{\beta}} - \boldsymbol{\gamma}^\top\mathbf{C}\widehat{\boldsymbol{\beta}} = \boldsymbol{\gamma}^\top(\mathbf{C}\widetilde{\boldsymbol{\beta}} - \mathbf{b}) - \boldsymbol{\gamma}^\top(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{b}) \geq 0,$$

Rearranging (3.50), we have

$$\begin{aligned}\lambda\|\widehat{\boldsymbol{\beta}}\|_1 - \lambda\langle\widetilde{\mathbf{z}},\widehat{\boldsymbol{\beta}}\rangle &\le \langle\bar{\ell}_n(\widehat{\boldsymbol{\beta}}) - \bar{\ell}_n(\widetilde{\boldsymbol{\beta}}), \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle + \lambda\langle\widehat{\mathbf{z}},\widetilde{\boldsymbol{\beta}}\rangle - \lambda\|\widetilde{\boldsymbol{\beta}}\|_1 \\ &\le \langle\bar{\ell}_n(\widehat{\boldsymbol{\beta}}) - \bar{\ell}_n(\widetilde{\boldsymbol{\beta}}), \widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\rangle \\ &\le \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1^2 - (\alpha-\mu)\|\boldsymbol{\delta}\|_2^k.\end{aligned} \tag{3.51}$$

Set $k = 2$, we have

$$\lambda\|\widehat{\boldsymbol{\beta}}\|_1 - \lambda\langle\widetilde{\mathbf{z}},\widehat{\boldsymbol{\beta}}\rangle \le \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1^2 - (\alpha-\mu)\|\boldsymbol{\delta}\|_2^2. \tag{3.52}$$

We claim that if $\|\widehat{\mathbf{z}}_{\mathcal{A}^c}\|_\infty \le 1-\nu$ for some $\nu \in (0,1]$ and $\lambda \ge \frac{2\tau}{\nu}\sqrt{\frac{\log p}{n}}$, then $\|\boldsymbol{\delta}\|_1 \le (4/\nu+2)\sqrt{s}\|\boldsymbol{\delta}\|_2$. From (3.51) with $k=1$, we know

$$\begin{aligned}(\alpha-\mu)\|\boldsymbol{\delta}\|_2^2 - \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1 &\le \langle\bar{\ell}_n(\widehat{\boldsymbol{\beta}}) - \bar{\ell}_n(\widetilde{\boldsymbol{\beta}}), \boldsymbol{\delta}\rangle \\ &\le \lambda\langle\widehat{\mathbf{z}},\widetilde{\boldsymbol{\beta}}\rangle - \lambda\|\widehat{\boldsymbol{\beta}}\|_1 + \lambda\langle\widetilde{\mathbf{z}},\boldsymbol{\delta}\rangle.\end{aligned} \tag{3.53}$$

We derive the upper bounds for $\lambda\langle\widehat{\mathbf{z}},\widetilde{\boldsymbol{\beta}}\rangle - \lambda\|\widehat{\boldsymbol{\beta}}\|_1$ and $\lambda\langle\widetilde{\mathbf{z}},\boldsymbol{\delta}\rangle$ separately. Note that $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}^c} = \mathbf{0}$,

$$\begin{aligned}\lambda\langle\widehat{\mathbf{z}},\widetilde{\boldsymbol{\beta}}\rangle - \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le \lambda\|\widetilde{\boldsymbol{\beta}}\|_1 - \lambda\|\widehat{\boldsymbol{\beta}}\|_1 &= \lambda\left(\|\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}\|_1 - \|\widehat{\boldsymbol{\beta}}_{\mathcal{A}}\|_1 - \|\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c}\|_1\right) \\ &= \lambda\left(\|\boldsymbol{\delta}_{\mathcal{A}}\|_1 - \|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1\right).\end{aligned} \tag{3.54}$$

For the term $\lambda\langle\widetilde{\mathbf{z}},\boldsymbol{\delta}\rangle$, we have

$$\begin{aligned}\lambda\langle\widetilde{\mathbf{z}},\boldsymbol{\delta}\rangle = \lambda\left(\langle\widetilde{\mathbf{z}}_{\mathcal{A}},\boldsymbol{\delta}_{\mathcal{A}}\rangle + \langle\widetilde{\mathbf{z}}_{\mathcal{A}^c},\boldsymbol{\delta}_{\mathcal{A}^c}\rangle\right) &\le \lambda(\|\widetilde{\mathbf{z}}_{\mathcal{A}}\|_\infty\|\boldsymbol{\delta}_{\mathcal{A}}\|_1 + \|\widetilde{\mathbf{z}}_{\mathcal{A}^c}\|_\infty\|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1) \\ &\le \lambda(\|\boldsymbol{\delta}_{\mathcal{A}}\|_1 + (1-\nu)\|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1).\end{aligned} \tag{3.55}$$

Combining (3.53), (3.54) and (3.55), we have

$$\begin{aligned}(\alpha-\mu)\|\boldsymbol{\delta}\|_2^2-\tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1 &\le \lambda(2(\|\boldsymbol{\delta}_{\mathcal{A}}\|_1-\nu\|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1)\\ -\tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1 &\le \lambda(2(\|\boldsymbol{\delta}_{\mathcal{A}}\|_1-\nu\|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1)\\ -\frac{\nu}{2}\lambda\|\boldsymbol{\delta}\|_1 &\le \lambda(2(\|\boldsymbol{\delta}_{\mathcal{A}}\|_1-\nu\|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1).\\ \frac{\nu}{2}\|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1 &\le (2+\frac{\nu}{2})\|\boldsymbol{\delta}_{\mathcal{A}}\|_1\end{aligned}$$

The third inequality is due to the assumption that $\lambda \ge \frac{2\tau}{\nu}\sqrt{\frac{\log p}{n}}$. Then we have

$$\|\boldsymbol{\delta}\|_1 = \|\boldsymbol{\delta}_{\mathcal{A}}\|_1 + \|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1 \le \|\boldsymbol{\delta}_{\mathcal{A}}\|_1 + (4/\nu+1)\|\boldsymbol{\delta}_{\mathcal{A}}\|_1 \le (4/\nu+2)\sqrt{s}\|\boldsymbol{\delta}\|_2,$$

which proves the claim. As a result, equation (3.52) implies that

$$\begin{aligned}\lambda\|\widehat{\boldsymbol{\beta}}\|_1-\lambda\langle\widetilde{\mathbf{z}},\widehat{\boldsymbol{\beta}}\rangle &\le \tau\sqrt{\frac{\log p}{n}}\|\boldsymbol{\delta}\|_1^2-(\alpha-\mu)\|\boldsymbol{\delta}\|_2^2\\ &\le \tau s\sqrt{\frac{\log p}{n}}\left(\frac{4}{\nu}+2\right)^2\|\boldsymbol{\delta}\|_2^2-(\alpha-\mu)\|\boldsymbol{\delta}\|_2^2.\end{aligned}$$

Since $n \ge \left(\frac{\tau s}{\alpha-\mu}\right)^2\left(\frac{4}{\nu}+2\right)^4\log p$, we have $\lambda\|\widehat{\boldsymbol{\beta}}\|_1-\lambda\langle\widetilde{\mathbf{z}},\widehat{\boldsymbol{\beta}}\rangle \le 0$. On the other hand, the Holder inequality implies that $\lambda\langle\widetilde{\mathbf{z}},\widehat{\boldsymbol{\beta}}\rangle \le \lambda\|\widehat{\boldsymbol{\beta}}\|_1$. Then we must have $\langle\widetilde{\mathbf{z}},\widehat{\boldsymbol{\beta}}\rangle = \|\widehat{\boldsymbol{\beta}}\|_1$. By assumption $\|\widetilde{\mathbf{z}}_{\mathcal{A}^c}\|_\infty < 1$, we conclude that $\widehat{\beta}_j = 0$ for $j \in \mathcal{A}^c$, as claimed.

(ii) By Lemma A.2, we know $\frac{\mu}{2}t^2 - J_\lambda(t)$ is convex and hence $\frac{\mu}{2}\boldsymbol{\beta}_{\mathcal{A}}^\top\boldsymbol{\beta}_{\mathcal{A}} - J_\lambda(\boldsymbol{\beta}_{\mathcal{A}})$ is convex too. Since $\lambda_{\min}(\mathbf{W}_{\mathcal{A}\mathcal{A}}) > \frac{\mu}{2}$, we know $\boldsymbol{\beta}_{\mathcal{A}}^\top\mathbf{W}_{\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} - \frac{\mu}{2}\boldsymbol{\beta}_{\mathcal{A}}^\top\boldsymbol{\beta}_{\mathcal{A}}$ is strictly convex. Therefore $\boldsymbol{\beta}_{\mathcal{A}}^\top\mathbf{W}_{\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} - J_\lambda(\boldsymbol{\beta}_{\mathcal{A}}) = \boldsymbol{\beta}_{\mathcal{A}}^\top\mathbf{W}_{\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} - \lambda\|\boldsymbol{\beta}_{\mathcal{A}}\|_1 + P_\lambda(\boldsymbol{\beta}_{\mathcal{A}})$ is also strictly convex. Consequently, $\boldsymbol{\beta}_{\mathcal{A}}^\top\mathbf{W}_{\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} - \mathbf{q}_{\mathcal{A}}^\top\boldsymbol{\beta}_{\mathcal{A}} + P_\lambda(\boldsymbol{\beta}_{\mathcal{A}})$ is strictly convex. Therefore the solution $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}$ to the restricted program (3.14) is unique. From part (i), we know all stationary points $\widehat{\boldsymbol{\beta}}$ are supported on $\mathcal{A}$ and can be written as $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\mathcal{A}}, \mathbf{0}_{\mathcal{A}^c})$. It is easy to verify that $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ is also a stationary point of the restricted program (3.14). Since the restricted program is strictly convex, the stationary point $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ is unique and so is $\widehat{\boldsymbol{\beta}}$. We know that $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ also satisfies the zero-gradient

condition in (3.15),

$$\mathbf{W}_{\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \mathbf{q}_{\mathcal{A}} + \nabla P_\lambda(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) + \mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma} = \mathbf{0}.$$

Subtracting $\mathbf{W}_{\mathcal{A}\mathcal{A}}\boldsymbol{\beta}^\star_{\mathcal{A}}$ on both sides, we have

$$\begin{aligned}
\mathbf{W}_{\mathcal{A}\mathcal{A}}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}) &= -\mathbf{W}_{\mathcal{A}\mathcal{A}}\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{q}_{\mathcal{A}} - \nabla P_\lambda(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) - \mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma} \\
\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}} &= -\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{q}_{\mathcal{A}} - \mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma}) - \mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}\nabla P_\lambda(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) \\
\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}\|_\infty &\le \|\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma} - \mathbf{q}_{\mathcal{A}})\|_\infty + \|\mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}\nabla P_\lambda(\widehat{\boldsymbol{\beta}}_{\mathcal{A}})\|_\infty \\
&\le \|\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma} - \mathbf{q}_{\mathcal{A}})\|_\infty + \lambda\|\mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty.
\end{aligned}$$

The last inequality is because $\|\nabla P_\lambda(\widehat{\boldsymbol{\beta}}_{\mathcal{A}})\|_\infty \le \lambda$. Furthermore, we have $\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c} = \boldsymbol{\beta}^\star_{\mathcal{A}^c} = \mathbf{0}$ and thus

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_\infty \le \|\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma} - \mathbf{q}_{\mathcal{A}})\|_\infty + \lambda\|\mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty.$$

(iii) From part (ii), we know for any $j \in \mathcal{A}$,

$$|\widehat{\beta}_j - \beta^\star_j| \le \|\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma} - \mathbf{q}_{\mathcal{A}})\|_\infty + \lambda\|\mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty.$$

Since $|\beta^\star_j| \ge \beta^\star_{\min} \ge \lambda(a + \|\mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}\|_\infty) + \|\mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{q}_{\mathcal{A}} + \mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma} - \boldsymbol{\beta}^\star_{\mathcal{A}})\|_\infty$, we have $|\widehat{\beta}_j| \ge a\lambda$ for all $j \in \mathcal{A}$. By condition (vi) in assumption 1, we know $\nabla P_\lambda(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) = \mathbf{0}$ and hence

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_\infty \le \|\boldsymbol{\beta}^\star_{\mathcal{A}} + \mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma} - \mathbf{q}_{\mathcal{A}})\|_\infty.$$

Since $\nabla P_\lambda(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}) = \mathbf{0}$, the zero-gradient condition (3.15) reduces to

$$\mathbf{W}_{\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}} - \mathbf{q}_{\mathcal{A}} + \mathbf{C}_{\mathcal{A}}^\top \boldsymbol{\gamma} = \mathbf{0}.$$

Since the restricted program (3.14) is strictly convex on $\mathbb{R}^{\mathcal{A}}$, this zero-gradient condition implies that $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}$ is the unique global minimum, hence $\widehat{\boldsymbol{\beta}}_{\mathcal{A}} = \widehat{\boldsymbol{\beta}}^{(o)}_{\mathcal{A}}$ and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(o)}$, as claimed.

□

Proof of Theorem 3.3.

Proof. Let $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, then we have $\boldsymbol{\beta}^\star = \gamma\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\beta}} = \widetilde{\gamma}\boldsymbol{\theta}$, where $\gamma = 1/\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$

and $\widetilde{\gamma} = 1/\bar{\mathbf{x}}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. $\widetilde{\gamma}$ is chosen such that $\widetilde{\boldsymbol{\beta}}$ satisfies the equality constraint, i.e., $\bar{\mathbf{x}}^\top \widetilde{\boldsymbol{\beta}} = 1$. According to Theorem 3.1, it suffices to show that with probability at least $1 - cp^{-1}$, we have

$$\|\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\beta}} - \widetilde{\gamma}\bar{\mathbf{x}}\|_\infty = |\widetilde{\gamma}| \cdot \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \bar{\mathbf{x}}\|_\infty \leq Ms\sqrt{\log p/n}.$$

We first deal with $\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \bar{\mathbf{x}}\|_\infty$. By triangle inequality, we have

$$\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \bar{\mathbf{x}}\|_\infty \leq \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \boldsymbol{\mu}\|_\infty + \|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_\infty.$$

Since $\mathbf{x}$ is sub-Gaussian random vector with norm K, we know that $\bar{\mathbf{x}}$ is also sub-Gaussian random vector with norm $K/\sqrt{n}$. Therefore for any $t > 0$, we have

$$\mathrm{P}(\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_\infty > t) \leq 2p \exp\left\{-\frac{nt^2}{cK^2}\right\}.$$

Take $t = M_1\sqrt{\log p/n}$ for some large $M_1 > 0$, we have

$$\mathrm{P}(\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_\infty < M_1\sqrt{\log p/n}) \geq 1 - 2p^{-1}. \tag{3.56}$$

From Lemma B.2, we know that there exists $M_2 > 0$ such that

$$\mathrm{P}(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \leq M_2\sqrt{\log p/n}) \geq 1 - 2p^{-1}.$$

Then with probability at least $1 - 2p^{-1}$, we have

$$\begin{aligned} \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \boldsymbol{\mu}\|_\infty = \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \boldsymbol{\Sigma}\boldsymbol{\theta}\|_\infty &\leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \|\boldsymbol{\theta}\|_1 \\ &\leq \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty s\|\boldsymbol{\theta}\|_\infty \leq M_2 C_2 s\sqrt{\log p/n}. \end{aligned} \tag{3.57}$$

Combine (3.57) and (3.56), we know that with probability at least $1 - 4p^{-1}$, we have

$$\|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \bar{\mathbf{x}}\|_\infty \leq sM'\sqrt{\log p/n},$$

with $M' = M_1 + C_2 M_2$. Next we show that $\widehat{\gamma}$ is bounded from below. Note that

$$|\bar{\mathbf{x}}^\top\boldsymbol{\theta} - \boldsymbol{\mu}^\top\boldsymbol{\theta}| \leq \|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_\infty \|\boldsymbol{\theta}\|_1 \leq M_1\sqrt{\frac{\log p}{n}}\|\boldsymbol{\theta}\|_1 \leq M_1 C_2 s\sqrt{\frac{\log p}{n}}.$$

By assumption, we know $C_2\lambda/C_1 \leq 1-\epsilon$ with the choices of $M = \max\{M_1, M'/(C_1\epsilon)\}$ and $\lambda = sM\sqrt{\log p/n}$. Thus $M_1C_2s\sqrt{\log p/n} \leq (1-\epsilon)C_1$ and $|\bar{\mathbf{x}}^\top\boldsymbol{\theta} - \boldsymbol{\mu}^\top\boldsymbol{\theta}| \leq (1-\epsilon)C_1$. Consequently,

$$|\bar{\mathbf{x}}^\top\boldsymbol{\theta}| \geq ||\boldsymbol{\mu}^\top\boldsymbol{\theta}| - |\bar{\mathbf{x}}^\top\boldsymbol{\theta} - \boldsymbol{\mu}^\top\boldsymbol{\theta}|| \geq |C_1 - (1-\epsilon)C_1| = \epsilon C_1.$$

Thus $|\widehat{\gamma}| = 1/|\bar{\mathbf{x}}^\top\boldsymbol{\theta}| \leq 1/(\epsilon C_1)$. Therefore

$$\|\widehat{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\beta}} - \widehat{\gamma}\bar{\mathbf{x}}\|_\infty = |\widehat{\gamma}| \cdot \|\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \bar{\mathbf{x}}\|_\infty \leq sM\sqrt{\log p/n},$$

According to Theorem 3.1, by the choice of $\lambda = sM\sqrt{\log p/n}$, the L_1 and L_2 error bounds in part (i) hold.

Now we move to part (ii). Note that $\|\boldsymbol{\beta}^\star\|_\infty$ is bounded too since

$$\|\boldsymbol{\beta}^\star\|_\infty = \|\boldsymbol{\theta}\|_\infty/\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \leq C_2/C_1 := C^\star.$$

Again by the triangle inequality, we have

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_k \leq \|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_k + \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_k, \text{ for } k = 1, 2.$$

The first term is bounded by the result in part(i). To bound the second term,

$$\begin{aligned}
\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_k &\leq \left\|\boldsymbol{\theta}\left(\frac{1}{\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} - \frac{1}{\bar{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\right)\right\|_k \\
&\leq \|\boldsymbol{\beta}^\star\|_k \left|\frac{(\bar{\mathbf{x}} - \boldsymbol{\mu})^\top\boldsymbol{\theta}}{\bar{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\right| \\
&\leq \frac{2M_1}{C_1}\|\boldsymbol{\beta}^\star\|_k\|\boldsymbol{\theta}\|_1\sqrt{\frac{\log p}{n}} \\
&\leq 2sM_1C^\star\|\boldsymbol{\beta}^\star\|_k\sqrt{\frac{\log p}{n}},
\end{aligned}$$

with probability at least $1 - 2p^{-1}$. When $k = 1$,

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \leq 2M_1C^{\star 2}s^2\sqrt{\frac{\log p}{n}} = O(s\lambda),$$

When $k = 2$,

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \leq 2M_1 C^{\star 2} s\sqrt{s}\sqrt{\frac{\log p}{n}} = O(\sqrt{s}\lambda),$$

which completes the proof of part(ii).

□

Proof of Corollary 3.1.

Proof. From the proof of Theorem 3.3, we know $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O(\sqrt{s}\lambda)$. By assumption, $\sqrt{s}\lambda/(\kappa\|\boldsymbol{\beta}^\star\|_2) = o(1)$, or equivalently, $\kappa\|\boldsymbol{\beta}^\star\|_2/\sqrt{s}\lambda \to \infty$. We have

$$\frac{\kappa\|\widetilde{\boldsymbol{\beta}}\|_2}{\sqrt{s}\lambda} \geq \frac{\kappa(\|\boldsymbol{\beta}^\star\|_2 - \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2)}{\sqrt{s}\lambda} = \frac{\kappa(\|\boldsymbol{\beta}^\star\|_2 - O(\sqrt{s}\lambda))}{\sqrt{s}\lambda} = \frac{\kappa\|\boldsymbol{\beta}^\star\|_2}{\sqrt{s}\lambda} - O(\kappa) \to \infty,$$

or equivalently, $\sqrt{s}\lambda/(\kappa\|\widetilde{\boldsymbol{\beta}}\|_2) = o(1)$. Note that $\cos\langle\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star\rangle = \cos\langle\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}\rangle$, we only need to show $\cos\langle\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}\rangle \to 1$. $\cos\langle\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}\rangle$ can be decomposed as follows,

$$\cos\langle\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}\rangle = \frac{\widehat{\boldsymbol{\beta}}^\top\widetilde{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2} = \frac{\widehat{\boldsymbol{\beta}}^\top\widetilde{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2} - \frac{\widetilde{\boldsymbol{\beta}}^\top\widetilde{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2} + \frac{\widetilde{\boldsymbol{\beta}}^\top\widetilde{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2}.$$

On one hand,

$$\begin{aligned}
\frac{\|\widetilde{\boldsymbol{\beta}}\|_2 - \|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2} \leq \frac{\|\widehat{\boldsymbol{\beta}}\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2} &\leq \frac{\|\widetilde{\boldsymbol{\beta}}\|_2 + \|\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2} \\
1 - O\left(\frac{\sqrt{s}\lambda}{\|\widetilde{\boldsymbol{\beta}}\|_2}\right) \leq \frac{\|\widehat{\boldsymbol{\beta}}\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2} &\leq 1 + O\left(\frac{\sqrt{s}\lambda}{\|\widetilde{\boldsymbol{\beta}}\|_2}\right) \\
1 - o(1) \leq \frac{\|\widehat{\boldsymbol{\beta}}\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2} &\leq 1 + o(1)
\end{aligned} \tag{3.58}$$

As a result, we know that $\frac{\|\widehat{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2}{\widetilde{\boldsymbol{\beta}}^\top\widetilde{\boldsymbol{\beta}}} = \frac{\|\widehat{\boldsymbol{\beta}}\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2} \to 1$. One the other hand,

$$\begin{aligned}
\frac{\widehat{\boldsymbol{\beta}}^\top\widetilde{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2} - \frac{\widetilde{\boldsymbol{\beta}}^\top\widetilde{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2} &= \frac{(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^\top\widetilde{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2} \leq \frac{\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2}{\|\widehat{\boldsymbol{\beta}}\|_2\|\widetilde{\boldsymbol{\beta}}\|_2} \\
&\leq \frac{\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2}{\|\widehat{\boldsymbol{\beta}}\|_2} = \frac{\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\|_2}{\|\widetilde{\boldsymbol{\beta}}\|_2}\frac{\|\widetilde{\boldsymbol{\beta}}\|_2}{\|\widehat{\boldsymbol{\beta}}\|_2} \\
&= O(\sqrt{s}\lambda/(\kappa\|\widetilde{\boldsymbol{\beta}}\|_2)) = o(1).
\end{aligned} \tag{3.59}$$

Combine (3.58) and (3.59), we know $\cos\langle\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star\rangle = \cos\langle\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}\rangle \to 1$. □

Proof of Theorem 3.7.

Proof. Without loss of generosity, we assume $\mathrm{E}(\mathbf{x}_i) = \mathbf{0}$. We first show with probability at least $1 - 2p^{-1}$, we have

$$\left\| \frac{1}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^\star) \right\|_\infty \leq M\sqrt{\log p/n}.$$

Note that $\frac{1}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^\star) = \frac{1}{n}\mathbf{X}^\top\boldsymbol{\varepsilon} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i$. Since $\|X_{ij}\|_{\psi_2} \leq K_2$, $\|\varepsilon_i\|_{\psi_2} \leq K_1$, we have $\|X_{ij}\varepsilon_i\|_{\psi_1} \leq K$ where $K = K_1K_2$, then

$$\mathrm{P}\left(\left| \frac{1}{n}\sum_{i=1}^n X_{ij}\varepsilon_i \right| > t \right) \leq 2\exp\left\{ -c_1 n \left(\frac{t^2}{K^2} \wedge \frac{t}{K} \right) \right\},$$

where c_1 is some absolute constant. By the union bound inequality,

$$\mathrm{P}\left(\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i \right\|_\infty > t \right) \leq 2p\exp\left\{ -c_1 n \left(\frac{t^2}{K^2} \wedge \frac{t}{K} \right) \right\}. \tag{3.60}$$

Take $t = M\sqrt{\log p/n}$ with $M \geq 2K/c_1 \vee \sqrt{2K^2/c_1}$, we have

$$\mathrm{P}\left(\left\| \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\varepsilon_i \right\|_\infty > t \right) \leq 2p^{-1}. \tag{3.61}$$

Therefore, $\mathrm{P}(\|\frac{1}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^\star)\|_\infty \leq M\sqrt{\log p/n}) \geq 1 - 2p^{-1}$. By Theorem 3.1, we know part (i) holds. For part (ii),

$$\|\widehat{\mathbf{y}} - \mathbf{y}^\star\|_2 = \sqrt{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)^\top\mathbf{X}^\top\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)} = O(\sqrt{ns}\lambda/\kappa).$$

by Theorem 3.1.

□

Chapter 4
Sparse Online Projection Test

4.1 Introduction

One-sample mean vector test or two-sample test on the equality of two means is a fundamental problem in high-dimensional statistics. These tests are commonly encountered in genome-wide association studies. For instance, Chen and Qin (2010) performed a hypothesis testing to identify sets of genes which are significant with respect to certain treatments in a genetics research. Xu et al. (2016) applied various tests to the bipolar disorder dataset from a genome-wide association study collected by Consortium (2007) in which one would like to test whether there is any association between a disease and a large number of genetic variants. In these applications, the dimension of the data p is often much larger than the sample size n. Traditional methods such as Hotelling's T^2 test (Hotelling 1931) either cannot be directly applied or have low power against the alternative. The Hotelling's T^2 requires that the sample covariance matrix is invertible and this is typically not true in high-dimensional setting where $p > n$. Despite the singularity of the sample covariance matrix, it has been observed that the power of the Hotelling's T^2 test can be adversely affected even when $p < n$, if the sample covariance matrix is nearly singular, see Bai and Saranadasa (1996) and Pan and Zhou (2011).

Several tests for high-dimensional data have been proposed recently. These tests can be roughly classified into three types. The first type is known as the sum-of-squares-type test or modified Hotelling's T^2 test. These tests simply replace the sample covariance matrix by some diagonal matrix such as identity matrix, leading to a sum-of-squares-type test statistic, see Bai and Saranadasa (1996) and

Chen and Qin (2010). To get rid of the unit effect, Srivastava and Du (2008) suggested replacing the sample covariance matrix by its diagonal matrix. These tests are typically asymptoticly normally distributed under null hypothesis. The second type is the maximum-type test. The idea is performing p individual tests and choose the most significant one as the final test statistic. For example, Cai et al. (2014) introduced a test that is based on a linear transformation of the data by the precision matrix which incorporates the correlations among the variables. Such a maximum-type test converges to a certain extreme value distribution. The third type test is the projection test. The idea is to project the high-dimensional vector $\mathbf{x}$ onto a space of low dimension and then traditional methods such as Hotelling's T^2 can be applied. Lauter (1996) proposed a test which projects high-dimensional data to a one-dimensional space by some weight vector. Lopes et al. (2011) proposed a test based on random projection where the the entries in the projection matrix are randomly generated from standard normal distribution. Instead of using random projection, Li et al. (2015) proposed a projection test based on the optimal projection direction.

Different types of tests are powerful only against certain alternatives. For example, if the true mean is dense in the sense that there is a large proportion of small to moderate nonzero components, then sum-of-squares-type test is more powerful. In contrast, if the true mean is sparse in the sense that there are only few nonzero components in the mean, the maximum-type test is more powerful. In practice, since the true alternative hypothesis is unknown, it is unclear how to choose a powerful test. Furthermore, there are some intermediate situations in which neither type of test is powerful (Xu et al. 2016).

Consider a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a p-dimensional population $\mathbf{x}$ with finite mean $\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$ and positive definite covariance matrix $\mathrm{cov}(\mathbf{x}) = \boldsymbol{\Sigma}$. Of interest is to test the following hypothesis

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0, \tag{4.1}$$

for some known vector $\boldsymbol{\mu}_0$. Without loss of generality, we assume $\boldsymbol{\mu}_0 = \mathbf{0}$ and the one-sample problem (4.1) becomes

$$H_0 : \boldsymbol{\mu} = \mathbf{0} \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}. \tag{4.2}$$

In most cases, the test statistic constructed for one-sample problem can be easily extended to two-sample problem. For this reason, we only focus on the one sample problem (4.2) and assume $\boldsymbol{\mu}_0 = \mathbf{0}$. Let $\bar{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}}$ be the sample mean vector and the sample covariance matrix respectively,

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top. \tag{4.3}$$

When $p < n$, the Hotelling's T^2 test statistic for problem (4.2) is given by

$$T^2 = n\bar{\mathbf{x}}^\top \widehat{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{x}}.$$

Under the normality assumption, it is well known that the test statistic $(n-p)/((n-1)p)T^2 \sim F_{p,n-p}(n\zeta)$, which is a non-central F-distribution with degrees of freedom $(p, n-p)$ and non-centrality parameter $n\zeta$, where $\zeta = \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. Without the normality assumption, T^2 asymptotically follows a χ^2-distribution with degree of freedom 1 as $n \to \infty$ with p being fixed. When $p \geq n$, the Hotelling's T^2 is not well defined since the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is not invertible. The idea of projection test is to project the high-dimensional vector $\mathbf{x}_i$ onto a space of low dimension and then traditional methods such as t-test or Hotelling's T^2 can be applied. Let $\mathbf{P}$ be a $p \times k$ matrix with $k \ll n$ and we can project the p-dimensional vector $\mathbf{x}_i$ to a k-dimensional space by left-multiplying the matrix $\mathbf{P}^\top$. More specifically, define $\mathbf{y}_i = \mathbf{P}^\top\mathbf{x}_i, i = 1, \ldots, n$, and thus $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^k$ are independent and identically distributed with mean $\mathbf{P}^\top\boldsymbol{\mu}$ and covariance matrix $\mathbf{P}^\top\boldsymbol{\Sigma}\mathbf{P}$. The Hotelling's $T^2_{\mathbf{P}}$ after projection is defined to be

$$T^2_{\mathbf{P}} = n\bar{\mathbf{x}}^\top\mathbf{P}(\mathbf{P}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{P})^{-1}\mathbf{P}^\top\bar{\mathbf{x}},$$

which is equivalent to the Hotelling's T^2 test based on $\mathbf{y}_1, \ldots, \mathbf{y}_n$.

Several methods have been proposed to determine the projection matrix $\mathbf{P}$. Lauter (1996) considered a test using the linear score $\mathbf{z} = (Z_1, \ldots, Z_n)^\top = \mathbf{X}\mathbf{d}$, where $\mathbf{d}$ is a $p \times 1$ projection vector depending on $\mathbf{X}$ only through $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{d} \neq \mathbf{0}$ with probability 1. Then one can perform the one-sample t-test based on $Z_1, \ldots, Z_n$. Lauter (1996) also proposed two different ways to obtain the projection vector $\mathbf{d}$.

For example, $\mathbf{d}$ can take the form of

$$\mathbf{d} = (\mathrm{diag}(\mathbf{X}^\top \mathbf{X}))^{-1/2},$$

or be the eigenvector corresponding to the largest eigenvalue $\lambda_{\max}$ for the following eigenvalue problem

$$(\mathbf{X}^\top \mathbf{X})\mathbf{d} = \mathrm{diag}(\mathbf{X}^\top \mathbf{X})\mathbf{d}\lambda_{\max}.$$

Lopes et al. (2011) proposed a random projection test where the entries in $\mathbf{P}$ are randomly drawn from the standard normal distribution. This random projection test is an exact test if $\mathbf{x}_i$'s are normally distributed. Instead of using random projection, Li et al. (2015) proposed a projection test using the optimal projection direction. Li et al. (2015) showed that the optimal choice of k in $\mathbf{P}$ is 1 and the optimal projection direction is $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ in the sense that the power of the test $T^2_{\mathbf{P}}$ is maximized. Let $y_i = \boldsymbol{\theta}^\top \mathbf{x}_i, i = 1, \ldots, n$. The projection test statistic is

$$T^2_{\boldsymbol{\theta}} = n\bar{\mathbf{x}}^\top \boldsymbol{\theta}(\boldsymbol{\theta}^\top \widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta})^{-1}\boldsymbol{\theta}^\top \bar{\mathbf{x}},$$

which follows $F_{1,n-1}$ distribution under H_0. It is equivalent to a one-sample t test based on $y_1, \ldots, y_n$. In order to control the type I error, Li et al. (2015) proposed a data-splitting strategy to estimate the optimal direction and obtain an exact t-test. They partition the random sample into two separate sets: $\mathcal{D}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}\}$ and $\mathcal{D}_2 = \{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n\}$. They use $\mathcal{D}_1$ to estimate the direction $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and use $\mathcal{D}_2$ to construct the test statistic $T^2_{\boldsymbol{\theta}}$. To estimate $\boldsymbol{\theta}$, they proposed a ridge-type estimator $\widehat{\boldsymbol{\theta}} = (\mathbf{S}_1 + \lambda \mathbf{D}_1)^{-1}\bar{\mathbf{x}}_1$, where $\bar{\mathbf{x}}_1$ and $\widehat{\boldsymbol{\Sigma}}_1$ are the sample mean vector and the sample covariance matrix computed from $\mathcal{D}_1$, $\mathbf{D}_1 = \mathrm{diag}(\widehat{\boldsymbol{\Sigma}}_1)$ is the diagonal matrix of $\widehat{\boldsymbol{\Sigma}}_1$. The test statistic $T^2_{\widehat{\boldsymbol{\theta}}}$ is constructed using $\widehat{\boldsymbol{\theta}}^\top \mathbf{x}_{n_1+1}, \ldots, \widehat{\boldsymbol{\theta}}^\top \mathbf{x}_n$. Li et al. (2015) also derived the asymptotic power function of the projection test $T^2_{\widehat{\boldsymbol{\theta}}}$ under the assumption that $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ in probability. However, there is no guarantee that the ridge-type estimator is consistent. In order to obtain a better estimation of $\boldsymbol{\theta}$, we assume the optimal projection direction is sparse. Under the sparsity assumption, we estimate $\boldsymbol{\theta}$ using regularized quadratic programming and it can be shown that the resulting estimator is consistent.

4.2 Sparse online projection test

4.2.1 Sparse projection test with data splitting

Let $\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and $\boldsymbol{\beta}^\star = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}/\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ be the optimal projection direction. We assume $\boldsymbol{\theta}$ and $\boldsymbol{\beta}^\star$ are sparse and have s nonzero elements, i.e., $\|\boldsymbol{\theta}\|_0 = \|\boldsymbol{\beta}^\star\|_0 = s$. Under the assumption that the optimal projection direction is sparse, we propose a new estimation vis the following regularized quadratic programming with nonconvex penalty and linear constraint,

$$\widehat{\boldsymbol{\beta}} = \underset{\bar{\mathbf{x}}^\top\boldsymbol{\beta}=1}{\arg\min}\ \frac{1}{2}\boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}). \tag{4.4}$$

We may follow the data splitting procedure proposed in Li et al. (2015). Given the dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, we partition the dataset into $\mathcal{D}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}\}$ and $\mathcal{D}_2 = \{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n\}$, where $n_1 = \lfloor \tau n \rfloor$. We use the first subset to estimate the optimal projection direction by the regularized quadratic programming

$$\widehat{\boldsymbol{\beta}} = \underset{\bar{\mathbf{x}}_1^\top\boldsymbol{\beta}=1}{\arg\min}\ \frac{1}{2}\boldsymbol{\beta}^\top\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \tag{4.5}$$

where $\bar{\mathbf{x}}_1$ and $\widehat{\boldsymbol{\Sigma}}_1$ are the sample mean and sample covariance matrix estimated from $\mathcal{D}_1$. Then we project the high-dimensional data in $\mathcal{D}_2$ to a 1-dimensional space using $\widehat{\boldsymbol{\beta}}$, i.e., $y_i = \mathbf{x}_i^\top\widehat{\boldsymbol{\beta}}$ for $i = n_1 + 1, \ldots, n$. Note that the estimated projection direction $\widehat{\boldsymbol{\beta}}$ is independent from $\mathcal{D}_2$ and thus $y_{n_1+1}, \ldots, y_n$ are independent and identically distributed with mean 0 and variance $\widehat{\boldsymbol{\beta}}^\top\boldsymbol{\Sigma}\widehat{\boldsymbol{\beta}}$ under H_0. As a result, we can perform one-sample t-test based on the new dataset $\{y_{n_1+1}, \ldots, y_n\}$. The advantage of the data splitting procedure is that we achieve an exact t-test and the type I error can be well controlled. However, we perform the t-test only using the data from $\mathcal{D}_2$ and the data in $\mathcal{D}_1$ is discarded, which may lead to some loss in power. Define

$$\bar{y} = \frac{1}{n_2}\sum_{i=n_1+1}^{n} y_i,\ \ s_y^2 = \frac{1}{n_2 - 1}\sum_{i=n_1+1}^{n}(y_i - \bar{y})^2,$$

where $n_2 = n - n_1$ This sparse projection test with data splitting procedure is summarized in Algorithm 5 and is referred to as SPT-DS. The ridge projection test with data splitting procedure in Li et al. (2015) is referred to as RPT-DS. Based on

the empirical study in Li et al. (2015), one may set $\tau \in (0.4, 0.6)$ to achieve high power.

Algorithm 5 Sparse projection test with data splitting

Input: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\tau \in (0, 1)$.

1. Partition data $\mathcal{D}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}\}$ and $\mathcal{D}_2 = \{\mathbf{x}_{n_1+1}, \ldots, \mathbf{x}_n\}$ where $n_1 = \lfloor \tau n \rfloor$.
2. Compute sample mean $\bar{\mathbf{x}}_1$ and sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_1$ based on $\mathcal{D}_1$.
3. Solve
$$\widehat{\boldsymbol{\beta}} = \arg\min_{\bar{\mathbf{x}}_1^\top \boldsymbol{\beta} = 1} \frac{1}{2} \boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}_1 \boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}).$$
4. Compute $y_{n_1+1}, \ldots, y_n$ where $y_i = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}, i = n_1 + 1, \ldots, n$.
5. Compute the test statistic $T_y = \sqrt{n - n_1} \bar{y} / s_y$.
6. Reject H_0 whenever $|T_y| > z_{\alpha/2}$.

We also derive the asymptotic power function for the proposed SPT-DS.

Proposition 4.1. *Suppose the conditions in Theorem 3.3 hold. Let $\widehat{\boldsymbol{\beta}}$ be a stationary point of program (4.5) with $\lambda = Ms\sqrt{\log p / n_1}$. Assume that $s^{3/2}\sqrt{\log p / n_1} = o(1)$ and both n_1 and n_2 go to infinity, we have*

$$\beta_1(\boldsymbol{\mu}) - \Phi(-z_{\alpha/2} + \sqrt{n_2}\sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}) \to 0,$$

$\beta_1(\boldsymbol{\mu}) = P(|T_y| > z_{\alpha/2})$ is the power function for the SPT-DS and $\Phi(\cdot)$ is the cdf for the standard normal distribution.

Unlike RPT-DS which assumes that the ridge-type estimator is consistent, we can show our estimator is consistent under the conditions of Proposition 4.1 and thus the power function we derive is reliable. With a sample of size n, one may expect the power function is of order $\Phi(-z_{\alpha/2} + \sqrt{n}\sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}})$. With the choice of $n_1 = \lfloor \tau n \rfloor$, the power of the projection test with data splitting becomes $\Phi(-z_{\alpha/2} + \sqrt{(1-\tau)n}\sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}})$ and is less powerful than the test using the whole dataset.

4.2.2 Sparse online projection test

To improve the power of data splitting projection test, we propose a sparse projection test via an online framework. Pretending the data arrives one by one in a temporal

manner (though it is not necessary), we keep updating the estimated projection direction when new observations arrive. Suppose we have $\mathbf{x}_1, \ldots, \mathbf{x}_t$ at time t and we use the current t data points to estimate the optimal projection direction

$$\widehat{\boldsymbol{\beta}}_t = \arg\min_{\bar{\mathbf{x}}_t^\top \boldsymbol{\beta}=1} \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}_t \boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}), \tag{4.6}$$

where $\bar{\mathbf{x}}_t$ and $\widehat{\boldsymbol{\Sigma}}_t$ are the sample mean and sample covariance matrix estimated from $\{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$. When a new data point $\mathbf{x}_{t+1}$ arrives, we project the new data point to a scalar by $y_{t+1} = \mathbf{x}_{t+1}^\top \widehat{\boldsymbol{\beta}}_t$ and update the estimation of $\boldsymbol{\beta}^\star$ by (4.6) with $\bar{\mathbf{x}}_{t+1}$ and $\widehat{\boldsymbol{\Sigma}}_{t+1}$. Given an integer $k_n \ll n$, we can obtain an initial estimator $\widehat{\boldsymbol{\beta}}_{k_n}$ based on the first k_n observations. As a result, we obtain a sequence of new data points $\{y_{k_n+1}, \ldots, y_n\}$ and we can carry out one-sample test based on $\{y_{k_n+1}, \ldots, y_n\}$. This sparse online projection test is summarized in Algorithm 6. Since this algorithm updates $\widehat{\boldsymbol{\beta}}_t$ whenever we have a new observation and we refer it as Sparse Online Projection Test - One by one (SOPT-O). This algorithm can be computationally expensive especially when n is large because we need to solve a regularized quadratic programming whenever a new observation arrives. To reduce the computational burden, we also propose a mini-batch version of Algorithm 6. We update the estimated projection direction only when a batch of observations arrive. More specifically, suppose we have $\mathbf{x}_1, \ldots, \mathbf{x}_t$ at time t and obtain $\widehat{\boldsymbol{\beta}}_t$ using $\mathbf{x}_1, \ldots, \mathbf{x}_t$. When the next b observations $\mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+b}$ arrive, we project the b observations to 1-dimensional space by multiplying $\widehat{\boldsymbol{\beta}}_t$, i.e., $y_{t+1} = \mathbf{x}_{t+1}^\top \widehat{\boldsymbol{\beta}}_t, \ldots, y_{t+b} = \mathbf{x}_{t+b}^\top \widehat{\boldsymbol{\beta}}_t$. Then we update the estimation of $\boldsymbol{\beta}^\star$ based on $\{\mathbf{x}_1, \ldots, \mathbf{x}_{t+b}\}$. This mini-batch version is summarized in Algorithm 7 and the corresponding test is referred to as Sparse Online Projection Test - mini Batch (SOPT-B).

Remark. In Algorithm 6 and Algorithm 7, instead of using the regularized quadratic programming to estimate $\boldsymbol{\beta}^\star$, one may also use the ridge-type estimator. In order to achieve stable numerical performance, we also standardize the ridge-type estimator by $\widehat{\boldsymbol{\beta}}_t \leftarrow \widehat{\boldsymbol{\beta}}_t / \bar{\mathbf{x}}_t^\top \widehat{\boldsymbol{\beta}}_t$, i.e., $\widehat{\boldsymbol{\beta}}_t$ satisfies the linear constraint $\bar{\mathbf{x}}_t^\top \widehat{\boldsymbol{\beta}}_t = 1$. The corresponding tests are referred to as Ridge Online Projection Test - one by one (ROPT-O) and Ridge Online Projection Test - mini Batch (ROPT-B). Note that the data splitting projection test can be regarded as a special case of the mini-batch online projection test where there is only one single batch.

Algorithm 6 Sparse online projection test - one by one

Input: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and integer $k_n \geq 2$.
for $t = k_n + 1$ **to** n **do**
1. Compute sample mean $\bar{\mathbf{x}}_{t-1}$ and sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_{t-1}$ based on $\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}$.
2. Update $\widehat{\boldsymbol{\beta}}_{t-1}$ by

$$\widehat{\boldsymbol{\beta}}_{t-1} = \underset{\boldsymbol{\beta}^\top \bar{\mathbf{x}}_{t-1}=1}{\arg\min} \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}_{t-1}\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}).$$

3. Compute $y_t = \mathbf{x}_t^\top \widehat{\boldsymbol{\beta}}_{t-1}$.

end for
Compute the test statistic $T_y = \sqrt{n - k_n}\bar{y}/s_y$.
Reject H_0 if $|T_y| > z_{\alpha/2}$.

Algorithm 7 Sparse online projection test - mini batch

Input: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, integer $k_n \geq 2$ and batch size b.
for $t = 0$ **to** $B = \lfloor (n - k_n)/b \rfloor$ **do**
1. Compute sample mean $\bar{\mathbf{x}}_t$ and sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_t$ based on $\mathbf{x}_1, \ldots, \mathbf{x}_{k_n+bt}$.
2. Update $\widehat{\boldsymbol{\beta}}_t$ by

$$\widehat{\boldsymbol{\beta}}_t = \underset{\boldsymbol{\beta}^\top \bar{\mathbf{x}}_t=1}{\arg\min} \frac{1}{2}\boldsymbol{\beta}^\top \widehat{\boldsymbol{\Sigma}}_t\boldsymbol{\beta} + P_\lambda(\boldsymbol{\beta}).$$

if $t < B$ **then**
Compute $y_{k_n+tb+i} = \mathbf{x}_{k_n+tb+i}^\top \widehat{\boldsymbol{\beta}}_t$ for $i = 1, \ldots, b$.
else
Compute $y_{k_n+bB+i} = \mathbf{x}_{k_n+bB+i}^\top \widehat{\boldsymbol{\beta}}_t$ for $i = 1, \ldots, n - k_n - bB$.
end if
end for
Compute the test statistic $T_y = \sqrt{n - k_n}\bar{y}/s_y$.
Reject H_0 if $|T_y| > z_{\alpha/2}$.

4.2.3 Asymptotic normality for sparse online projection test

In this section, we establish the asymptotic normality of the proposed test statistic under both null hypothesis and alternative hypothesis. Let z_t be the centralized version of y_t, i.e., $z_t = (\mathbf{x}_t - \boldsymbol{\mu})^\top \widehat{\boldsymbol{\beta}}_{t-1}$ for $t = k_n+1, \ldots, n$. Note that $\{z_{k_n+1}, \ldots, z_n\}$

is actually a martingale difference since

$$\begin{aligned}\mathrm{E}(z_{t+1}|z_t, z_{t-1}, \dots) &= \mathrm{E}((\mathbf{x}_{t+1} - \boldsymbol{\mu})^\top \widehat{\boldsymbol{\beta}}_t | z_t, z_{t-1}, \dots) \\ &= \mathrm{E}((\mathbf{x}_{t+1} - \boldsymbol{\mu})^\top)\mathrm{E}(\widehat{\boldsymbol{\beta}}_t | z_t, z_{t-1}, \dots) = 0.\end{aligned}$$

The second equality is due to the fact that the new observation $\mathbf{x}_{t+1}$ is independent from $\widehat{\boldsymbol{\beta}}_t$. Similarly, we know that $\{y_{k_n+1}, \dots, y_n\}$ is also a martingale difference under H_0. Next we establish the asymptotic null distribution using the technique of central limit theorem for martingale difference. Let $\bar{z}$ and s_z^2 be the sample mean and sample variance based on $\{z_{k_n+1}, \dots, z_n\}$, i.e.,

$$\bar{z} = \frac{1}{n - k_n} \sum_{t=k_n+1}^{n} z_t \text{ and } s_z^2 = \frac{1}{n - k_n} \sum_{t=k_n+1}^{n} z_t^2.$$

Theorem 4.1. *Suppose the conditions in Theorem 3.3 hold. Let $\widehat{\boldsymbol{\beta}}_t$ be a stationary point of program (4.6) with $\lambda = Ms\sqrt{\log p/t}$. Assume that $s\sqrt{\log p}/n^{\frac{\tau}{6}} = o(1)$ for some $\tau \in (0, 1)$, then with the choice of $k_n = n^\tau$, we have*

(i) (Normality under alternative) $T_z = \sqrt{n - k_n}\bar{z}/s_z \to N(0, 1)$.

(ii) (Power function) Further assume $\|\boldsymbol{\mu}\|_\infty s^2 \sqrt{\frac{\log p}{n^\tau}} = o(1)$, then

$$\beta_2(\boldsymbol{\mu}) - \Phi(-z_{\alpha/2} + \sqrt{n}\sqrt{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}) \to 0,$$

where $\beta_2(\boldsymbol{\mu}) = P(|T_y| \geq z_\alpha)$ is the power function for the proposed SOPT-O.

Theorem 4.1 establishes the normality of the proposed test statistic under the alternative $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \geq C_1$ for some $C_1 > 0$ (see conditions in Theorem 3.3), and based on which we derive the power function. Clearly, we have $\beta_2(\boldsymbol{\mu}) > \beta_1(\boldsymbol{\mu})$. The SOPT-O improves the performance of SPT-DS in the following two ways: (1) SOPT-O keeps updating the estimation of $\boldsymbol{\beta}^\star$ and obtains a more and more accurate estimation of $\boldsymbol{\beta}^\star$ as more observations arrive. (2) Less observations is discarded in SOPT-O than SPT-DS when performing the test. For SOPT-O, only $k_n = o(n)$ data points is discarded and is negligible when deriving the asymptotic power function. However, the results in Theorem 4.1 does not hold for $H_0 : \boldsymbol{\mu} = \mathbf{0}$ since the condition $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \geq C_1$ is not satisfied. In order to establish the central

limit theorem for martingale difference, according to Lemme D.1, we need to show s_z^2 converge to some positive constant. Under the conditions of Theorem 4.1, we can show that the estimated direction $\widehat{\boldsymbol{\beta}}_t$ at time t is close enough to the true direction $\boldsymbol{\beta}^\star$ for $t = k_n + 1, \ldots, n$ and thus s_y^2 converges to $(\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu})^{-1}$. Thanks to the fact that $\widehat{\boldsymbol{\beta}}_t$ does not converge to $\boldsymbol{\beta}^\star = \mathbf{0}$, otherwise the the variance s_y^2 would diverge and consequently the martingale difference central limit theorem does not hold. Actually, this also explains why we do not use the unconstrained version to estimate the projection direction. In the unconstrained version, we dot not require $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \geq C_1$ and thus we have $\widehat{\boldsymbol{\beta}}_t \to \mathbf{0}$ under H_0, which is problematic. In fact, the linear equality constraint $\widehat{\boldsymbol{\beta}}_t^\top \bar{\mathbf{x}}_t = 1$ forces that $\widehat{\boldsymbol{\beta}}_t$ cannot be too close to $\mathbf{0}$. We establish the normality under H_0 assuming that $\widehat{\boldsymbol{\beta}}_t$ lies in the neighborhood of some nonzero constant vector. We have the following theorem.

Theorem 4.2. *(Normality under null) Let $\widehat{\boldsymbol{\beta}}_t$ be a stationary point of program (4.6) with $\lambda = Ms\sqrt{\log p/t}$ for some large $M > 0$. Under H_0, assume that there exists $\boldsymbol{\beta}_0 \neq \mathbf{0} \in \mathbb{R}^p$ such that $\|\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_0\|_1 = O(a_n)$ for some sequence a_n and all $t \geq k_n$. If $\|\boldsymbol{\beta}_0\|_2^2 \sqrt{\frac{\log p}{n}} = o(1)$ and $a_n \log p(\|\boldsymbol{\beta}_0\|_1 \vee 1) = o(1)$, then we have*

$$\frac{\sqrt{n - k_n}\bar{y}}{s_y} \to N(0, 1).$$

The proof of Theorem 4.2 is the same as the proof of part (i) in Theorem 4.1 by substituting $\boldsymbol{\beta}_0$ for $\boldsymbol{\beta}^\star$ and thus is omitted. Theorem 4.2 shows that the asymptotic null distribution is the standard normal distribution. Thus we reject the null hypothesis (4.1) if and only if $|T_y| \geq z_{\alpha/2}$, where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of standard normal distribution. We also numerically examine the normality under both H_0 and H_1. We consider two covariance matrix structures: (1) Autocorrelation structure where $\boldsymbol{\Sigma} = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$ and (2) compound symmetry structure where $\boldsymbol{\Sigma} = (\sigma_{ij})$ with $\sigma_{ij} = \rho$ if $i \neq j$ and $\sigma_{ij} = 1$ if $i = j$. We set $n = 160$, $p = 400, 1600$ and $\rho = 0.5$. Figure 4.1 shows the distributions of the test statistic T_y under the null hypothesis $\boldsymbol{\mu} = \mathbf{0}$. The upper two panels are the histograms for autocorrelation structure and the lower two panels are the histograms for compound symmetry structure. The red curves are the probability density function of standard normal distribution. Figure 4.1 shows that the distribution of the test statistic

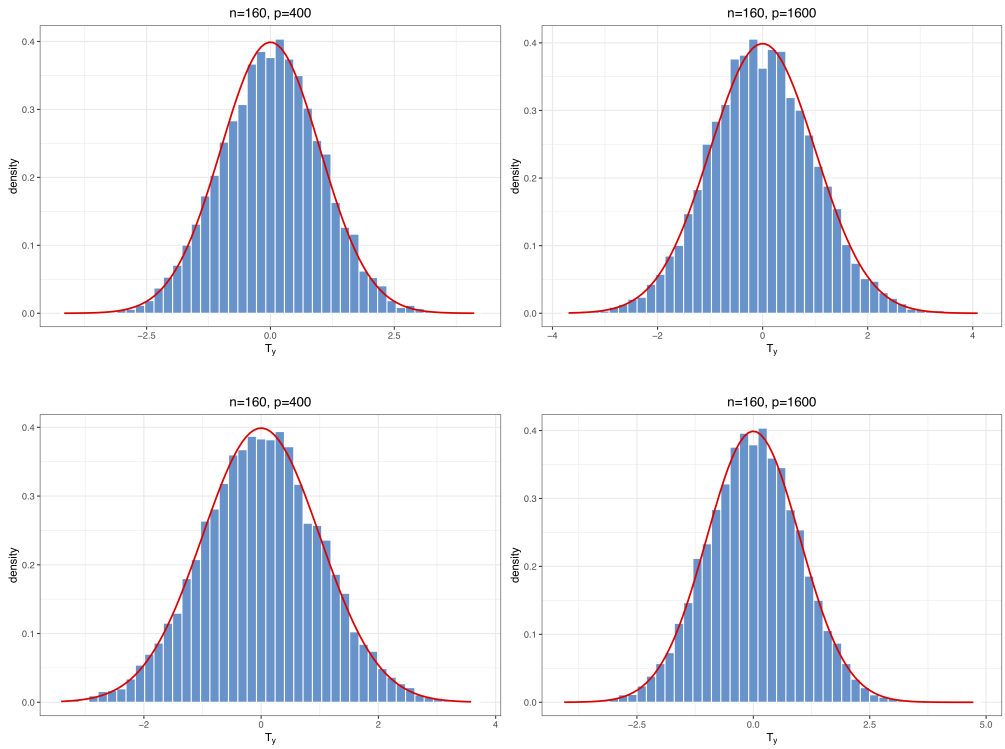


Figure 4.1: The distribution of test statistic T_y under H_0. We set $n = 160, \rho = 0.5$ and consider $p = 400, 1600$. The upper two panels are the histograms for autocorrelation structure and the lower two panels are the histograms for compound symmetry structure. The red curves are the probability density function of standard normal distribution.

is very close to standard normal distribution under H_0 when n, p are relatively large. Figure 4.2 shows the distributions of the statistic T_z under the alternative hypothesis $\boldsymbol{\mu} = 0.5(\mathbf{1}_{10}, \mathbf{0}_{p-10})$. It shows that the distribution of the T_z is very close to standard normal distribution under H_1 as claimed in Theorem 4.1

4.3 Numerical studies

4.3.1 Choice of k_n

According to Algorithm 6, we use the first $k_n = n^\tau$ data to obtain an initial estimate of the optimal projection direction $\boldsymbol{\beta}^\star$ and update the estimate when new data arrives. In order to obtain a good initial estimate of $\boldsymbol{\beta}^\star$, a large value of k_n is

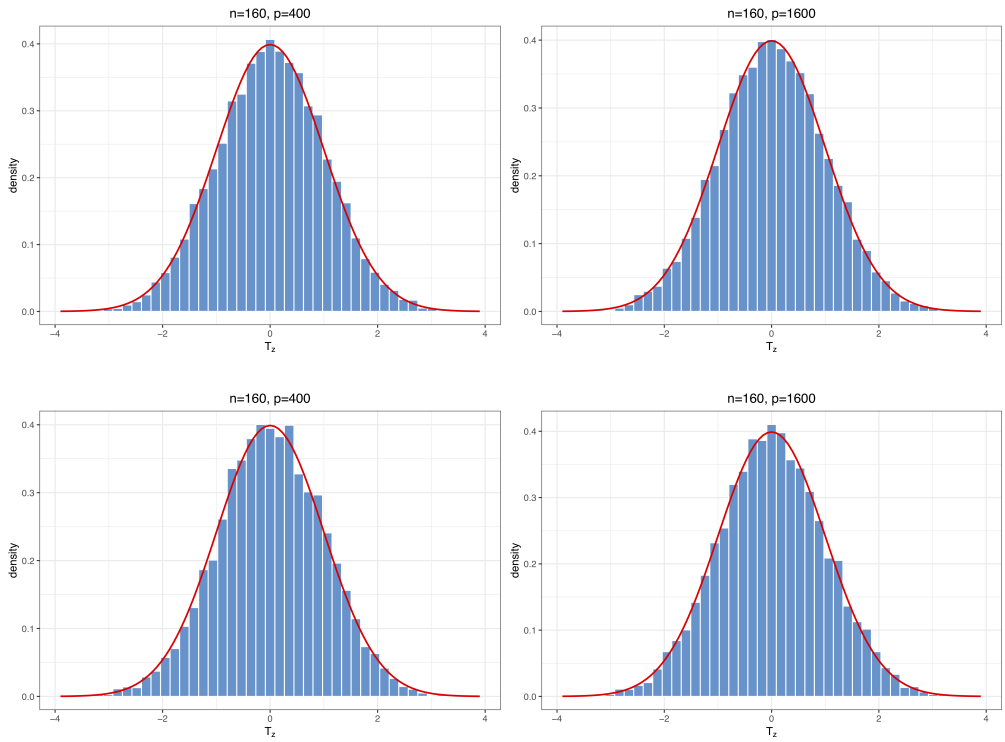


Figure 4.2: The distribution of test statistic T_z under H_1. We set $n = 160, \rho = 0.5$ and consider $p = 400, 1600$. The upper two panels are the histograms for autocorrelation structure and the lower two panels are the histograms for compound symmetry structure. The red curves are the probability density function of standard normal distribution.

preferred. However, the value of k_n cannot be too large since the first k_n data points are discarded when we perform the online projection test. A large choice of k_n may lead to a significant loss in power. In this section, we examine how the choice of k_n or equivalently the choice of parameter τ affect the power of the proposed online projection test. We consider $k_n = n^\tau$ with $\tau = (0.2, 0.25, \ldots, 0.95)$. We set $(n, p, c) = (100, 1600, 0.25)$ and $(40, 1600, 0.5)$ for both autocorrelation and compound symmetry covariance matrix structures with $\rho \in \{0.25, 0.50, 0.75, 0.95\}$. Figure 4.3 depicts the power against the choice of τ. The upper two panels are the power for autocorrelation structure and the lower two panels are the power for compound symmetry structure. Figure 4.3 shows that the power of the test increases very slowly as τ increases and then drops quickly as τ further increases. In other words, the power of the online projection test is not very sensitive to the

choice of τ when τ is relatively small ($\tau \leq 0.8$). When k_n is large, we perform the online projection test only based on a dataset with a relatively small sample size and expect to see some loss in power. In practice, we suggest choosing $\tau \in [0.4, 0.8]$ and we simply set $\tau = 0.6$ in the rest of this chapter.

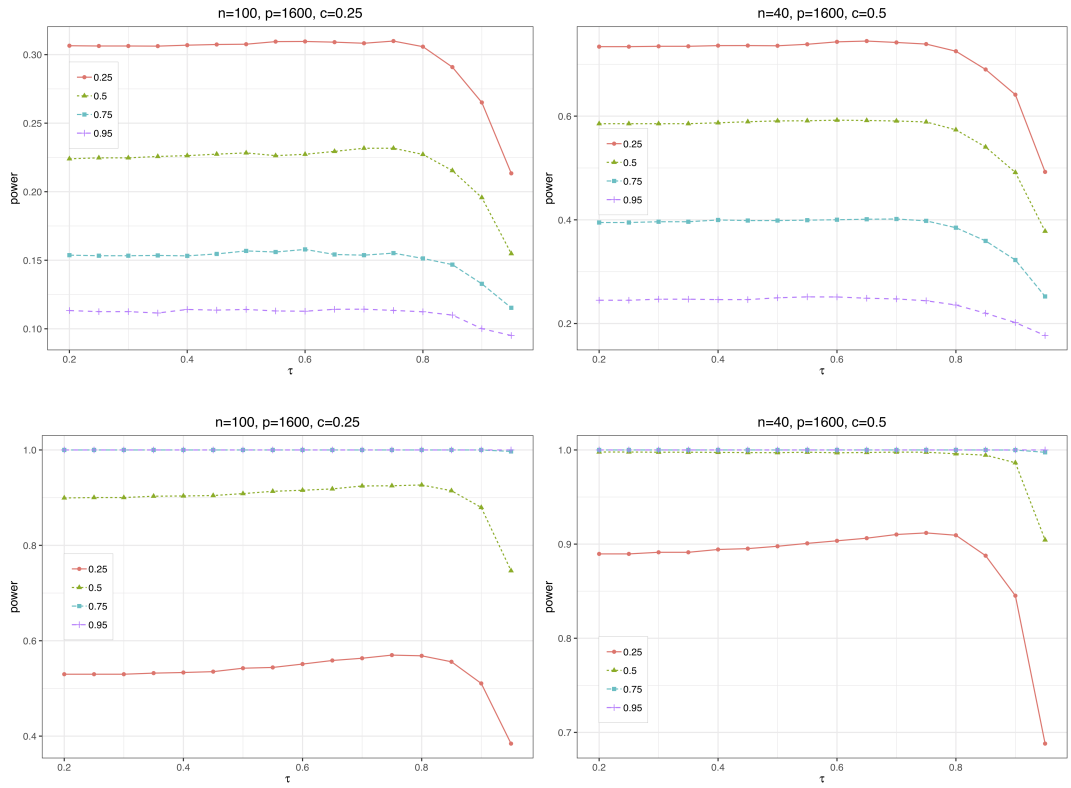


Figure 4.3: Power of the online projection test against the choice of τ. We set $(n, p, c) = (100, 1600, 0.25)$ and $(40, 1600, 0.5)$ for both autocorrelation and compound symmetry covariance matrix structure with $\rho \in \{0.25, 0.50, 0.75, 0.95\}$. The upper two panels are the power for autocorrelation structure and the lower two panels are the power for compound symmetry structure.

4.3.2 Size and power comparison for multivariate normal distribution

In this section, we conduct numerical studies to examine the finite sample performance of different tests for one-sample high-dimensional mean vector problem. These methods include the proposed projection tests using the online framework,

data splitting projectio tests, the sum-of-squares-type test, the maximum-type test as well as the random projection test. We consider two methods to estimate the optimal projection direction: the proposed sparse estimator using regularized quadratic programming with linear constraint and the ridge-type estimator proposed in Li et al. (2015). For ease of presentation, we use the following notations to denote the proposed online projection tests.

- SOPT-O: Sparse Online Projection Test with One-by-one update. The projection direction is estimated by regularized quadratic programming and is updated according to Algorithm 6.
- SOPT-B: Sparse Online Projection Test with mini-Batch update. The projection direction is estimated by regularized quadratic programming and is updated according to Algorithm 7. The batch size is set to be 10.
- ROPT-O: Ridge Online Projection Test with One-by-one update. The projection direction is estimated by the ridge estimator and is updated in a one-by-one manner.
- ROPT-B: Ridge Online Projection Test with mini-Batch update. The projection direction is estimated by the ridge estimator and is updated in a mini-batch manner.
- SPT-DS: Sparse Projection Test with Data Splitting. The projection direction is estimated by regularized quadratic programming with data splitting.
- SPT-DS: Ridge Projection Test with Data Splitting. The projection direction is estimated by the ridge estimator, i.e., the test proposed in Li et al. (2015).

The sum-of-squares-type tests include D1958 test (Dempster 1958), BS1996 test (Bai and Saranadasa 1996), CQ2010 test (Chen and Qin 2010). Srivastava and Du (2008) considered two versions of their test, one with modification and one without modification, and we denote them by SD2008w and SD2008wo, respectively. The maximum-type test we compare with is the one from Cai et al. (2014) without data transformation. We also include two other projection tests L1996 test (Lauter 1996) and LWJ2011 test (Lopes et al. 2011).

We generate a random sample of size n from $N(c\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (\mathbf{1}_s^\top, \mathbf{0}_{p-s}^\top)^\top$ and $s = 10$. We set $c = 0, 0.5$ and 1 to examine the Type I error rate and the power of these tests. For $\rho \in (0, 1)$, we consider the following two covariance structures:

(1) Compound symmetry with $\boldsymbol{\Sigma}_1 = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^\top$.

(2) Autocorrelation with $\boldsymbol{\Sigma}_2 = (\rho^{|i-j|})_{i,j}$.

We consider $\rho = 0.25, 0.5, 0.75$ and 0.95 to examine the influence of correlation on the power of these tests. We set sample size $n = 160, 40$ and dimension $p = 400, 1600$. For SOPT-O, SOPT-B, ROPT-O and ROPT-B, we set $k_n = \lfloor n^\tau \rfloor$ with $\tau = 0.6$, where $\lfloor \cdot \rfloor$ is the rounding operator. For SPT-DS and RPT-DS, we split the data set by setting $n_1 = \lfloor n\tau \rfloor$ with $\tau = 0.4$. To this end, we replace sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ by $\widehat{\boldsymbol{\Sigma}}_\phi = \widehat{\boldsymbol{\Sigma}} + \phi\mathbf{I}_p$ with a small positive number $\phi = \sqrt{\log p/n}$. Such a perturbation does not noticeably affect the computational accuracy of the final solution and all the theoretical properties hold when $\phi \leq \sqrt{\log p/n}$. We set the type I error rate $\alpha = 0.05$ and use $z_{\alpha/2}$ as the critical value. All simulation results are based on 10,000 independent replicates. Tables 4.1 and 4.2 summarize the type I error rate and power for compound symmetry structure when $n = 160$ and $n = 40$, respectively. Tables 4.3 and 4.4 summarize the type I error rate and power for autocorrelation structure when $n = 160$ and $n = 40$, respectively. In order to obtain a good estimate of $\boldsymbol{\beta}^\star$, we assume the optimal projection direction $\boldsymbol{\beta}^\star$ is sparse. When $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$, the compound symmetry structure, $\boldsymbol{\Sigma}_1^{-1}$ is an approximately sparse matrix in the sense that the off-diagonal entries are of order p^{-1} and are dominated by its diagonal entires. The corresponding optimal projection correlation $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is also approximately sparse in the sense that the first s entries dominate the rest entries. When $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_2$, the autocorrelation structure, $\boldsymbol{\Sigma}^{-1}$ is a 3-sparse matrix, meaning that only the diagonal and the first off-diagonal entries are nonzero, and optimal projection direction $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is sparse too.

We first examine the type I error rate. Among all these tests, SPT-DS, RPT-DS, L1996 test and LJW2011 test are exact tests under the normality assumption and thus control the type I error rate very well. All the sum-of-squares-type tests and our proposed online projection tests (SOPT-O, SOPT-B, ROPT-O and ROPT-B) have asymptotic normal distribution under H_0. These online projection tests also keep the type I error very well when $n = 160$ and may have slightly higher type I error rate than 0.05 when $n = 40$. We also observe the same phenomenon for

Table 4.1: Size and power comparison for $N(c\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ (values are in percentage).

	$c = 0$				$c = 0.5$				$c = 1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
					$n = 160$, $p = 400$							
SOPT-O	5.02	5.48	5.28	5.24	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SOPT-B	5.00	5.38	5.15	5.26	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
ROPT-O	5.01	4.87	5.16	5.28	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
ROPT-B	5.07	5.02	5.10	5.33	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SPT-DS	4.77	5.10	4.96	4.83	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
RPT-DS	4.97	4.89	4.80	4.99	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
D1958	6.34	5.32	5.10	5.42	87.48	20.32	12.16	9.70	100.0	100.0	99.94	89.00
BS1996	5.64	5.24	5.32	5.98	81.46	20.16	12.64	10.60	100.0	100.0	99.98	92.76
CQ2010	5.64	5.24	5.32	5.98	81.44	20.16	12.64	10.60	100.0	100.0	99.98	92.80
SD2008w	2.49	0.68	0.09	0.01	33.48	2.46	0.45	0.08	100.0	73.10	4.11	0.64
SD2008wo	5.98	5.60	5.49	5.70	83.70	21.07	12.63	10.83	100.0	100.0	99.96	92.77
L1996	5.32	4.84	4.86	5.36	5.36	5.72	5.12	5.06	7.34	6.06	5.78	5.40
LJW2011	4.98	5.42	4.30	5.32	98.26	100.0	100.0	100.0	100.0	100.0	100.0	100.0
CLX2013	4.84	3.02	0.82	0.12	100.0	100.0	100.0	99.80	100.0	100.0	100.0	100.0
					$n = 160$, $p = 1600$							
SOPT-O	5.22	5.20	4.65	5.12	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SOPT-B	5.14	5.45	5.08	4.76	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
ROPT-O	4.91	4.62	4.90	4.87	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
ROPT-B	4.76	4.65	4.82	5.00	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SPT-DS	4.63	4.99	4.79	4.96	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
RPT-DS	4.91	5.43	5.40	4.74	98.84	99.92	100.0	100.0	100.0	100.0	100.0	100.0
D1958	6.14	5.58	5.28	4.94	11.14	6.88	6.44	6.04	92.62	19.50	12.18	9.94
BS1996	5.46	5.54	5.56	5.38	10.02	6.76	6.84	6.46	87.52	19.44	12.94	10.64
CQ2010	5.46	5.54	5.54	5.38	10.02	6.78	6.82	6.46	87.54	19.44	12.96	10.64
SD2008w	1.56	0.22	0.00	0.01	2.71	0.22	0.03	0.00	17.53	0.58	0.03	0.01
SD2008wo	5.71	5.32	5.42	5.41	10.80	7.29	6.95	6.62	90.39	20.01	13.01	10.78
L1996.	5.30	5.26	5.18	4.92	4.98	4.44	5.18	5.26	5.26	5.36	5.46	5.42
LJW2011	5.24	5.80	5.08	5.32	34.28	56.34	91.70	100.0	98.26	99.94	100.0	100.0
CLX2013	5.64	3.04	0.80	0.12	100.0	100.0	99.94	99.44	100.0	100.0	100.0	100.0

D1958 test, BS1996 test, CQ2010 test and SD2008wo. All these tests tend to have slightly higher type I error rate than the online projection tests. SD2008w does not control the type I error well and is very sensitive to the correlation level ρ. The maximum-type test CLX2013 test converges to a type I extreme-value distribution under H_0 and may suffer from a slower convergence rate. It turns out that CLX2013 test cannot control the type I error and is also sensitive to the correlation level ρ. The type I error rate of CLX2013 test decreases as ρ increases and can be much greater than the pre-specified level α when ρ is small. For example, the type I error rate of CLX2013 test can be as large as 23.06% in the setting of autocorrelation with $(n, p) = (40, 1600)$.

Table 4.2: Size and power comparison for $N(c\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ (values are in percentage).

	$c = 0$				$c = 0.5$				$c = 1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
					$n = 40$, $p = 400$							
SOPT-O	5.92	5.49	5.31	5.23	97.51	99.96	100.0	100.0	100.0	100.0	100.0	100.0
SOPT-B	5.71	5.73	5.61	5.73	91.38	98.46	99.89	99.99	100.0	100.0	100.0	100.0
ROPT-O	5.53	5.23	5.35	5.24	80.76	96.92	99.99	100.0	100.0	100.0	100.0	100.0
ROPT-B	5.41	5.81	5.04	5.54	66.37	85.71	98.65	100.0	99.98	99.95	100.0	100.0
SPT-DS	4.98	4.50	4.94	5.19	71.53	89.92	99.00	99.96	99.97	99.88	99.99	100.0
RPT-DS	5.16	4.47	4.88	4.90	50.22	70.74	94.04	100.0	98.61	99.53	100.0	100.0
D1958	6.36	5.74	5.20	4.96	13.14	8.10	6.86	6.28	81.56	23.18	13.56	10.60
BS1996	5.74	5.86	5.84	5.82	11.82	8.22	7.60	7.22	77.98	23.38	15.18	12.30
CQ2010	5.72	5.86	5.82	5.82	11.82	8.24	7.58	7.22	77.98	23.40	15.22	12.28
SD2008w	3.29	1.11	0.29	0.09	6.65	1.56	0.37	0.11	45.52	4.40	0.76	0.17
SD2008wo	7.63	6.96	6.47	6.22	15.24	9.77	8.17	7.47	86.63	27.57	16.20	12.74
L1996	4.70	4.76	4.74	4.72	4.74	4.66	4.52	4.54	5.06	4.66	4.48	4.50
LJW2011	4.90	4.34	4.88	4.92	14.66	20.02	42.06	98.36	55.18	74.68	95.92	100.0
CLX2013	11.94	6.18	2.06	0.28	88.08	76.62	59.42	40.26	100.0	100.0	100.0	99.74
					$n = 40$, $p = 1600$							
SOPT-O	5.70	5.83	5.07	5.11	90.35	99.70	100.0	100.0	100.0	100.0	100.0	100.0
SOPT-B	6.11	5.74	5.68	5.41	77.18	95.28	99.80	99.97	100.0	100.0	100.0	100.0
ROPT-O	5.20	5.03	5.27	5.05	15.82	28.90	82.01	100.0	97.97	99.79	99.99	100.0
ROPT-B	5.71	5.06	5.26	5.14	16.25	25.86	64.16	99.67	86.72	92.29	99.11	100.0
SPT-DS	5.22	4.99	5.21	5.08	50.43	79.97	98.51	99.98	99.92	99.94	99.99	100.0
RPT-DS	5.01	4.71	5.06	4.94	14.62	23.71	54.68	98.14	71.49	81.98	95.74	100.0
D1958	6.82	6.16	5.64	5.38	7.92	6.62	5.96	5.60	12.30	8.30	7.02	6.48
BS1996	6.08	6.20	6.22	6.22	7.28	6.72	6.58	6.50	11.18	8.46	7.70	7.34
CQ2010	6.02	6.14	6.14	6.12	7.30	6.62	6.54	6.46	11.18	8.46	7.58	7.32
SD2008w	2.29	0.56	0.11	0.01	2.50	0.58	0.12	0.01	4.06	0.72	0.14	0.01
SD2008wo	7.72	6.92	6.45	6.28	8.94	7.47	6.82	6.53	14.30	9.30	7.95	7.32
L1996	5.16	5.18	5.14	5.16	5.04	5.06	5.08	5.06	5.12	5.08	5.02	5.02
LJW2011	5.12	5.24	5.04	4.90	6.88	8.00	11.78	51.76	14.56	20.94	42.22	98.38
CLX2013	15.64	7.50	2.56	0.22	78.46	64.74	46.54	28.60	100.0	100.0	99.96	99.08

Next we compare the power of these tests. Tables 4.1 - 4.4 show that the power of these tests strongly relies on the covariance structure as well as the values of ρ and c. We first examine the 6 tests based on the optimal projection, i.e., SOPT-O, SOPT-B, ROPT-O, ROPT-O, SPT-DS and RPT-DS. In summary, the one-by-one online projection test is slightly more powerful than the mini-batch online projection test and improves the power of data splitting projection test a lot. This is not surprising since the one-by-one online projection test keeps updating the estimated projection whenever a new data point arrives and thus in general has more accurate estimation than the mini-batch version. The mini-batch projection test sacrifices the accuracy a little bit to reduce the computation burden. The data splitting

Table 4.3: Size and power comparison for $N(c\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$ (values are in percentage).

	$c=0$				$c=0.5$				$c=1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
					$n=160$, $p=400$							
SOPT-O	5.14	5.44	5.34	5.20	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SOPT-B	5.38	5.30	5.20	4.77	100.0	100.0	99.99	100.0	100.0	100.0	100.0	100.0
ROPT-O	4.97	4.93	5.02	5.44	100.0	99.98	98.27	99.98	100.0	100.0	100.0	100.0
ROPT-B	5.05	5.20	5.20	4.98	100.0	99.97	98.09	99.94	100.0	100.0	100.0	100.0
SPT-DS	5.10	4.94	4.82	5.03	100.0	99.97	99.37	99.98	100.0	100.0	100.0	100.0
RPT-DS	5.33	4.68	5.03	5.16	99.99	99.43	89.97	96.04	100.0	100.0	100.0	100.0
D1958	5.20	5.24	4.92	5.40	100.0	100.0	99.98	85.90	100.0	100.0	100.0	100.0
BS1996	4.82	4.96	4.52	4.48	100.0	100.0	99.98	82.52	100.0	100.0	100.0	100.0
CQ2010	4.82	4.96	4.50	4.46	100.0	100.0	99.98	82.54	100.0	100.0	100.0	100.0
SD2008w	4.15	3.80	2.96	1.45	100.0	100.0	99.92	61.45	100.0	100.0	100.0	100.0
SD2008wo	5.97	5.85	5.53	5.04	100.0	100.0	99.95	84.16	100.0	100.0	100.0	100.0
L1996	5.12	4.84	4.46	5.78	57.58	36.30	19.24	7.32	93.70	73.74	39.54	11.26
LJW2011	5.00	5.02	5.84	4.96	89.76	86.02	80.22	98.60	100.0	100.0	100.0	100.0
CLX2013	5.72	5.90	5.30	2.38	100.0	100.0	100.0	99.98	100.0	100.0	100.0	100.0
					$n=160$, $p=1600$							
SOPT-O	5.51	5.35	5.41	5.21	100.0	100.0	99.94	99.99	100.0	100.0	100.0	100.0
SOPT-B	5.42	5.34	5.47	5.35	100.0	100.0	99.92	99.98	100.0	100.0	100.0	100.0
ROPT-O	5.24	5.53	4.84	5.13	99.94	98.80	84.90	70.08	100.0	100.0	100.0	100.0
ROPT-B	5.38	5.60	4.99	5.25	99.89	98.51	83.39	66.93	100.0	100.0	100.0	100.0
SPT-DS	4.85	5.04	4.92	4.79	100.0	99.94	97.61	93.27	100.0	100.0	100.0	100.0
RPT-DS	5.24	4.83	4.97	5.01	97.18	88.69	61.66	35.37	100.0	100.0	100.0	99.60
D1958	5.18	4.56	4.78	5.92	100.0	99.94	94.74	43.40	100.0	100.0	100.0	99.98
BS1996	5.52	4.72	4.90	4.82	100.0	99.80	91.82	35.86	100.0	100.0	100.0	99.98
CQ2010	5.52	4.70	4.90	4.84	100.0	99.80	91.82	35.84	100.0	100.0	100.0	99.98
SD2008w	4.10	3.73	2.99	1.74	99.99	99.72	91.28	22.26	100.0	100.0	100.0	99.87
SD2008wo	7.07	6.37	5.67	5.47	100.0	99.90	94.50	39.96	100.0	100.0	100.0	100.0
L1996	5.02	5.06	5.10	5.40	19.08	12.20	8.46	5.70	40.38	25.06	13.58	6.28
LJW2011	4.50	5.24	4.94	5.10	25.60	25.04	23.18	37.16	92.12	91.08	90.62	98.80
CLX2013	7.34	7.26	6.20	3.20	100.0	100.0	100.0	99.86	100.0	100.0	100.0	100.0

projection test is less powerful is because it discards too many data points comparing to the online projection tests. Under the settings of compound symmetry and autocorrelation, the optimal projection direction is sparse or approximately sparse, thus the tests based on regularized quadratic programming are more powerful than the tests using ridge-type estimators.

In the setting of $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_2, n = 40, c = 0.5$, SD2008wo test and CLX2013 test cannot control the type I error and their type I error rate can be as high as 20%. Thus their high power are not reliable. SOPT-O is the most powerful test among those who have control on the type I error rate. In the setting of compound symmetry, the power of tests based on the optimal projection increases as ρ increases

Table 4.4: Size and power comparison for $N(c\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$ (values are in percentage).

	$c=0$				$c=0.5$				$c=1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
					$n=40,\ p=400$							
SOPT-O	5.82	5.88	5.73	6.09	91.05	77.42	54.81	43.46	100.0	100.0	100.0	100.0
SOPT-B	5.87	6.03	5.99	6.39	83.98	69.45	47.97	36.29	100.0	100.0	99.97	99.88
ROPT-O	5.70	6.36	6.07	6.04	73.80	55.54	33.10	22.50	100.0	100.0	99.02	95.32
ROPT-B	5.77	6.07	5.98	6.37	64.99	47.91	28.92	19.00	100.0	99.93	97.89	87.22
SPT-DS	5.18	5.19	5.26	4.78	61.15	50.17	36.46	27.19	100.0	99.99	99.67	97.72
RPT-DS	5.29	4.46	5.16	4.81	46.27	35.27	21.13	13.86	99.98	99.53	91.08	68.03
D1958	4.90	4.98	4.78	5.44	88.86	77.00	50.34	17.52	100.0	100.0	99.94	85.12
BS1996	5.16	4.64	4.66	4.72	83.98	69.96	43.02	14.66	100.0	100.0	99.86	82.12
CQ2010	5.24	4.68	4.62	4.70	84.00	69.96	43.02	14.82	100.0	100.0	99.84	81.98
SD2008w	6.91	5.66	3.84	1.92	89.91	75.60	43.69	7.14	100.0	100.0	99.84	60.80
SD2008wo	14.36	12.21	9.54	7.15	95.01	86.40	59.98	21.70	100.0	100.0	99.95	85.66
L1996	4.72	4.86	4.80	4.88	18.74	12.26	7.70	5.18	40.40	24.78	12.98	5.96
LJW2011	5.20	4.78	5.20	5.14	11.82	11.30	11.78	15.34	44.40	42.08	42.08	59.82
CLX2013	14.72	13.78	12.58	6.24	93.82	89.08	77.90	52.68	100.0	100.0	100.0	99.96
					$n=40,\ p=1600$							
SOPT-O	5.87	5.99	5.76	5.88	74.33	59.22	40.02	25.12	100.0	100.0	99.98	99.84
SOPT-B	5.48	5.90	6.06	5.98	63.21	49.59	32.92	20.84	100.0	100.0	99.87	98.33
ROPT-O	6.14	5.45	5.96	5.47	32.44	24.99	15.73	8.37	99.89	98.48	84.55	40.02
ROPT-B	5.95	5.74	6.27	5.49	27.08	21.14	13.53	7.44	99.37	96.33	77.47	32.85
SPT-DS	5.25	5.19	5.09	5.12	38.03	30.96	22.88	16.49	100.0	99.94	98.81	91.04
RPT-DS	4.61	4.95	5.30	4.92	17.85	14.57	9.55	6.10	94.90	84.59	58.09	22.43
D1958	5.28	5.36	5.34	5.18	49.08	38.84	24.40	11.12	100.0	99.98	94.56	42.64
BS1996	5.18	4.96	5.24	4.46	38.02	29.42	17.74	8.02	100.0	99.84	91.44	33.82
CQ2010	5.16	5.06	5.24	4.44	38.08	29.46	17.72	8.06	100.0	99.84	91.48	33.92
SD2008w	11.73	8.22	4.08	1.65	64.17	45.15	20.19	3.45	100.0	99.82	91.48	20.90
SD2008wo	32.71	25.60	16.07	9.19	85.92	71.63	45.15	15.73	100.0	99.98	98.15	51.92
L1996	4.90	4.66	5.20	5.10	8.70	6.42	5.94	5.14	14.64	9.48	6.98	5.36
LJW2011	4.86	4.98	5.16	4.90	6.34	6.30	6.60	6.86	11.88	12.16	11.52	13.36
CLX2013	23.06	22.64	19.72	10.06	87.86	81.14	69.82	44.96	100.0	100.0	100.0	99.76

when $c = 0.5$. As the value of c increases from 0.5 to 1, the power of the these tests increases dramatically. As the dimension p increases, there is a downward trend for these tests. However, even in the most challenging case $(n, p, c) = (40, 1600, 0.5)$, the proposed SOPT-O has power between 90% and 100%. The sum-of-squares-type tests tend to become less powerful when ρ increases. This is because theses tests ignore the correlation among the variables and therefore their overall performances are not satisfactory. In the setting of autocorrelation, the power of all the tests decrease as ρ increases. We notice that some of the sum-of-squares-type tests may have more satisfactory performance than the data splitting projection tests when ρ is small. This is because $\boldsymbol{\Sigma}_2^{-1}$ is a 3-sparse matrix and it does not hurt too much if the dependences among variables are ignored. It is also observed that the power of

Table 4.5: Size and power comparison for $t_6(c\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ (values are in percentage).

	$c=0$				$c=0.5$				$c=1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
					$n=160$, $p=400$							
SOPT-O	5.22	4.78	5.94	5.35	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SOPT-B	5.08	5.20	5.62	4.99	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
ROPT-O	4.78	4.99	5.43	5.20	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
ROPT-O	4.81	4.92	5.51	5.28	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SPT-DS	4.65	4.77	5.00	5.14	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
RPT-DS	5.03	5.06	4.72	4.96	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
D1958	5.56	5.49	5.05	5.24	36.56	13.47	9.03	8.00	99.95	98.63	65.71	34.93
BS1996	4.83	5.43	5.24	5.68	31.78	13.42	9.61	8.62	99.95	98.78	69.46	38.23
CQ2010	5.35	5.60	5.31	5.74	34.68	13.78	9.76	8.72	100.0	99.10	70.35	38.80
SD2008w	2.24	0.70	0.22	0.04	12.87	1.62	0.22	0.07	99.75	22.01	1.59	0.22
SD2008wo	5.21	5.57	5.34	5.71	35.33	14.04	9.76	8.63	100.0	99.00	71.08	38.39
L1996	4.68	5.16	5.00	5.16	5.58	5.50	5.11	5.03	6.70	5.58	5.77	5.08
LJW2011	4.40	4.13	4.45	4.42	95.01	99.50	100.0	100.0	100.0	100.0	100.0	100.0
CLX2013	4.38	2.66	1.07	0.17	100.0	99.91	99.01	94.66	100.0	100.0	100.0	100.0
					$n=160$, $p=1600$							
SOPT-O	5.11	4.92	4.83	4.96	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SOPT-B	5.14	5.19	5.06	4.94	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
ROPT-O	4.92	4.77	4.61	4.76	99.99	100.0	100.0	100.0	100.0	100.0	100.0	100.0
ROPT-B	4.74	4.75	4.61	4.80	99.92	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SPT-DS	5.19	4.82	4.98	4.75	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
RPT-DS	5.05	4.59	4.80	4.41	95.94	99.61	100.0	100.0	100.0	100.0	100.0	100.0
D1958	5.64	5.25	5.28	5.14	8.93	6.67	5.62	5.87	37.23	13.27	8.74	7.52
BS1996	5.10	5.17	5.54	5.43	7.92	6.61	5.90	6.22	31.38	13.15	9.22	8.12
CQ2010	5.42	5.34	5.62	5.45	8.59	6.77	5.98	6.29	35.14	13.48	9.40	8.17
SD2008w	1.37	0.15	0.03	0.00	2.02	0.24	0.04	0.00	6.59	0.38	0.02	0.00
SD2008wo	5.46	5.42	5.63	5.42	8.59	6.85	5.98	6.23	35.82	13.75	9.43	8.16
L1996	5.11	4.84	5.18	5.11	5.20	5.11	4.73	5.20	5.07	4.93	5.00	4.45
LJW2011	4.44	4.35	4.27	4.14	27.76	45.62	85.50	100.0	94.92	99.58	99.99	100.0
CLX2013	4.41	2.08	0.71	0.09	99.99	99.60	97.52	90.55	100.0	100.0	100.0	100.0

these sum-of-squares-type tests decreases significantly as the correlation increases and become less powerful than the data splitting projection tests when $\rho = 0.95$.

4.3.3 Size and power comparison for multivariate t-distribution

We also investigate the numerical performance of type I error and power of the proposed online projection tests without the normality assumption. To this end, we generate random samples from the multivariate t-distribution with degrees of freedom 6. Again, we consider bothn compound symmetry and autocorrelation covariance structures. We use the same critical values as those used with normality assumption to examine the robustness of the online projection tests. Simulation

Table 4.6: Size and power comparison for $t_6(c\boldsymbol{\mu}, \boldsymbol{\Sigma}_1)$ (values are in percentage).

	$c = 0$				$c = 0.5$				$c = 1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
					$n = 40$, $p = 400$							
SOPT-O	5.59	5.37	5.54	5.16	88.41	98.30	100.0	100.0	100.0	99.99	100.0	100.0
SOPT-B	5.89	5.62	5.52	5.30	78.57	94.22	99.50	99.96	100.0	99.97	100.0	100.0
ROPT-O	5.70	5.58	5.49	5.26	65.20	87.21	99.54	100.0	99.96	99.97	99.99	100.0
ROPT-B	6.03	5.63	5.42	5.40	52.33	73.24	95.99	99.97	99.59	99.75	99.98	100.0
SPT-DS	4.79	5.14	4.53	4.65	59.04	82.06	97.16	99.85	99.89	99.87	99.98	99.97
RPT-DS	5.02	5.10	4.59	4.78	38.22	58.97	88.91	99.96	96.86	98.47	99.93	100.0
D1958	5.25	5.57	5.34	5.12	8.42	6.86	6.22	5.88	38.68	14.42	9.83	8.23
BS1996	4.72	5.65	5.90	5.96	7.61	7.03	6.77	6.65	34.53	14.66	10.84	9.48
CQ2010	5.98	6.03	6.09	6.09	9.53	7.47	6.99	6.78	44.09	15.69	11.25	9.72
SD2008w	2.27	1.13	0.31	0.08	3.87	1.40	0.34	0.10	17.70	2.93	0.60	0.14
SD2008wo	6.18	6.57	6.27	6.05	9.97	8.06	7.21	6.72	48.77	17.70	11.66	9.67
L1996	4.90	4.90	4.87	4.87	5.16	5.01	4.91	4.86	5.29	5.16	5.07	4.98
LJW2011	4.24	4.42	4.34	4.20	12.17	17.69	35.17	96.12	47.07	65.95	92.81	100.0
CLX2013	9.61	5.44	2.04	0.34	68.31	54.94	39.89	24.08	99.98	99.81	98.70	94.09
					$n = 40$, $p = 1600$							
SOPT-O	5.81	5.39	5.18	4.89	75.58	95.94	99.90	100.0	100.0	100.0	100.0	100.0
SOPT-B	6.21	5.32	5.11	4.79	61.58	87.35	98.92	100.0	100.0	99.98	99.99	100.0
ROPT-O	4.93	4.79	4.75	5.74	12.71	22.99	67.67	99.95	89.86	96.65	99.74	100.0
ROPT-B	5.03	4.95	4.78	4.99	12.45	20.61	53.67	99.00	75.52	85.93	97.51	100.0
SPT-DS	5.28	5.09	4.83	4.62	40.44	69.29	95.49	99.97	99.82	99.83	99.99	99.99
RPT-DS	4.88	5.06	5.02	4.74	11.40	18.51	46.12	96.50	61.99	74.62	92.62	99.98
D1958	5.44	5.59	5.43	5.23	5.91	5.92	5.64	5.35	8.29	6.96	6.18	5.85
BS1996	4.92	5.75	5.93	5.98	5.45	6.03	6.09	6.12	7.48	7.13	6.82	6.65
CQ2010	6.16	6.19	6.18	6.13	6.75	6.44	6.31	6.28	9.42	7.57	7.12	6.89
SD2008w	1.19	0.43	0.05	0.02	1.34	0.43	0.05	0.02	1.89	0.53	0.05	0.02
SD2008wo	6.24	6.51	6.34	6.05	7.07	6.89	6.56	6.20	9.92	8.22	7.32	6.80
L1996	4.95	4.97	4.93	4.96	4.98	4.95	4.97	4.96	4.97	4.93	4.92	4.91
LJW2011	4.43	4.61	4.26	4.15	5.71	7.08	9.87	44.45	12.53	18.11	35.10	96.15
CLX2013	11.57	5.67	1.71	0.14	55.24	42.53	29.76	16.47	99.98	99.55	96.81	89.51

results are summarized in Tables 4.5 - 4.8, from which we observe that all the online projection tests and the data splitting projection tests can retain the type I error rate very well. This implies that these projection tests are not very sensitive to the normality assumption. All other alternative tests except for the CQ2010 test fail to retain the type I error. The pattern of power is very similar to that of multivariate normal setting. When the covariance matrix is compound symmetry, the proposed SOPT-O is the most powerful test including those that cannot retain the type I error. When the covariance matrix is autocorrelation structure, the proposed SOPT-O is the most powerful one among those who can retain the type I error rate.

Table 4.7: Size and power comparison for $t_6(c\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$ (values are in percentage).

	$c = 0$				$c = 0.5$				$c = 1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
					$n = 160$, $p = 400$							
SOPT-O	4.89	5.37	5.26	5.13	99.99	99.99	99.59	100.0	100.0	100.0	100.0	100.0
SOPT-B	4.97	5.33	5.14	5.19	99.99	99.98	99.56	100.0	100.0	100.0	100.0	100.0
ROPT-O	4.95	5.27	5.54	5.40	100.0	99.51	92.75	99.31	100.0	100.0	100.0	100.0
ROPT-B	4.93	5.12	5.39	5.32	99.99	99.47	92.06	99.06	100.0	100.0	100.0	100.0
SPT-DS	4.68	4.94	4.29	4.69	100.0	99.58	94.11	99.66	100.0	100.0	99.98	100.0
RPT-DS	4.77	4.98	4.82	4.95	99.58	96.61	78.27	88.18	100.0	100.0	99.97	100.0
D1958	0.54	1.09	2.45	4.73	99.56	98.96	95.90	53.54	99.98	100.0	99.98	99.96
BS1996	0.28	0.56	1.75	3.87	99.36	98.61	94.01	48.96	99.98	100.0	99.98	99.95
CQ2010	4.77	4.80	4.91	4.53	100.0	99.96	97.68	52.31	100.0	100.0	100.0	99.99
SD2008w	0.21	0.35	1.04	1.27	99.31	98.38	91.41	26.90	99.99	100.0	99.98	99.78
DS2008wo	0.49	0.88	2.17	4.34	99.74	99.21	95.30	52.32	100.0	100.0	99.98	99.97
L1996	4.69	5.27	4.75	5.25	45.62	28.58	14.71	6.44	87.08	63.57	33.29	9.76
LJW2011	4.11	4.61	4.68	4.23	82.77	76.28	70.63	95.51	100.0	100.0	100.0	100.0
CLX2013	5.30	5.14	4.74	1.99	100.0	100.0	99.85	97.34	100.0	100.0	100.0	100.0
					$n = 160$, $p = 1600$							
SOPT-O	5.14	5.23	4.95	4.72	100.0	99.96	98.67	99.53	100.0	100.0	100.0	100.0
SOPT-B	5.48	5.17	5.13	4.73	100.0	99.94	98.37	99.17	100.0	100.0	100.0	100.0
ROPT-O	5.45	4.92	4.79	4.70	98.87	92.61	70.22	52.48	100.0	100.0	100.0	99.98
ROPT-B	5.28	4.93	5.21	5.03	98.72	91.84	68.45	49.92	100.0	100.0	100.0	99.99
SPT-DS	5.19	5.08	5.35	4.85	99.95	98.95	89.73	79.71	100.0	100.0	100.0	100.0
RPT-DS	5.19	4.68	4.39	4.81	88.97	75.32	45.68	27.00	100.0	100.0	99.90	97.5
D1958	0.00	0.04	0.49	3.10	43.79	40.97	33.62	17.27	99.68	99.66	99.56	96.05
BS1996	0.00	0.02	0.22	2.26	27.97	27.81	22.73	12.69	99.50	99.60	99.43	93.99
CQ2010	4.92	5.17	5.35	4.97	97.87	91.43	66.06	19.72	100.0	100.0	100.0	97.68
SD2008w	0.00	0.02	0.10	0.59	21.42	21.18	16.14	4.90	99.20	99.38	98.86	81.67
SD2008wo	0.00	0.05	0.42	3.00	41.38	39.18	31.28	15.92	99.85	99.84	99.73	94.36
L1996	4.97	5.09	5.55	4.95	15.09	10.91	7.08	5.33	34.25	21.43	11.64	5.99
LJW2011	4.46	4.24	4.48	4.45	20.00	19.15	19.03	30.21	84.88	83.68	83.38	96.87
CLX2013	5.87	5.93	5.38	2.53	99.99	99.97	99.73	95.36	100.0	100.0	100.0	100.0

4.3.4 Real data example

In this section, we apply the proposed sparse projection tests to a real dataset of high resolution micro-computed tomography. This dataset contains the bone density of 58 mice's skull of three different genotypes ("T0A0", "T0A1", "T1A1") measured at different bone density levels in a genetic mutation study. For each mouse, bone density is measured for 16 different areas of its skull. For each area, bone volume is measured at density levels from 130 - 249. This dataset was collected at Center for Quantitative X-Ray Imaging at the Pennsylvania State University. See Percival et al. (2014) for a detailed description of protocols. In this empirical

Table 4.8: Size and power comparison for $t_6(c\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)$ (values are in percentage).

	$c = 0$				$c = 0.5$				$c = 1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
					$n = 40$, $p = 400$							
SOPT-O	6.19	5.82	6.00	5.74	76.39	62.21	42.33	31.76	100.0	99.97	99.66	99.52
SOPT-B	6.04	5.89	5.68	6.23	68.29	55.25	36.58	27.53	100.0	99.95	99.37	98.20
ROPT-O	6.00	5.90	6.02	5.87	57.47	43.44	25.83	18.85	99.98	99.61	94.46	84.48
ROPT-B	6.05	5.67	6.03	5.94	49.53	38.07	22.86	15.87	99.87	99.19	91.50	74.71
SPT-DS	4.74	4.44	4.81	5.19	46.94	38.23	27.23	19.02	99.95	99.52	96.50	91.21
RPT-DS	4.77	5.00	4.53	5.52	36.00	26.34	16.44	11.40	99.27	95.97	79.71	55.13
D1958	0.03	0.16	0.85	3.63	9.72	10.63	10.93	8.87	92.90	91.23	84.55	47.77
BS1996	0.01	0.06	0.51	3.20	5.43	6.92	7.71	7.76	90.47	88.62	80.54	43.69
CQ2010	5.04	4.89	4.73	4.73	57.67	44.93	27.18	10.75	99.99	99.82	96.78	53.52
SD2008w	0.02	0.04	0.30	0.99	6.14	6.69	6.11	2.94	89.35	86.42	75.20	24.96
SD2008wo	0.24	0.52	1.66	5.09	20.25	20.14	18.04	11.63	97.27	95.81	90.66	56.35
L1996	5.04	4.83	4.88	4.84	15.39	10.85	7.66	5.36	34.28	21.55	11.88	6.45
LJW2011	3.99	4.58	4.53	4.20	9.89	9.84	10.32	12.91	37.22	35.84	35.41	52.02
CLX2013	11.38	10.96	9.60	4.58	75.57	69.57	57.40	34.03	100.0	100.0	99.86	97.08
					$n = 40$, $p = 1600$							
SOPT-O	5.93	5.86	5.91	6.03	58.34	45.07	30.92	17.82	99.99	99.96	99.36	96.20
SOPT-B	5.46	5.82	6.02	6.24	48.12	36.98	26.30	16.15	99.99	99.90	98.59	91.16
ROPT-O	5.48	5.96	6.20	6.19	23.81	18.51	12.64	7.43	97.26	91.16	70.57	30.22
ROPT-B	5.54	5.52	5.81	5.82	20.15	16.04	11.49	7.24	94.57	85.87	63.29	25.07
SPT-DS	4.89	4.61	4.85	4.77	28.90	24.90	17.90	12.35	99.89	99.19	94.29	78.77
RPT-DS	5.24	4.58	5.08	5.37	13.82	11.03	8.69	5.40	83.00	69.65	44.60	17.05
D1958	0.00	0.00	0.02	1.28	0.01	0.01	0.15	2.25	6.76	7.62	8.96	9.62
BS1996	0.00	0.00	0.01	0.78	0.00	0.01	0.06	1.36	3.05	3.60	5.24	6.85
CQ2010	5.14	5.17	5.09	5.25	21.73	17.47	11.39	7.08	97.03	90.64	66.82	20.99
SD2008w	0.00	0.00	0.00	0.06	0.00	0.00	0.01	0.18	1.31	1.82	2.21	1.66
SD2008wo	0.00	0.00	0.15	2.31	0.08	0.22	0.85	4.14	21.14	21.05	21.04	16.37
L1996	5.07	5.12	4.76	5.09	7.72	6.65	5.08	5.27	12.20	9.06	6.38	5.44
LJW2011	3.89	4.63	4.00	4.34	5.47	5.81	5.14	5.78	10.12	10.39	10.17	11.02
CLX2013	14.61	15.40	14.07	6.98	62.50	57.26	47.60	26.83	100.0	99.95	99.44	94.29

analysis, we are interested in comparing the bone density patterns of two different areas in mice's skull. We compare the performance of the proposed SOPT-O and SPT-DS with several existing methods. To emphasize the high-dimensionality nature of this dataset, we only use a subset of the dataset. We select the mice of the genotype "T0A1" and there are 29 observations available in the dataset, i.e., sample size $n = 29$. The two areas of the skull "Mandible" and "Nasal" are selected. We use all density levels from 130 - 249 for our analysis, hence dimension $p = 120$. We first take the difference of the bone density of the selected two areas at the corresponding density level for each subject since the two bones come from the same mouse. Then we normalize the bone density in the sense that $\frac{1}{29}\sum_{i=1}^{29} X_{ij}^2 = 1$ for all $1 \leq j \leq 120$.

Table 4.9: P-values of different tests for bone density dataset.

δ	1.0	0.8	0.6	0.4	0.3	0.2
SOPT-O	0	0	4.6×10^{-13}	4.0×10^{-10}	1.6×10^{-4}	0.0325
SPT-DS	7.9×10^{-10}	6.7×10^{-9}	3.3×10^{-7}	9.9×10^{-6}	3.4×10^{-4}	0.0418
RPT-DS	2.4×10^{-8}	5.1×10^{-7}	1.9×10^{-5}	0.0014	0.0140	0.1462
D1958	9.0×10^{-9}	1.4×10^{-6}	1.7×10^{-4}	1.2×10^{-2}	0.0697	0.2705
BS1996	0	0	0	1.2×10^{-4}	0.0763	0.7684
CQ2010	0	0	0	1.6×10^{-4}	0.0810	0.7717
SD2008w	0	0	2.0×10^{-9}	1.6×10^{-2}	0.2494	0.7995
SD2008wo	0	0	0	6.9×10^{-9}	0.0056	0.5414
L1996	1.1×10^{-10}	3.1×10^{-8}	1.6×10^{-3}	0.1265	0.2625	0.4574
LJW2011	3.8×10^{-9}	4.2×10^{-8}	4.5×10^{-6}	8.5×10^{-4}	9.6×10^{-3}	0.2031
LCX2013	0	1.1×10^{-14}	6.4×10^{-8}	4.3×10^{-3}	0.1885	0.9651

We apply SOPT-O, SPT-DS and several other existing tests to perform the one-sample test and compute the p-values. The p-values are reported in the firs row in Table 4.9. The p-values of all methods are very small ($\ll 0.05$), implying that the bone volume is significantly different. To compare the power of different tests, we also compute the p-values of different tests as we decrease the signals. Let $\bar{\mathbf{x}}$ be the sample mean and $\mathbf{r}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ is the residual for the ith subject. Then a new observation $\mathbf{z}_i = \delta\bar{\mathbf{x}} + \mathbf{r}_i$ is constructed for the ith subject. By the construction, a smaller δ leads to a weaker signal and would make the test more challenging. Table 4.9 reports the p-values of all tests with $\delta = 1, 0.8, 0.6, 0.4, 0.3, 0.2$. As expected, the p-values all tests increase as δ decreases. When $\delta = 0.8$ and 0.6, all the tests perform well and reject the null hypothesis at level 0.05. When $\delta = 0.4$, the Lauter's test starts to fail to reject the null hypothesis. When $\delta = 0.3$, three projections SOPT-O, SPT-DS and RPT-DS are able to reject the null hypothesis while all other tests except the LJW2011 test fail to reject null hypothesis. When $\delta = 0.2$, only SOPT-O and SPT-DS reject the null hypothesis, which suggests that our proposed projection tests can still perform well even though the signal is weak. Among those tests that fail to reject the null hypothesis at $\delta = 0.2$, RPT-DS has the smallest p-value.

We plot the histogram of absolute values of paired sample correlations among all bone density levels in Figure 4.4. The histogram indicates that some bone density levels are highly correlated. This may explain why the proposed projection tests are more powerful than the D1958 test, the BS1996 test test and SD2008w test since these tests do not take the dependence among variables into account.

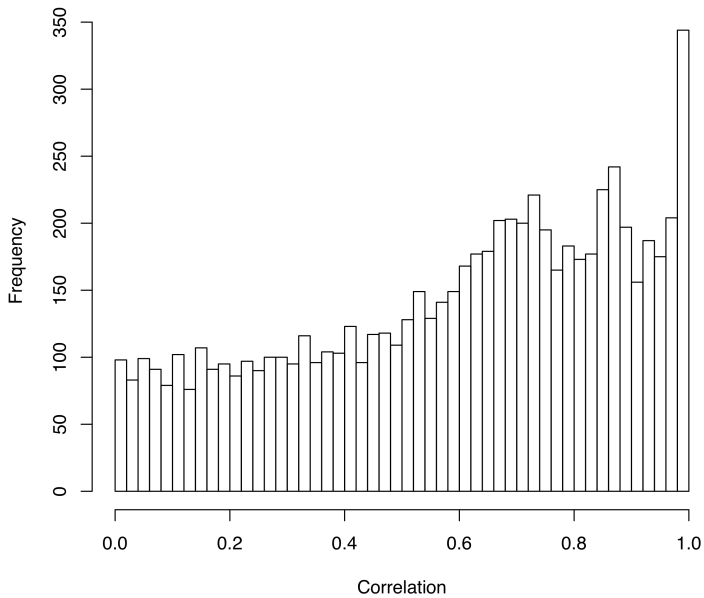


Figure 4.4: Histogram of absolute values of paired sample correlations among bone densities at all different bone density levels.

4.4 Proofs

Proof of Theorem 4.1

(i) Define $\widetilde{\mathbf{x}}_t = \mathbf{x}_t - \boldsymbol{\mu}$, thus $z_t = \widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\beta}}_{t-1}$ and let $m_n = n - k_n$. Since $k_n = n^\tau = o(n)$, we have $m_n/n \to 1$. By the construction, we know that $\{z_{k_n+1}, \ldots, z_n\}$ is a martingale difference. According to the central limit theorem for martingale difference in Lemma D.1, we need to show (1) $\frac{1}{m_n}\sum_{t=k_n+1}^{n} z_t^2 \to \sigma^2$ for some $\sigma > 0$ and (2) $\frac{1}{\sqrt{m_n}}\mathrm{E}(\max_{k_n+1\le t\le n}|z_t|) \to 0$. Note that

$$
\begin{aligned}
\frac{1}{m_n}\sum_{t=k_n+1}^{n} z_t^2 =& \frac{1}{m_n}\sum_{t=k_n+1}^{n}(\widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\beta}}_{t-1})^2 \\
=& \frac{1}{m_n}\sum_{t=k_n+1}^{n}\left(\widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\beta}}_{t-1} - \widetilde{\mathbf{x}}_t^\top \boldsymbol{\beta}^\star + \widetilde{\mathbf{x}}_t^\top \boldsymbol{\beta}^\star\right)^2 \\
=& \underbrace{\frac{1}{m_n}\sum_{t=k_n+1}^{n}\left[\widetilde{\mathbf{x}}_t^\top(\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^\star)\right]^2}_{①} + \underbrace{\frac{2}{m_n}\sum_{t=k_n+1}^{n}\left(\widetilde{\mathbf{x}}_t^\top \boldsymbol{\beta}^\star\right)\left[\widetilde{\mathbf{x}}_t^\top(\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^\star)\right]}_{②} + \\
& \underbrace{\boldsymbol{\beta}^{\star\top}\left[\frac{1}{m_n}\sum_{t=k_n+1}^{n}\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top\right]\boldsymbol{\beta}^\star}_{③}
\end{aligned}
$$

We first deal with term ③. Let

$$
\widetilde{\boldsymbol{\Sigma}} = \frac{1}{m_n}\sum_{t=k_n+1}^{n}\widetilde{\mathbf{x}}_t\widetilde{\mathbf{x}}_t^\top = \frac{1}{m_n}\sum_{t=k_n+1}^{n}(\mathbf{x}_t - \boldsymbol{\mu})(\mathbf{x}_t - \boldsymbol{\mu})^\top.
$$

Since $\mathbf{x}_t$ is sub-Gaussian, by Lemma 4.3 we have $\|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_\infty \le M_2\sqrt{\log p/m_n}$ for some $M_2 > 0$ with high probability. Therefore

$$
\begin{aligned}
③ = \boldsymbol{\beta}^{\star\top}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\boldsymbol{\beta}^\star + \boldsymbol{\beta}^{\star\top}\boldsymbol{\Sigma}\boldsymbol{\beta}^\star &\le M_2\sqrt{\frac{\log p}{m_n}}\|\boldsymbol{\beta}^\star\|_2^2 + \frac{1}{\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} \\
&\le C_2^2 M_2 s\sqrt{\frac{\log p}{m_n}} + \frac{1}{\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}.
\end{aligned}
$$

The last inequality is because of the assumptions $\|\boldsymbol{\beta}^\star\|_0 = s$ and $\|\boldsymbol{\beta}^\star\|_\infty \le C_2$. Since $s\sqrt{\log p/m_n} \simeq s\sqrt{\log p/n} = o(1)$, we have ③ $\to (\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^{-1} > 0$. Now we deal with term ①. Since $\widetilde{\mathbf{x}}_t$ is sub-Gaussian, we know $\|\widetilde{\mathbf{x}}_t\|_\infty \le M_1\sqrt{\log p}$ for some $M_1 > 0$ with high probability. From the result of Theorem 3.3, we have $\|\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^\star\|_1 = O\left(\frac{s^2}{\kappa}\sqrt{\frac{\log p}{t-1}}\right)$ for $t = k_n + 1, \ldots, n$. With the choice $k_n = n^\tau$, we

further have $\|\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^\star\|_1 = O\left(\frac{s^2}{\kappa}\sqrt{\frac{\log p}{n^\tau}}\right)$. As a result,

$$\widetilde{\mathbf{x}}_t^\top(\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^\star) \leq \|\widetilde{\mathbf{x}}_t\|_\infty \|\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^\star\|_1 = O\left(\frac{s^2 \log p}{\kappa n^{\tau/2}}\right) = o(1),$$

implying that ① $\to 0$. As for term ②, note that

$$\widetilde{\mathbf{x}}_t^\top \boldsymbol{\beta}^\star \leq \|\widetilde{\mathbf{x}}_t\|_\infty \|\boldsymbol{\beta}^\star\|_1 \leq M_1 C_2 s\sqrt{\log p}.$$

Hence we have

$$② = \frac{1}{m_n}\sum_{t=k_n+1}^{n} \widetilde{\mathbf{x}}_t^\top \boldsymbol{\beta}^\star \cdot O\left(\frac{s^2 \log p}{\kappa n^{\tau/2}}\right) = O\left(\frac{s^3(\log p)^{3/2}}{\kappa n^{\tau/2}}\right) = o(1).$$

Combine ①, ② and ③, we have

$$\frac{1}{m_n}\sum_{t=k_n+1}^{n} z_t^2 = \frac{1}{m_n}\sum_{t=k_n+1}^{n} (\widetilde{\mathbf{x}}_t^\top \widehat{\boldsymbol{\beta}}_{t-1})^2 \to \frac{1}{\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}.$$

We next show $\frac{1}{\sqrt{m_n}}\mathrm{E}(\max_{k_n+1\leq t\leq n}|z_t|) \to 0$. Note that

$$\begin{aligned}
&\frac{1}{\sqrt{m_n}}\mathrm{E}\left(\max_{k_n+1\leq t\leq n}|z_t|\right) \\
=&\frac{1}{\sqrt{m_n}}\mathrm{E}\left(\max_{k_n+1\leq t\leq n}|\widehat{\boldsymbol{\beta}}_{t-1}^\top \widetilde{\mathbf{x}}_t|\right) \\
\leq&\frac{1}{\sqrt{m_n}}\mathrm{E}\left(\max_{k_n+1\leq t\leq n}|\boldsymbol{\beta}^{\star\top} \widetilde{\mathbf{x}}_t|\right) + \frac{1}{\sqrt{m_n}}\mathrm{E}\left(\max_{k_n+1\leq t\leq n}|(\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^\star)^\top \widetilde{\mathbf{x}}_t|\right)
\end{aligned}$$

Since $\widetilde{\mathbf{x}}_t$ is a sub-Gaussian random vector with variance proxy σ^2, then $\boldsymbol{\beta}^{\star\top}\widetilde{\mathbf{x}}_t$ is also a sub-Gaussian random variable with variance proxy $\|\boldsymbol{\beta}^\star\|_2^2\sigma^2$, which is bounded by $sC_2^2\sigma^2$. From Lemma B.1, we know that

$$\frac{1}{\sqrt{m_n}}\mathrm{E}\left(\max_{k_n+1\leq t\leq n}|\boldsymbol{\beta}^{\star\top}\widetilde{\mathbf{x}}_t|\right) \leq \sigma\|\boldsymbol{\beta}^\star\|_2\sqrt{\frac{2\log(2m_n)}{m_n}} = O\left(\sqrt{\frac{s\log n}{n}}\right) \to 0. \quad (4.7)$$

From the result of Theorem 3.3, we have $\|\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^\star\|_2^2 = O\left(\frac{s^3 \log p}{\kappa^2 n^\tau}\right)$. Again by

Lemma B.1, we have

$$\frac{1}{\sqrt{m}}\mathrm{E}\left(\max_{1\le i\le n}|(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star)^\top\mathbf{x}_t|\right)=O\left(\frac{s^3\log p}{\kappa^2 n^\tau}\sqrt{\frac{\log n}{n}}\right)\to 0. \tag{4.8}$$

Combine (4.7) and (4.8), we know $\frac{1}{\sqrt{m_n}}\mathrm{E}(\max_{k_n+1\le t\le n}|z_t|)\to 0$. According to Lemma D.1, we conclude

$$\frac{\sqrt{n-n^\tau}\bar{z}}{s_z}\to N(0,1).$$

(ii) We derive the power function for the proposed test statistic T_y. We first show s_y^2 and s_z^2 are asymptotically equivalent. By definition, we have

$$\begin{aligned}
s_y^2-s_z^2&=\frac{1}{m_n}\sum_{t=k_n+1}^{n}y_t^2-\bar{y}^2-s_z^2\\
&=\frac{1}{m_n}\sum_{t=k_n+1}^{n}(\mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}_{t-1})^2-\bar{y}^2-s_z^2\\
&=\frac{1}{m_n}\sum_{t=k_n+1}^{n}((\mathbf{x}_t-\boldsymbol{\mu})^\top\widehat{\boldsymbol{\beta}}_{t-1}+\boldsymbol{\mu}^\top\widehat{\boldsymbol{\beta}}_{t-1})^2-\bar{y}^2-s_z^2\\
&=\underbrace{\frac{1}{m_n}\sum_{t=k_n+1}^{n}(\boldsymbol{\mu}^\top\widehat{\boldsymbol{\beta}}_{t-1})^2-\bar{y}^2}_{④}+\underbrace{\frac{2}{m_n}\sum_{t=k_n+1}^{n}\widehat{\boldsymbol{\beta}}_{t-1}^\top\boldsymbol{\mu}(\mathbf{x}_t-\boldsymbol{\mu})^\top\widehat{\boldsymbol{\beta}}_{t-1}}_{⑤}.
\end{aligned}$$

We will show both of the terms ④ and ⑤ converge to 0 in probability. Before we show ④ $\to 0$, wee first show $\bar{y}\to 1$. Note that $\boldsymbol{\mu}^\top\boldsymbol{\beta}^\star=1$, we have

$$\begin{aligned}
\bar{y}-1&=\frac{1}{m_n}\sum_{t=k_n+1}^{n}y_t-\boldsymbol{\mu}^\top\boldsymbol{\beta}^\star\\
&=\frac{1}{m_n}\sum_{t=k_n+1}^{n}\mathbf{x}_t^\top\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\mu}^\top\boldsymbol{\beta}^\star\\
&=\frac{1}{m_n}\sum_{t=k_n+1}^{n}(\mathbf{x}_t-\boldsymbol{\mu})^\top\widehat{\boldsymbol{\beta}}_{t-1}+\frac{1}{m_n}\sum_{t=k_n+1}^{n}\boldsymbol{\mu}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star).
\end{aligned}\tag{4.9}$$

For the second term of the right hand side of (4.9), we have

$$\begin{aligned}\frac{1}{m_n}\sum_{t=k_n+1}^{n}\boldsymbol{\mu}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star) &\le \frac{1}{m_n}\sum_{t=k_n+1}^{n}\|\boldsymbol{\mu}\|_\infty\|\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star\|_1 \\ &= O\left(\|\boldsymbol{\mu}\|_\infty s^2\sqrt{\frac{\log p}{n^\tau}}\right)\to 0.\end{aligned} \tag{4.10}$$

Let $\bar{\mathbf{x}}_{m_n}=\frac{1}{m_n}\sum_{t=k_n+1}^{n}\mathbf{x}_t$, then the first term of the right hand side of (4.9) can be written as

$$\begin{aligned}&\frac{1}{m_n}\sum_{t=k_n+1}^{n}(\mathbf{x}_t-\boldsymbol{\mu})^\top\widehat{\boldsymbol{\beta}}_{t-1}\\ =&\frac{1}{m_n}\sum_{t=k_n+1}^{n}(\mathbf{x}_t-\boldsymbol{\mu})^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star)+\frac{1}{m_n}\sum_{t=k_n+1}^{n}(\mathbf{x}_t-\boldsymbol{\mu})^\top\boldsymbol{\beta}^\star\\ =&\frac{1}{m_n}\sum_{t=k_n+1}^{n}(\mathbf{x}_t-\boldsymbol{\mu})^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star)+(\bar{\mathbf{x}}_{m_n}-\boldsymbol{\mu})^\top\boldsymbol{\beta}^\star\\ \le&\frac{1}{m_n}\sum_{t=k_n+1}^{n}\|\mathbf{x}_t-\boldsymbol{\mu}\|_\infty\|\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star\|_1+\|\bar{\mathbf{x}}_{m_n}-\boldsymbol{\mu}\|_\infty\|\boldsymbol{\beta}^\star\|_1\\ =&O\left(\frac{s^2\log p}{n^{\tau/2}}\right)+O\left(s\sqrt{\frac{\log p}{m_n}}\right)\to 0.\end{aligned} \tag{4.11}$$

The last equality is due to the fact the $\mathbf{x}_t$ and $\bar{\mathbf{x}}_{m_n}$ are sub-Gaussian random vectors and $\|\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star\|_1=O\left(s^2\sqrt{\frac{\log p}{n^\tau}}\right)$. Combine (4.9), (4.10) and (4.11), we have $\bar{y}-1\to 0$ and hence $\bar{y}^2\to 1$. In order to show ④ $\to 0$, it suffices to show that

$\frac{1}{m_n}\sum_{t=k_n+1}^{n}\left(\boldsymbol{\mu}^\top\widehat{\boldsymbol{\beta}}_{t-1}\right)^2 \to 1$. To see this,

$$
\begin{aligned}
&\frac{1}{m_n}\sum_{t=k_n+1}^{n}\left(\boldsymbol{\mu}^\top\widehat{\boldsymbol{\beta}}_{t-1}\right)^2\\
=&\frac{1}{m_n}\sum_{t=k_n+1}^{n}\left(\boldsymbol{\mu}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star)-1\right)^2\\
=&\frac{1}{m_n}\sum_{t=k_n+1}^{n}\left(\boldsymbol{\mu}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star)\right)^2-\frac{2}{m_n}\sum_{t=k_n+1}^{n}\boldsymbol{\mu}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star)+1\\
\le&\frac{1}{m_n}\sum_{t=k_n+1}^{n}\left(\|\boldsymbol{\mu}\|_\infty\|\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star\|_1\right)^2-\frac{2}{m_n}\sum_{t=k_n+1}^{n}\|\boldsymbol{\mu}\|_\infty\|\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star\|_1+1\\
=&1+o(1).
\end{aligned}
$$

The last equality is due to

$$
\boldsymbol{\mu}^\top(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star)\le\|\boldsymbol{\mu}\|_\infty\|\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star\|_1=O\left(\|\boldsymbol{\mu}\|_\infty s^2\sqrt{\frac{\log p}{n^\tau}}\right)\to 0. \tag{4.12}
$$

Thus we have ④ $\to 0$. Now we deal with term ⑤, which can be rewritten as

$$
\begin{aligned}
⑤ &= \frac{2}{m_n}\sum_{t=k_n+1}^{n}\left\{(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star)^\top\boldsymbol{\mu}(\mathbf{x}_t-\boldsymbol{\mu})^\top\widehat{\boldsymbol{\beta}}_{t-1}+(\mathbf{x}_t-\boldsymbol{\mu})^\top\widehat{\boldsymbol{\beta}}_{t-1}\right\}\\
&= \frac{2}{m_n}\sum_{t=k_n+1}^{n}\left\{(\widehat{\boldsymbol{\beta}}_{t-1}-\boldsymbol{\beta}^\star)^\top\boldsymbol{\mu}+1)(\mathbf{x}_t-\boldsymbol{\mu})^\top\widehat{\boldsymbol{\beta}}_{t-1}\right\}
\end{aligned}
$$

According to (4.12) and (4.11), we know that

$$
⑤ = \frac{2}{m_n}\sum_{t=k_n+1}^{n}\left\{(o(1)+1)(\mathbf{x}_t-\boldsymbol{\mu})^\top\widehat{\boldsymbol{\beta}}_{t-1}\right\}\to 0.
$$

As a result, we have $s_y^2/s_z^2\to 1$ and hence

$$
\frac{\sqrt{m_n}\bar{z}}{s_y}\to N(0,1) \text{ and } s_y\to\frac{1}{\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}.
$$

Note that $\bar{y}-\bar{z}=\frac{1}{m_n}\sum_{t=k_n+1}^{n}\boldsymbol{\mu}^\top\widehat{\boldsymbol{\beta}}_{t-1}$, (4.12) also implies that $\bar{y}-\bar{z}=1+o(1)$.

Therefore, we have

$$
\begin{aligned}
\mathrm{P}(|T_y| \geq z_{\alpha/2}) &= \mathrm{P}\left(\left|\frac{\sqrt{m_n}\bar{y}}{s_y}\right| \geq z_{\alpha/2}\right) \\
&= \mathrm{P}\left(\frac{\sqrt{m_n}\bar{y}}{s_y} \geq z_{\alpha/2}\right) + \mathrm{P}\left(\frac{\sqrt{m_n}\bar{y}}{s_y} \leq -z_{\alpha/2}\right) \\
&\simeq \mathrm{P}\left(\frac{\sqrt{m_n}\bar{z}}{s_z} \geq z_{\alpha/2} - \frac{\sqrt{m_n}}{s_z}\right) \\
&\simeq \mathrm{P}\left(\frac{\sqrt{m_n}\bar{z}}{s_z} \geq z_{\alpha/2} - \sqrt{n}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\
&= \Phi(\sqrt{n}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - z_{\alpha/2}),
\end{aligned}
$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal.

Chapter 5 Model Free Feature Screening via Projection Correlation

5.1 Introduction

The technological development has made the data collection and storage easy and cheap in diverse fields. Datasets with ultra-high dimensional features characterize many contemporary research problems in computational neuroscience, machine learning, statistics, engineering, social science, finance and so on. When the features contain redundant or noisy information, estimating their functional relationship with the response can become quite challenging in terms of computational expediency, statistical accuracy and algorithmic stability (Fan et al. 2009). To overcome such challenges caused by ultra-high dimensionality, Fan and Lv (2008) proposed a sure independence screening (SIS) method which aims to screen out the redundant features by ranking their marginal Pearson correlations. The SIS method is named after the sure independence screening property which states the selected subset of features contains all the active ones with probability approaching one. The promising numerical performance soon made SIS popular among ultra-high dimensional studies. The sure screening idea has be applied to many important regression problems including generalized linear model (Fan and Song 2010), multi-index semi-parametric models (Zhu et al. 2011), nonparametric models (Fan et al. 2011, Liu et al. 2014) and quantile regression (He et al. 2013, Wu and Yin 2015) among others.

The idea of screening is to deliver a computationally efficient way to reduce the dimensionality of the feature space from a very high scale to a moderate

one. The researchers will then benefit both computationally and statistically from learning the data in a much reduced feature space. Besides the sure screening property, we argue an appealing screening method should satisfy the following two properties. First, the screening method should be model free which means it can be implemented without specifying a regression model. In ultra-high dimensional regime, it is challenging if not impossible to specify a correct regression model with existence of the huge number of redundant features. Hence the model free property is desired as it guarantees the effectiveness of the screening method in the presence of model mis-specification. The model free screening method becomes a hot research topic in recent years, see (Zhu et al. 2011, Li et al. 2012, Mai and Zou 2015, He et al. 2013, Cui et al. 2015) and the references therein. The second property is data adaptive which means the screening method should be insensitive to the assumptions like independence, sub-Gaussianity and uni-variate response. Such assumptions are usually not realistic in ultra-high dimensional applications. Even when the assumptions are satisfied on the population level, they can be easily violated in the realized sample simply due to ultra-high dimensionality. Therefore the screening method which is sensitive to such assumptions may perform poorly in real applications. The data adaptive screening method also draws certain amount of attention recently. He et al. (2013), Wu and Yin (2015) and Ma et al. (2017) among others considered quantile based screening which adapts to heavy-tailed data. Wang (2012) and Fan, Ke and Wang (2016) developed screening methods for strongly correlated features.

Unfortunately, none of the aforementioned screening methods enjoy sure screening, model free and data adaptive properties simultaneously. For example, the SIS is tailored to the linear regression and depends on the Gaussian assumption. Li et al. (2012) developed a sure independence screening procedure based on the distance correlation which is model free. However its sure screening property requires sub-exponential assumptions for features and response. The Kolmogorov distance based screening method proposed in Mai and Zou (2012) is robust against heavy tailed data but only works for binary classification problems. Pan et al. (2016) proposed a pairwise sure screening procedure for linear discriminant analysis which requires balanced categories and is sensitive to the tail behavior of the features.

In this chapter, we propose a model free and data adaptive feature screening method. The proposed method is based on ranking the projection correlations

between features and the response. The projection correlation, proposed by Zhu et al. (2017), is a measure of dependence between two random vectors which enjoys several nice probability properties. Assume that two random vectors have continuous joint and marginal densities, the projection correlation equals zero if and only if the two random vectors are independent and is invariant to orthogonal transformations. The proposed screening procedure does not require specifying any regression model and is insensitive to the moment conditions of the dataset. As the projection correlation is dimension free to both random vectors, the proposed screening method can be applied to grouped features and multivariate response. For instance, we can find a parsimonious set of features that are jointly dependent with multivariate response. The theoretical analysis demonstrates the proposed method enjoys not only the sure screening property but also a stronger result called rank consistency property. The only condition required is a minimum signal gap between active and inactive features. The extensive simulated experiments show the proposed method wins the horse racing against its competitors on various scenarios.

The rest of chpater is organized as follows. The Section 5.2 introduces the projection correlation and its properties. We study the sample projection correlation and demonstrate its non-asymptotic properties. In Section 5.3, we propose a screening procedure based on the projection correlation. We show the proposed screening method satisfies sure screening and rank consistency properties under very mild conditions. The Section 5.4 provides various simulated experiments to assess the performance of the proposed screening procedure. The Section 5.5 further demonstrates the proposed screening method by a real data example.

5.2 Projection correlation

To begin with, we give some background on the projection correlation and its properties introduced in Zhu et al. (2017) to pave the way for the proposed screening procedure. Let $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$ be two random vectors. The projection correlation is elicited by the following independence testing problem.

$$H_0 : \mathbf{x} \text{ and } \mathbf{y} \text{ are independent} \quad \text{versus} \quad H_1 : \text{ otherwise.}$$

The null hypothesis holds if and only if $U = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}$ and $V = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{y}$ are independent for all unit vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Let $F_{U,V}(u,v)$ be the joint distribution of $(U,\ V)$, and let $F_U(u)$ and $F_V(v)$ be the marginal distributions of U and V. We can define the squared projection covariance as follows

$$\begin{aligned}\mathrm{Pcov}(\mathbf{x},\mathbf{y})^2 &= \iiint (F_{U,V}(u,v) - F_U(u)F_V(v))^2\, dF_{U,V}(u,v)\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta} \\ &= \iiint \mathbf{Cov}^2\{I(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} \le u),\ I(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{y} \le v)\}\, dF_{U,V}(u,v)\, d\boldsymbol{\alpha}\, d\boldsymbol{\beta},\end{aligned} \tag{5.1}$$

where $I(\cdot)$ is the indicator function and $\mathbf{Cov}(\cdot,\cdot)$ is the Pearson covariance. Furthermore, we define the projection correlation between $\mathbf{x}$ and $\mathbf{y}$ as the square root of

$$\mathrm{PC}(\mathbf{x},\mathbf{y})^2 = \frac{\mathrm{Pcov}(\mathbf{x},\mathbf{y})^2}{\mathrm{Pcov}(\mathbf{x},\mathbf{x})\mathrm{Pcov}(\mathbf{y},\mathbf{y})}, \tag{5.2}$$

and we follow the convention $0/0 = 0$. In general $0 \le \mathrm{PC}(\mathbf{x},\mathbf{y}) \le 1$, testing whether $\mathbf{x}$ and $\mathbf{y}$ are independent amounts testing whether $\mathrm{PC}(\mathbf{x},\mathbf{y}) = 0$. The following lemma in Zhu et al. (2017) presents some appealing properties of the projection correlation at the population level.

Lemma 5.1. *Let $(\mathbf{x},\mathbf{y})$ belong to the class of random vectors with continuous joint and marginal probability densities. Then we have*

(1) $\mathrm{PC}(\mathbf{x},\mathbf{y}) = 0$ *if and only if* $\mathbf{x}$ *and* $\mathbf{y}$ *are independent.*

(2) $\mathrm{PC}(\mathbf{x},\mathbf{x}) = 0$ *if and only if* $\mathbf{x} = E(\mathbf{x})$ *almost surely.*

(3) Let $\mathbf{D}_1 \in \mathbb{R}^{p\times p}$ *and* $\mathbf{D}_2 \in \mathbb{R}^{q\times q}$ *be two orthonormal matrices,* $\mathbf{a}_1 \in \mathbb{R}^p$ *and* $\mathbf{a}_2 \in \mathbb{R}^q$ *be two vectors, and* b_1 *and* b_2 *be two scalars. Then* $\mathrm{PC}(\mathbf{x},\mathbf{y}) = \mathrm{PC}(\mathbf{a}_1 + b_1\mathbf{D}_1\mathbf{x}, \mathbf{a}_2 + b_2\mathbf{D}_2\mathbf{y})$.

Remark. The first two properties state that the projection correlation is a measure of dependence between two random vectors. The third one indicates the projection correlation is invariant with respect to orthogonal transformations. The first property in Lemma 5.1 does not hold in general without the assumption that $(\mathbf{x},\mathbf{y})$ is jointly and marginally continuous random vectors. Hoeffding (1948) constructed a counterexample in which $\mathbf{x}$ and $\mathbf{y}$ are two dependent discrete random variables

but $\mathrm{PC}(\mathbf{x},\mathbf{y}) = 0$. However, our numerical results show that the PC may still work well and serve the purpose of screening when the response is discrete.

Zhu et al. (2017) gives an explicit formula for the squared projection covariance in (5.1). Let $(\mathbf{x}_1,\mathbf{y}_1),\ \ldots,\ (\mathbf{x}_5,\mathbf{y}_5)$ be 5 independent random copies of $(\mathbf{x},\mathbf{y})$, then

$$\begin{aligned}
\mathrm{Pcov}(\mathbf{x},\mathbf{y})^2 =& S_1 + S_2 - 2S_3 \\
=& E\left[\arccos\left\{\frac{(\mathbf{x}_1-\mathbf{x}_3)^{\mathrm{T}}(\mathbf{x}_4-\mathbf{x}_3)}{\|\mathbf{x}_1-\mathbf{x}_3\|\|\mathbf{x}_4-\mathbf{x}_3\|}\right\}\arccos\left\{\frac{(\mathbf{y}_1-\mathbf{y}_3)^{\mathrm{T}}(\mathbf{y}_4-\mathbf{y}_3)}{\|\mathbf{y}_1-\mathbf{y}_3\|\|\mathbf{y}_4-\mathbf{y}_3\|}\right\}\right] \\
&+E\left[\arccos\left\{\frac{(\mathbf{x}_1-\mathbf{x}_3)^{\mathrm{T}}(\mathbf{x}_4-\mathbf{x}_3)}{\|\mathbf{x}_1-\mathbf{x}_3\|\|\mathbf{x}_4-\mathbf{x}_3\|}\right\}\arccos\left\{\frac{(\mathbf{y}_2-\mathbf{y}_3)^{\mathrm{T}}(\mathbf{y}_5-\mathbf{y}_3)}{\|\mathbf{y}_2-\mathbf{y}_3\|\|\mathbf{y}_5-\mathbf{y}_3\|}\right\}\right] \\
&-2E\left[\arccos\left\{\frac{(\mathbf{x}_1-\mathbf{x}_3)^{\mathrm{T}}(\mathbf{x}_4-\mathbf{x}_3)}{\|\mathbf{x}_1-\mathbf{x}_3\|\|\mathbf{x}_4-\mathbf{x}_3\|}\right\}\arccos\left\{\frac{(\mathbf{y}_2-\mathbf{y}_3)^{\mathrm{T}}(\mathbf{y}_4-\mathbf{y}_3)}{\|\mathbf{y}_2-\mathbf{y}_3\|\|\mathbf{y}_4-\mathbf{y}_3\|}\right\}\right],
\end{aligned} \tag{5.3}$$

where S_1, S_2 and S_3 are defined in an obvious manner and $\|\cdot\|$ represents the L_2 Euclidean norm. Equation (5.3) shows that the projection covariance only depends on the vectors through $(\mathbf{x}_k-\mathbf{x}_l)/\|\mathbf{x}_k-\mathbf{x}_l\|$ and $(\mathbf{y}_k-\mathbf{y}_l)/\|\mathbf{y}_k-\mathbf{y}_l\|$ whose second moments are unity. This gives us the intuition that the projection covariance is free of the moment conditions on $(\mathbf{x},\mathbf{y})$ which are usually required by some other measurements, such as distance correlation (Li et al. 2012).

Let $\mathbf{X} = (\mathbf{x}_1,\ \ldots\ ,\mathbf{x}_n)^{\mathrm{T}}$ and $\mathbf{Y} = (\mathbf{x}_1,\ \ldots\ ,\mathbf{y}_n)^{\mathrm{T}}$ be an observed sample of $(\mathbf{x},\mathbf{y})$. Equation (5.3) leads to a straightforward empirical estimate of $\mathrm{Pcov}(\mathbf{x},\mathbf{y})^2$ based on U-statistic, yet it is difficult to calculate (Székely and Rizzo 2010). An equivalent form of the U-statistic is given in Zhu et al. (2017). In particular, the sample projection variance and covariance of $\mathbf{X}$ and $\mathbf{Y}$ can be calculated as

$$\begin{aligned}
\widehat{\mathrm{Pcov}}(\mathbf{X},\mathbf{Y}) &= \left\{n^{-3}\sum_{k,l,r=1}^{n} A_{klr}B_{klr}\right\}^{1/2}, \\
\widehat{\mathrm{Pcov}}(\mathbf{X},\mathbf{X}) &= \left\{n^{-3}\sum_{k,l,r=1}^{n} A_{klr}^2\right\}^{1/2}, \quad \text{and} \\
\widehat{\mathrm{Pcov}}(\mathbf{Y},\mathbf{Y}) &= \left\{n^{-3}\sum_{k,l,r=1}^{n} B_{klr}^2\right\}^{1/2},
\end{aligned} \tag{5.4}$$

where for $k, l, r = 1, \cdots, n$,

$$
\begin{aligned}
a_{klr} &= \arccos\left\{\frac{(\mathbf{x}_k - \mathbf{x}_r)^{\mathrm{T}}(\mathbf{x}_l - \mathbf{x}_r)}{\|\mathbf{x}_k - \mathbf{x}_r\|\|\mathbf{x}_l - \mathbf{x}_r\|}\right\}, \quad a_{klr} = 0 \text{ if } k = r \text{ or } l = r, \\
\bar{a}_{k\cdot r} &= n^{-1}\sum_{l=1}^{n} a_{klr}, \quad \bar{a}_{\cdot lr} = n^{-1}\sum_{k=1}^{n} a_{klr}, \quad \bar{a}_{\cdot\cdot r} = n^{-2}\sum_{k=1}^{n}\sum_{l=1}^{n} a_{klr}, \\
A_{klr} &= a_{klr} - \bar{a}_{k\cdot r} - \bar{a}_{\cdot lr} + \bar{a}_{\cdot\cdot r}, \\
b_{klr} &= \arccos\left\{\frac{(\mathbf{y}_k - \mathbf{y}_r)^{\mathrm{T}}(\mathbf{y}_l - \mathbf{y}_r)}{\|\mathbf{y}_k - \mathbf{y}_r\|\|\mathbf{y}_l - \mathbf{y}_r\|}\right\}, \quad b_{klr} = 0 \text{ if } k = r \text{ or } l = r, \\
\bar{b}_{k\cdot r} &= n^{-1}\sum_{l=1}^{n} b_{klr}, \quad \bar{b}_{\cdot lr} = n^{-1}\sum_{k=1}^{n} b_{klr}, \quad \bar{b}_{\cdot\cdot r} = n^{-2}\sum_{k=1}^{n}\sum_{l=1}^{n} b_{klr}, \\
B_{klr} &= b_{klr} - \bar{b}_{k\cdot r} - \bar{b}_{\cdot lr} + \bar{b}_{\cdot\cdot r}.
\end{aligned}
$$

Then the sample projection correlation between $\mathbf{X}$ and $\mathbf{Y}$ is defined as the square root of

$$
\widehat{\mathrm{PC}}(\mathbf{X}, \mathbf{Y})^2 = \frac{\widehat{\mathrm{Pcov}}(\mathbf{X}, \mathbf{Y})^2}{\widehat{\mathrm{Pcov}}(\mathbf{X}, \mathbf{X})\widehat{\mathrm{Pcov}}(\mathbf{Y}, \mathbf{Y})}. \tag{5.5}
$$

Based on (5.4), the sample projection correlation can be computed in $O(n^3)$.

We first provide exponential-type deviation inequalities for sample projection covariance and correlation.

Theorem 5.1. *For any $0 < \varepsilon < 1$, as long as $n \geq 10\pi^2/\varepsilon$, there exists positive constants c_1 and c_2, such that*

$$
\Pr\left\{|\widehat{\mathrm{Pcov}}(\mathbf{X}, \mathbf{Y})^2 - \mathrm{Pcov}(\mathbf{x}, \mathbf{y})^2| > \varepsilon\right\} \leq c_1 \exp\{-c_2 n\varepsilon^2\},
$$

and

$$
\Pr\left\{|\widehat{\mathrm{PC}}(\mathbf{X}, \mathbf{Y})^2 - \mathrm{PC}(\mathbf{x}, \mathbf{y})^2| > \varepsilon\right\} \leq 5c_1 \exp\{-c_2 \sigma n\varepsilon^2\},
$$

where $\sigma = \min\{\sigma_x^3\sigma_y^3/4M^4, \sigma_x^2\sigma_y^2/4M^4, \sigma_x\sigma_y/4\}$, $\sigma_x = \mathrm{Pcov}(\mathbf{x}, \mathbf{x})^2$, $\sigma_y = \mathrm{Pcov}(\mathbf{y}, \mathbf{y})^2$ and $M = (2\pi)^2$.

The proof of Theorem 5.1 is based on an exponential-type deviation inequality for U-statistic and can be found in Section 5.6. The above exponential-type inequalities do not depend on the dimensionality and moment conditions of both random vectors. The exception probability decays exponentially with sample size n which guarantees good finite sample performance of the proposed estimators.

5.3 A model free feature screening procedure

In the section, we propose a model free screening procedure utilizing the nice properties of projection correlation. Let $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}}$ be vector of q response variables and $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be a vector of p features. To avoid the trivial discussion, we restrict ourselves to the random designed case, i.e. $\min_{1\le k\le p} \mathrm{Pcov}(X_k, X_k)^2 \ge \sigma^2$ and $\min_{1\le k\le q} \mathrm{Pcov}(Y_k, Y_k)^2 \ge \sigma^2$ for some $\sigma > 0$.

Denote by $F(\mathbf{y}|\mathbf{x})$ the conditional distribution function of $\mathbf{y}$ given $\mathbf{x}$. Without specifying any regression model on $\mathbf{y}$ and $\mathbf{x}$, we define the index set of active features by

$$\mathcal{A} = \{k : F(\mathbf{y}|\mathbf{x}) \text{ functionally depends on } X_k, k = 1, \ldots, p\}.$$

The number of active features is $s = |\mathcal{A}|$, where $|\mathcal{A}|$ denotes the the cardinality of $\mathcal{A}$. The features that do not belong to $\mathcal{A}$ are called inactive features. We use $\mathcal{A}^c$, the complement of $\mathcal{A}$, to denote the index set of all inactive features. The above setting abstracts a large number of sparse regression problems including linear model, generalized linear model, additive model, semi-parametric model, non-linear model and so on. In addition, it allows multivariate response and grouped predictor.

In practice, one observes a random sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}$, $i = 1, \ldots, n$ of $(\mathbf{x}, \mathbf{y})$ and let $\mathbf{X} = (\mathbf{x}_1, \ \ldots \ , \mathbf{x}_n)^{\mathrm{T}}$ and $\mathbf{Y} = (\mathbf{x}_1, \ \ldots \ , \mathbf{y}_n)^{\mathrm{T}}$. In ultra-high dimensional regime, it is natural to assume that the number of features p greatly exceeds the sample size n but the number of active features s is smaller than n. For a given feature X_k, $k = 1, \ldots, p$, a sufficient condition for X_k to be an inactive feature is the independence between X_k and $\mathbf{y}$. This intuition together with Theorem 5.1 encourages us to screen out the features whose projection correlations with $\mathbf{y}$ are small. As a result, we select a set contains active features as follows

$$\widehat{\mathcal{A}}(\delta) = \{k : \widehat{\mathrm{PC}}(\mathbf{X}_k, \mathbf{Y}) \ge \delta, 1 \le k \le p\},$$

where δ is a pre-specified threshold value, $\mathbf{X}_k$ is the kth column of $\mathbf{X}$. With a proper choice of δ, we show that the proposed feature screening procedure enjoys the sure screening property, which states that with probability approaching to 1, all active features are selected in $\widehat{\mathcal{A}}(\delta)$. We first impose the following minimum signal strength type conditions.

Condition 1 (Minimum signal strength)

Denote $\omega_k \equiv \mathrm{PC}(X_k, \mathbf{y})^2$ the population squared projection correlation between the kth feature and the response.

(a) For some $c_3 > 0$ and $0 \leq \kappa < 1/2$, $\min_{k \in \mathcal{A}} \omega_k \geq 2c_3 n^{-\kappa}$.

(b) For some $c_3 > 0$ and $0 \leq \kappa < 1/2$, $\min_{k \in \mathcal{A}} \omega_k - \max_{k \in \mathcal{A}^c} \omega_k > 2c_3 n^{-\kappa}$.

Remark. The Condition 1 (a) is a minimum signal strength condition that assumes the squared projection correlations between the active features and response should be uniformly lower bounded and do not converge to zero too fast as n diverges. The Condition 1 (b) puts an assumption on the gap of signal strength between active and inactive features. Condition 1 (a) is implied by Condition 1 (b) since ω_k is always non-negative.

The following two theorems state the sure screening property and rank consistency property of the proposed screening procedure. We refer the proposed feature screening procedure as PC-SIS.

Theorem 5.2 (Sure screening property)**.** *Under Condition 1 (a), choose* $\delta \leq \min_{k \in \mathcal{A}} \omega_k/2$*, then we have*

$$\Pr\Big(\mathcal{A} \subseteq \widehat{\mathcal{A}}(\delta)\Big) \geq 1 - O\Big(s \exp\{-c_4 n \delta^2\}\Big), \tag{5.6}$$

where c_4 *is a positive constant.*

Remark. In Theorem 5.2, if we set $\delta = c_3 n^{-\kappa}$, which satisfies the condition $\delta \leq \min_{k \in \mathcal{A}} \omega_k/2$, then we have

$$\Pr\Big(\mathcal{A} \subseteq \widehat{\mathcal{A}}(\delta)\Big) \geq 1 - O(s \exp\{-c_3^2 c_4 n^{1-2\kappa}\}). \tag{5.7}$$

From equation (5.7), we know that with the choice $\delta = c_3 n^{-\kappa}$, all active features are selected with probability approaching to 1 as $n \to \infty$. In fact, any choice of $\delta \leq c_3 n^{-\kappa}$ leads to the sure screening property. With the same choice of $\delta = c_3 n^{-\kappa}$, Li et al. (2012) showed that the distance correlation based screening method (DC-SIS) satisfies

$$\Pr\Big(\mathcal{A} \subseteq \widehat{\mathcal{A}}(\delta)\Big) \geq 1 - O(s \exp\{-c_4' n^{1-2(\kappa+\eta)}\} + n \exp\{-c_4'' n^{\eta}\}),$$

where c_4', c_4'' and η are positive constants. Our PC-SIS achieves a faster rate since it does not have the extra second term $n\exp\{-c_4''n^{\eta}\}$ and the extra η in the power of the first term. The faster rate of PC-SIS is due to the fact that projection correlation does not require the existence of any moments.

Theorem 5.3 (Rank consistency property)**.** *Under Condition 1 (b), we have*

$$\Pr\left(\min_{k\in\mathcal{A}}\widehat{\omega}_k-\max_{k\in\mathcal{A}^c}\widehat{\omega}_k>0\right)>1-O(p\exp\{-c_5n^{1-2\kappa}\}),$$

where c_5 is some positive constant. If $\log p=o(n^{1-2\kappa})$ with $0\le\kappa<1/2$, then we have

$$\liminf_{n\to\infty}\left(\min_{k\in\mathcal{A}}\widehat{\omega}_k-\max_{k\in\mathcal{A}^c}\widehat{\omega}_k\right)>0,\ \textit{almost surely.}$$

Remark. The rank consistency in Theorem 5.3 is a stronger result than sure screening property. When the signal gap between active and inactive features satisfies Condition 1 (b), the active features are always ranked ahead of the inactive ones with high probability. In other words, there exists a choice of δ on the solution path that can perfectly separate the active and inactive sets with high probability.

5.4 Simulated experiments

In this section, we asses the finite sample performance of the proposed projection correlation based feature screening procedure and compare it with sure independence screening (Fan and Lv 2008), distance correlation based screening (Li et al. 2012) and the bias-corrected distance correlation based screening (Székely and Rizzo 2014) under various scenarios. We denote our method by PC-SIS and the competitors by SIS, DC-SIS and bcDC-SIS respectively. In the following regression models, we set $n=100$ and $p=5000,10000$ to mimic the ultra-high dimensional setting. For each scenario, we repeat 200 replications. For each replication, we rank the features in descending order by the above four screening criteria and record the minimum model size that contains all active features. The screening performance is measured by the $5\%,25\%,50\%,75\%,95\%$ quantiles of the minimum model size of each screening method. Throughout this section, we denote $\mathbf{\Sigma}=(\sigma_{ij})_{p\times p}$, where $\sigma_{ij}=0.5^{|i-j|}$.

5.4.1 Linear and generalized linear model

Consider the following linear model

$$Y = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon,$$

where $\boldsymbol{\beta} = (\mathbf{1}_5^{\mathrm{T}}, \mathbf{0}_{p-5}^{\mathrm{T}})^{\mathrm{T}}$. We generate covariates $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ and ϵ from the following 4 scenarios.

Model 1.a: $\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\epsilon \sim N(0, 1)$.

Model 1.b: $\mathbf{x} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\epsilon \sim \mathrm{Cauchy}(0, 1)$.

Model 1.c: $\mathbf{u} \sim \mathrm{Cauchy}(\mathbf{0}, \mathbf{I})$, $\mathbf{x} = \boldsymbol{\Sigma}^{1/2}\mathbf{u}$ and $\epsilon \sim N(0, 1)$.

Model 1.d: $\mathbf{u} \sim \mathrm{Cauchy}(\mathbf{0}, \mathbf{I})$, $\mathbf{x} = \boldsymbol{\Sigma}^{1/2}\mathbf{u}$ and $\epsilon \sim \mathrm{Cauchy}(0, 1)$.

$\mathrm{Cauchy}(0, 1)$ stands for the standard Cauchy distribution. The Cauchy distribution is a heavy-tailed distribution and does not have any finite moment. In Model 1.a, both covariates and error are normally distributed and no heavy-tailed distribution is involved. In Models 1.b - 1.d, at least one of $\mathbf{x}$ and ϵ is a heavy-tailed distribution, i.e., Cauchy distribution, making the task of feature screening more challenging. We also consider the following two generalized linear models

Model 1.e: $Y = \exp\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\} + \epsilon$, where $\epsilon \sim N(0, 1)$.

Model 1.f: (Poisson regression) $Y \sim \mathrm{Poisson}(\exp\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\})$.

We generate $\mathbf{x}$ from $N(\mathbf{0}, \boldsymbol{\Sigma})$ and set $\boldsymbol{\beta} = (\mathbf{2}_5^{\mathrm{T}}, \mathbf{0}_{p-5}^{\mathrm{T}})^{\mathrm{T}}$. Model 1.e is the well known Poisson regression model in which the response Y is the count data. The difference between Model 1.e and Model 1.f is the that the response in Model 1.e is continuous while the response in Model 1.f is discrete. Notice that the property $\mathrm{PC}(X_j, Y) = 0$ if and only if X_j and Y are independent may not hold if X_j and Y are discrete random variables (see Lemma 5.1). The numerical results show that the proposed PC-SIS still has good performance when the response is discrete.

There are 5 active features in all the scenarios. The different quantiles of minimum model size are summarized in Table 5.1. For Model 1.a, all the four methods can select all active variables almost perfectly since both covariates $\mathbf{x}$ and

error term ϵ are normal distribution. For Models 1.b-1.d, PC-SIS outperforms all other 3 methods. The SIS completely fails at the presence of Cauchy distribution since the Pearson correlation requires the existence of first and second moments. The bcDC-SIS has a better performance than DC-SIS. The 75% and 95% quantiles of minimum model size of PC-SIS is roughly half of that of bcDC-SIS. Even for the most challenging scenario where both $\mathbf{x}$ and ϵ are Cauchy distribution, the PC-SIS still works reasonably well which indicates that the PC-SIS is more robust to heavy-tailed distribution and outliers. For the generalized linear models 1.e and 1.f, the PC-SIS performs much better than the other three methods. The PC-SIS can recover the true active set with a model size close to 5 while the other three methods perform as bad as random guesses.

5.4.2 Nonparametric model

Consider the following four non-linear models

Model 2.a: $Y = 5X_1 + 2\sin(\pi X_2/2) + 2X_3\mathbf{1}\{X_3 > 0\} + 2\exp\{5X_4\} + \epsilon.$

Model 2.b: $Y = 3X_1 + 3X_2^3 + 3X_3^{-1} + 5\mathbf{1}\{X_4 > 0\} + \epsilon.$

Model 2.c: $Y = 1 - 5(X_2 + X_3)^3 \exp\{5(X_1 + X_4^2)\} + \epsilon.$

Model 2.d: $Y = 1 - 5(X_2 + X_3)^{-3} \exp\{1 + 10\sin(\pi X_1/2) + 5X_4\} + \epsilon.$

Models 2.a - 2.b are also known as additive model and Models 2.c - 32.d are more general nonparametric models. We generate $\mathbf{x} \sim N(\mathbf{0}, \mathbf{\Sigma})$ and $\epsilon \sim N(0, 1)$. The number of active features in the four models are 4 and the simulation results are summarized in Table 5.2. For Model 2.a, the median of the minimum model size of the PC-SIS is exactly the same as the number of active features while the other three methods need a much larger model size to recover the active set. The PC-SIS performs slightly worse for Model 2.b due to the high intrinsic difficulty of the model, but still outperforms other three methods by a big margin. For the nonparametric models 2.c - 2.d, the PC-SIS performs reasonably well while the other methods fail to effectively screen out the inactive features. The 95% quantile of SIS, DC-SIS and bcDC-SIS are almost as large as p. This shows, in the worst case scenario, SIS, DC-SIS and bcDC-SIS are hopeless to reduce the dimensionality without screening out active features.

Table 5.1: The quantiles of minimum model size for linear and generalized liner models out of 200 replications.

	Model 1.a					**Model 1.b**				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$p = 5000$										
PC-SIS	5.0	5.0	5.0	5.0	8.0	5.0	5.0	6.0	14.0	125.0
DC-SIS	5.0	5.0	5.0	5.0	6.0	5.0	5.0	10.0	81.2	1305.0
bcDC-SIS	5.0	5.0	5.0	5.0	6.0	5.0	5.0	6.0	20.5	231.4
SIS	5.0	5.0	5.0	5.0	5.0	6.0	238.0	1833.0	3878.5	4915.0
$p = 10000$										
PC-SIS	5.0	5.0	5.0	5.0	7.0	5.0	5.0	8.0	23.0	233.5
DC-SIS	5.0	5.0	5.0	5.0	6.0	5.0	6.8	21.0	204.2	3511.2
bcDC-SIS	5.0	5.0	5.0	5.0	6.0	5.0	5.0	10.0	43.2	718.8
SIS	5.0	5.0	5.0	5.0	5.0	16.0	703.2	3418.5	7432.0	9651.0
	Model 1.c					**Model 1.d**				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$p = 5000$										
PC-SIS	5.0	5.0	6.0	8.0	50.9	5.0	5.0	6.0	10.2	139.9
DC-SIS	5.0	8.0	42.5	143.0	722.5	5.0	16.0	54.0	189.0	701.6
bcDC-SIS	5.0	5.0	6.0	14.2	80.4	5.0	5.0	8.0	21.8	156.8
SIS	5.0	39.0	81.5	374.0	3244.4	5.0	45.8	130.5	523.5	3241.0
$p = 10000$										
PC-SIS	5.0	5.0	6.0	8.0	92.4	5.0	6.0	6.0	13.0	176.2
DC-SIS	5.0	18.2	78.0	277.5	1567.5	5.0	30.8	113.0	410.0	2036.5
bcDC-SIS	5.0	5.0	7.0	19.2	179.1	5.0	5.0	10.0	27.0	412.6
SIS	6.0	58.8	180.5	773.2	4189.0	8.9	73.0	244.5	959.8	5777.7
	Model 1.e					**Model 1.f**				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$p = 5000$										
PC-SIS	5.0	5.0	5.0	5.0	17.2	5.0	5.0	5.0	5.0	7.0
DC-SIS	90.3	396.0	897.5	1762.5	3787.3	76.3	396.0	893.5	1764.8	3471.8
bcDC-SIS	22.7	82.2	259.5	669.2	2358.7	15.0	87.2	266.5	821.2	2471.5
SIS	178.6	604.8	1137.0	2319.8	4303.4	186.0	606.2	1210.5	2253.0	4261.6
$p = 10000$										
PC-SIS	5.0	5.0	5.0	6.0	23.0	5.0	5.0	5.0	5.0	12.0
DC-SIS	138.0	788.8	1878.5	3301.0	6831.9	154.2	729.2	1725.5	3424.0	6832.8
bcDC-SIS	45.7	163.5	534.5	1520.5	5415.2	30.0	175.8	509.0	1513.2	5580.0
SIS	462.8	1276.8	2460.0	4164.5	8622.5	512.6	1271.2	2484.5	4281.0	8416.3

Table 5.2: The quantiles of minimum model size for nonparametric model out of 200 replications.

	Model 2.a					Model 2.b				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$p = 5000$										
PC-SIS	4.0	4.0	4.0	5.2	19.1	4.0	5.0	9.5	26.5	261.9
DC-SIS	600.0	1880.2	2994.5	4052.2	4701.3	4.0	7.8	61.0	541.5	2310.9
bcDC-SIS	488.4	1480.8	2863.0	3967.8	4855.9	4.0	6.0	21.5	88.0	893.8
SIS	709.0	2065.8	3062.5	4160.0	4869.4	54.0	658.5	2692.5	4213.0	4829.1
$p = 10000$										
PC-SIS	4.0	4.0	4.0	5.0	31.0	4.0	5.8	13.0	48.8	393.9
DC-SIS	664.0	3162.5	5655.5	7605.8	9490.1	4.0	13.8	86.0	843.2	5529.2
bcDC-SIS	663.4	2605.8	5578.5	7745.2	9208.4	4.0	8.0	24.5	150.5	1403.9
SIS	986.2	4048.2	6068.0	8298.5	9752.6	64.5	1193.8	4488.0	8139.5	9725.6
	Model 2.c					**Model 2.d**				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$p = 5000$										
PC-SIS	4.0	4.0	4.0	6.0	21.3	4.0	5.0	9.5	36.0	169.4
DC-SIS	397.6	1536.8	2750.5	3930.0	4721.8	1639.1	3138.5	3851.5	4434.8	4902.4
bcDC-SIS	196.5	1267.5	2774.0	4236.0	4986.2	605.5	1251.8	2073.5	2972.5	4209.1
SIS	421.2	1615.8	2920.0	4057.0	4761.0	2065.0	3487.8	4083.5	4659.0	4950.3
$p = 10000$										
PC-SIS	4.0	4.0	4.0	7.2	25.0	4.0	5.0	13.0	62.5	297.4
DC-SIS	668.8	3628.2	6243.5	7920.5	9531.6	3409.6	6380.8	7705.0	8756.5	9790.6
bcDC-SIS	288.2	2006.5	4714.5	7370.2	9869.9	776.4	2567.8	4154.5	5517.8	8251.0
SIS	760.9	3609.5	6155.5	8129.8	9577.1	3333.4	6387.8	8007.5	9252.8	9847.2

5.4.3 Multivariate response model

In the last experiment, we investigate the performance of the PC-SIS for a multivariate response problem. Here we ignore the the SIS as it cannot be directly applied to multivariate response problem. We generate $\mathbf{y} = (Y_1, Y_2)^{\mathrm{T}}$ from a bivariate normal distribution with conditional mean $\boldsymbol{\mu}_{\mathbf{y}|\mathbf{x}} = (\mu_1, \mu_2)^{\mathrm{T}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} = (\sigma_{ij})_{2\times 2}$, where $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \sigma_{21} = \sigma(\mathbf{x})$, a function of $\mathbf{x}$. Following the settings in Li et al. (2012), we consider the following two models.

Model 3.a: $\mu_1 = \exp\{2(X_1 + X_2)\}$, $\mu_2 = X_3 + X_4$ and $\sigma(\mathbf{x}) = \sin(\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta})$.

Model 3.b: $\mu_1 = 2\sin(\pi X_1/2) + X_3 + \exp\{1 + X_4\}$, $\mu_2 = X_1^{-2} + X_2$ and $\sigma(\mathbf{x}) = (\exp\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}\} - 1)/(\exp\{\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta}\} + 1)$.

We set $\boldsymbol{\theta} = (\mathbf{2}_4^{\mathrm{T}}, \mathbf{0}_{p-4}^{\mathrm{T}})^{\mathrm{T}}$. The above two models allow μ_1 and μ_2 to depend on different active sets and the union of their active sets contains 4 features. We also

Table 5.3: The quantiles of minimum model size for multivariate response model out of 200 replications.

	Model 3.a					Model 3.b				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$p = 5000$										
PC-SIS	4.0	4.0	4.0	4.0	6.0	4.0	4.0	6.0	18.0	114.4
DC-SIS	53.0	463.0	1211.0	2349.0	3774.2	641.1	2308.8	3307.0	4257.8	4838.0
bcDC-SIS	24.0	215.5	758.5	1965.0	3999.8	225.2	1270.2	2494.5	3596.8	4709.0
$p = 10000$										
PC-SIS	4.0	4.0	4.0	4.0	9.0	4.0	4.0	8.0	30.2	302.9
DC-SIS	136.8	978.5	2237.5	4506.0	8106.6	1804.7	4510.5	6445.5	8153.5	9707.6
bcDC-SIS	83.0	546.0	1828.5	4264.5	7711.3	444.4	2217.0	4695.0	7122.5	9076.4

allow the covariance matrix $\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}}$ to depend on the covariates. The simulation results are summarized in Table 5.3. Again, the PC-SIS method performs strikingly well compared to the other methods.

5.5 Real data example

In this section, we apply the proposed PC-SIS to a microarray dataset, which is from a transgenic mouse model of dilated cardiomyopathy. The mice overexpress a G protein-coupled receptor, designated Ro1, that is a mutated form of the human kappa opioid receptor. The aim was to determine which are the influential genes for overexpression of Ro1 in mice. The research is related to understanding different types of human heart disease. The expression of Ro1 (the response) was measured for $n = 30$ mice and $p = 6319$ genetic expression levels were obtained for each mice. This dataset was also analyzed by Segal et al. (2003) Hall and Miller (2009) and Li et al. (2012).

We apply both PC-SIS and DC-SIS to the dataset and rank the features based on projection correlation and distance correlation respectively. PC-SIS ranks Msa.5799.0 and Msa.21346.0 as the two most important features while DC-SIS ranks Msa.21346.0 and Msa.28772.0 as the two most important features. Msa.21346.0 is shared by the two methods. In practice, we may choose the top $\lceil n/\log n \rceil = 9$ (Fan and Lv 2008) features as the important features for further analysis. Table 5.4 shows the top 9 features selected by PC-SIS and DC-SIS. Among the top 9 features, 6 features are selected by both methods. We fit linear regression model

and nonparametric additive model with the selected 9 features for each method. The adjusted R^2 is reported in Table 5.5, which shows PC-SIS works slightly better than PC-SIS for both linear model and additive model with top 9 selected features in terms of prediction.

Table 5.4: Top 9 features identified by PC-SIS and DC-SIS. The gene names in bold are the common genes selected by both methods.

Ranking	1	2	3	4	5
PC-SIS	Msa.5799.0	**Msa.21346.0**	Msa.702.0	**Msa.11662.0**	**Msa.1545.0**
DC-SIS	**Msa.21346.0**	**Msa.28772.0**	Msa.2603.0	Msa.559.0	**Msa.1591.0**
Ranking	6	7	8	9	
PC-SIS	**Msa.1591.0**	**Msa.1011.0**	Msa.573.0	**Msa.28772.0**	
DC-SIS	**Msa.11662.0**	Msa.24000.0	**Msa.1545.0**	**Msa.1011.0**	

Table 5.5: The adjusted R^2 for linear and additive models with the top 9 features identified by PC-SIS and DC-SIS.

	Linear model	**Additive model**
PC-SIS	0.7780	0.7780
DC-SIS	0.7716	0.7720

Figure 5.1 gives the scatter plots for Msa.5799.0 and Msa.21346.0, the most important features detected by PC-SIS. The left panel of Figure 5.1 clearly shows that there is a nonlinear relationship between Ro1 and Msa.5799.0: when Msa.5799.0 expression level is relatively low (< 0.5), Ro1 expression decreases dramatically as Msa.5799.0 increases; when Msa.5799.0 expression level is relatively high (> 0.5), the Ro1 expression level stays roughly flat. DC-SIS misses Msa.5799.0 and only ranks it as the 18th most important feature. The right panel of Figure 5.1 also shows a clear nonlinear relationship between Ro1 and Msa.21346.0. In addition, Figure 5.1 indicates that there are potential outliers, that are marked in red triangles. The red dash curves and blue solid curves are fitted regression lines by local polynomial regression with and without potential outliers. After the removal of potential outliers, the nonlinear relationships become more clear, see the blue curves in Figure 5.1.

The existence of influential points motivates us to examine the robustness of the PC-SIS and the DC-SIS. To see how the potential outliers affect the ranking of Msa.5799.0, we remove one data point at one time and then obtain the ranking of

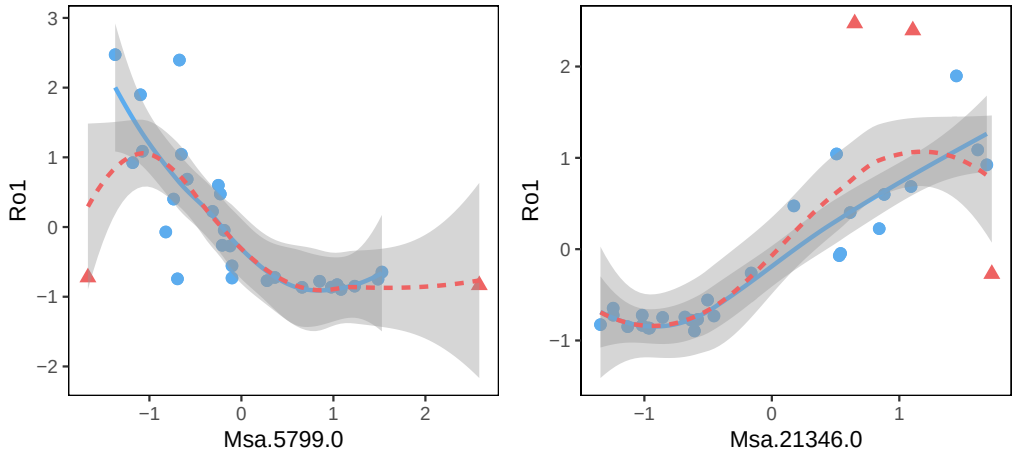


Figure 5.1: Scatter plots for Msa.5799.0 and Msa.21346.0. The red triangles are the potential outliers. The red dash curves and blue solid curves are fitted regression lines by local polynomial regression with and without the potential outliers. The gray shadows are the 95% confidence interval.

Msa.5799.0 by the PC-SIS and the DC-SIS respectively using the remaining 29 data points. We repeat this procedure for $n = 30$ times and thus obtain 30 rankings of Msa.5799.0 for each method. The left panel in Figure 5.2 shows the boxplot of the 30 rankings. As we can see, the PC-SIS always ranks Msa.5799.0 as the 1st or 2nd most important feature while its rankings given by the DC-SIS varies from 5 to 35. To further distinguish the PC-SIS and the DC-SIS, instead of removing one data point, we contaminate one data point at one time for Msa.21346.0. More specifically, for each $i = 1, \ldots, 30$, we replace the ith data point (X_i, Y_i) by (X_i', Y_i') where X_i' is generated from Uniform$(-1.5, -1)$ and Y_i' is generated from Uniform$(2.0, 2.5)$ such that (X_i', Y_i') is very likely to be a potential outlier. The right panel in Figure (5.2) shows that the rankings given by the PC-SIS has less variance than the DC-SIS. Both boxplots show that the PC-SIS is much more robust to potential outliers than the DC-SIS.

5.6 Proofs

Before proving Theorem 5.1, we introduce two useful lemmas. The first lemma is based on Theorem 5.6.1.A in Serfling (1980), which gives a probability equality for U-statistics.

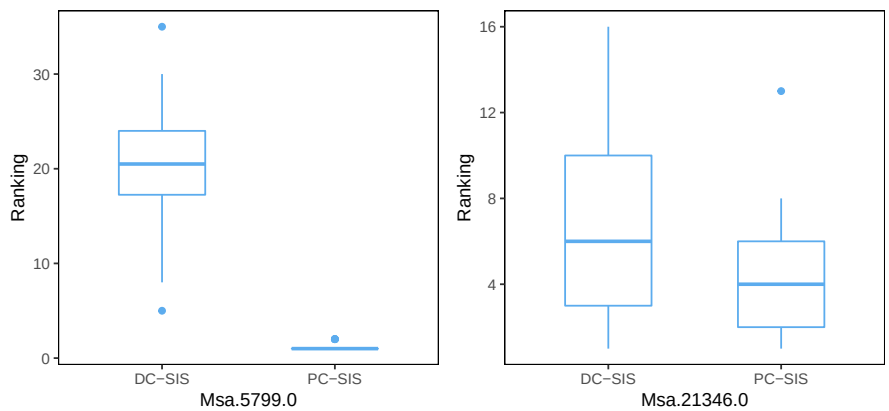


Figure 5.2: Boxplots of rankings for Msa.5799.0 and Msa.21346.0.

Lemma 5.2. *Let $h(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ be a kernel of the U-statistic U_n, and $\theta = E\{h(\mathbf{x}_1, \ldots, \mathbf{x}_m)\}$. If $a \leq h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \leq b$, then for any $t > 0$ and $n \geq m$,*

$$\Pr(|U_n - \theta| \geq t) \leq 2\exp\{-2\lfloor n/m \rfloor t^2/(b-a)^2\},$$

where $\lfloor n/m \rfloor$ denotes the integer part of n/m.

Proof. By Theorem 5.6.1.A of Serfling (1980), we have

$$\Pr(U_n - \theta \geq t) \leq \exp\{-2\lfloor n/m \rfloor t^2/(b-a)^2\}.$$

Due to the symmetry of U-statistic, we have

$$\Pr(|U_n - \theta| \geq t) \leq 2\exp\{-2\lfloor n/m \rfloor t^2/(b-a)^2\}.$$

□

The next theorem establishes the connection between the exponential-type deviation inequalities for sample covariance and sample correlation.

Lemma 5.3. *Suppose $\widehat{\gamma}_1$, $\widehat{\gamma}_2$ and $\widehat{\gamma}_3$ are estimates of parameters γ_1, γ_2 and γ_3 based on a size-n sample, respectively. Assume $\gamma_2 > 0$, $\gamma_3 > 0$ and $M \geq 2\max\{\gamma_1, \gamma_2, \gamma_3, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\gamma}_3\}$. If*

$$\Pr(|\widehat{\gamma}_k - \gamma_k| > \varepsilon) \leq c_1 \exp\{-c_2 n\varepsilon^2\}, \ k = 1, 2, 3,$$

for some positive constants c_1, c_2. Then we have

$$\Pr\left\{\left|\frac{\widehat{\gamma}_1}{\sqrt{\widehat{\gamma}_2\widehat{\gamma}_3}} - \frac{\gamma_1}{\sqrt{\gamma_2\gamma_3}}\right| > \varepsilon\right\} \leq 5c_1 \exp\{-c_2 n\varepsilon^2\gamma_0\},$$

where $\gamma_0 = \min\{\gamma_2^2\gamma_3^2/4M^4, \gamma_2^3\gamma_3^3/4M^4, \gamma_2\gamma_3/4\}$.

Proof. Since $\gamma_1, \gamma_2, \gamma_3, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\gamma}_3$ are bounded by $M/2$, it is easy to verify that

$$\Pr(|\widehat{\gamma}_2\widehat{\gamma}_3 - \gamma_2\gamma_3| > 2\varepsilon) \leq 2c_1 \exp\{-c_2 n\varepsilon^2/4M^2\}, \text{ and}$$

$$\Pr(|\sqrt{\widehat{\gamma}_2\widehat{\gamma}_3} - \sqrt{\gamma_2\gamma_3}| > 2\varepsilon) \leq 2c_1 \exp\{-c_2 n\varepsilon^2\gamma_2\gamma_3/4M^2\}. \tag{5.8}$$

Let $\gamma = \sqrt{\gamma_2\gamma_3}$ and $\widehat{\gamma} = \sqrt{\widehat{\gamma}_2\widehat{\gamma}_3}$. For any $0 < \varepsilon < 1$, we have

$$\begin{aligned}
\Pr\{|1/\widehat{\gamma} - 1/\gamma| > \varepsilon\} =& \Pr(|\widehat{\gamma} - \gamma| > |\widehat{\gamma}\gamma|\varepsilon) \\
\leq& \Pr\{|\widehat{\gamma} - \gamma| > |\widehat{\gamma}\gamma|\varepsilon, |\widehat{\gamma}| \geq \gamma/2\} + \Pr\{|\widehat{\gamma}| < \gamma/2\} \\
\leq& \Pr\{|\widehat{\gamma} - \gamma| > \gamma^2\varepsilon/2\} + \Pr\{|\widehat{\gamma} - \gamma| > \gamma/2\} \\
\leq& 2\Pr\{|\widehat{\gamma} - \gamma| > \min\{\gamma, \gamma^2\}\varepsilon/2\}.
\end{aligned}$$

From (5.8), we know

$$\Pr\{|1/\widehat{\gamma} - 1/\gamma| > \varepsilon\} \leq 4c_1 \exp\{-c_2 n\varepsilon^2\gamma'/16M^2\},$$

where $\gamma' = \min\{\gamma_2^2\gamma_3^2, \gamma_2^3\gamma_3^3\}$. As a result,

$$\begin{aligned}
\Pr\left\{\left|\frac{\widehat{\gamma}_1}{\sqrt{\widehat{\gamma}_2\widehat{\gamma}_3}} - \frac{\gamma_1}{\sqrt{\gamma_2\gamma_3}}\right| > \varepsilon\right\} =& \Pr\{|\widehat{\gamma}_1/\widehat{\gamma} - \gamma_1/\gamma| > \varepsilon\} \\
\leq& \Pr\{|\widehat{\gamma}_1/\widehat{\gamma} - \widehat{\gamma}_1/\gamma| > \varepsilon/2\} + \Pr(|\widehat{\gamma}_1/\gamma - \gamma_1/\gamma| > \varepsilon/2\} \\
\leq& \Pr\{|1/\widehat{\gamma} - 1/\gamma| > \varepsilon/M\} + \Pr\{|\widehat{\gamma}_1 - \gamma_1| > \varepsilon\gamma/2\} \\
\leq& 4c_1 \exp\{-c_2 n\varepsilon^2\gamma'/4M^4\} + c_1 \exp\{-c_2 n\varepsilon^2\gamma_2\gamma_3/4\} \\
\leq& 5c_1 \exp\{-c_2 n\varepsilon^2\gamma_0\},
\end{aligned}$$

where $\gamma_0 = \min\{\gamma_2^2\gamma_3^2/4M^4, \gamma_2^3\gamma_3^3/4M^4, \gamma_2\gamma_3/4\}$. □

Proof of Theorem 5.1

Let $\mathbf{z}_i = (\mathbf{z}_i, \mathbf{y}_i), i = 1, \ldots, n$ and define

$$
\begin{aligned}
h_1(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) &= \left[\arccos\left\{\frac{(\mathbf{x}_i - \mathbf{x}_k)^{\mathrm{T}}(\mathbf{x}_j - \mathbf{x}_k)}{\|\mathbf{x}_i - \mathbf{x}_k\|\|\mathbf{x}_j - \mathbf{x}_k\|}\right\}\arccos\left\{\frac{(\mathbf{y}_i - \mathbf{y}_k)^{\mathrm{T}}(\mathbf{y}_j - \mathbf{y}_k)}{\|\mathbf{y}_i - \mathbf{y}_k\|\|\mathbf{y}_j - \mathbf{y}_k\|}\right\}\right], \\
h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r) &= \left[\arccos\left\{\frac{(\mathbf{x}_i - \mathbf{x}_k)^{\mathrm{T}}(\mathbf{x}_l - \mathbf{x}_k)}{\|\mathbf{x}_i - \mathbf{x}_k\|\|\mathbf{x}_l - \mathbf{x}_k\|}\right\}\arccos\left\{\frac{(\mathbf{y}_j - \mathbf{y}_k)^{\mathrm{T}}(\mathbf{y}_r - \mathbf{y}_k)}{\|\mathbf{y}_j - \mathbf{y}_k\|\|\mathbf{y}_r - \mathbf{y}_k\|}\right\}\right], \\
h_3(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) &= \left[\arccos\left\{\frac{(\mathbf{x}_i - \mathbf{x}_k)^{\mathrm{T}}(\mathbf{x}_l - \mathbf{x}_k)}{\|\mathbf{x}_i - \mathbf{x}_k\|\|\mathbf{x}_l - \mathbf{x}_k\|}\right\}\arccos\left\{\frac{(\mathbf{y}_j - \mathbf{y}_k)^{\mathrm{T}}(\mathbf{y}_l - \mathbf{y}_k)}{\|\mathbf{y}_j - \mathbf{y}_k\|\|\mathbf{y}_l - \mathbf{y}_k\|}\right\}\right].
\end{aligned}
$$

Recall that the squared population projection covariance and correlation are defined as

$$
\begin{aligned}
\mathrm{Pcov}(\mathbf{x}, \mathbf{y})^2 &= S_1 + S_2 - 2S_3, \\
\mathrm{PC}(\mathbf{x}, \mathbf{y})^2 &= \mathrm{Pcov}(\mathbf{x}, \mathbf{y})^2 / \mathrm{Pcov}(\mathbf{x}, \mathbf{x})\mathrm{Pcov}(\mathbf{y}, \mathbf{y}),
\end{aligned}
$$

where S_1, S_2, S_3 are defined in (5.3). Their sample counterparts are given by

$$
\begin{aligned}
\widehat{\mathrm{Pcov}}(\mathbf{X}, \mathbf{Y})^2 &= \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3, \\
\widehat{\mathrm{PC}}(\mathbf{X}, \mathbf{Y})^2 &= \widehat{\mathrm{Pcov}}(\mathbf{X}, \mathbf{Y})^2 / \widehat{\mathrm{Pcov}}(\mathbf{X}, \mathbf{X})\widehat{\mathrm{Pcov}}(\mathbf{Y}, \mathbf{Y}),
\end{aligned}
$$

where $\widehat{S}_1, \widehat{S}_2, \widehat{S}_3$ are defined as follows,

$$
\begin{aligned}
\widehat{S}_1 =& n^{-3} \sum_{i,j,k=1}^{n} h_1(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k), \\
\widehat{S}_2 =& n^{-5} \sum_{i,j,k,l,r=1}^{n} h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r), \\
\widehat{S}_3 =& n^{-4} \sum_{i,j,k,l=1}^{n} h_3(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l).
\end{aligned}
$$

Clearly $\widehat{S}_1, \widehat{S}_2$ and $\widehat{S}_3$ are V-statistics. Denote by $\pi(i_1, \ldots, i_m)$ the set of all

permutations of $(i_1, \ldots, i_m)$. Define

$$\widehat{S}_1^* = \binom{n}{3}^{-1} \sum_{i<j<k} h_1^*(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k),$$
$$\widehat{S}_2^* = \binom{n}{5}^{-1} \sum_{i<j<k<l<r} h_2^*(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r),$$
$$\widehat{S}_3^* = \binom{n}{4}^{-1} \sum_{i<j<k<l} h_3^*(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l),$$

where

$$h_1^*(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) = \frac{1}{3!} \sum_{\pi(i,j,k)} h_1(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k),$$
$$h_2^*(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r) = \frac{1}{5!} \sum_{\pi(i,j,k,l,r)} h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r),$$
$$h_3^*(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) = \frac{1}{4!} \sum_{\pi(i,j,k,l)} h_3(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l).$$

By definition, we know $\widehat{S}_1^*, \widehat{S}_2^*, \widehat{S}_3^*$ are the corresponding U-statistics with associated kernels h_1^*, h_2^* and h_3^*, respectively. We have $0 \le S_k, \widehat{S}_k, \widehat{S}_k^* \le \pi^2$ and $0 \le h_k^* \le \pi^2$, $k = 1, 2, 3$.

We first deal with $\widehat{S}_1$. Note that

$$\sum_{i,j,k} h_1(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) = \sum_{i \ne j \ne k} h_1(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k),$$

which implies $n^3 \widehat{S}_1 = n(n-1)(n-2)\widehat{S}_1^*$. For any given $\varepsilon > 0$, take $n \ge 3\pi^2/\varepsilon$ such that $3S_1/n \le \varepsilon$, we have

$$\begin{aligned}
&\Pr(|\widehat{S}_1 - S_1| \ge 2\varepsilon) \\
=&\Pr\{|\widehat{S}_1^*(n-1)(n-2)/n^2 - S_1(n-1)(n-2)/n^2 - S_1(3n-2)/n^2| \ge 2\varepsilon\} \\
\le&\Pr\{|\widehat{S}_1^* - S_1|(n-1)(n-2)/n^2 \ge 2\varepsilon - S_1(3n-2)/n^2\} \\
\le&\Pr\{|\widehat{S}_1^* - S_1| \ge \varepsilon\}.
\end{aligned}$$

Since $0 \le h_1^* \le \pi^2$, applying Lemma 5.2, we have

$$\Pr(|\widehat{S}_1 - S_1| \ge 2\varepsilon) \le 2\exp\{-2\lfloor n/3 \rfloor \varepsilon^2/\pi^4\}. \tag{5.9}$$

Now we move to the third term $\widehat{S}_3$. Note that

$$\sum_{i,j,k,l} h_3(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) = \sum_{i\neq j\neq k\neq l} h_3(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l) + \sum_{i=j\neq k\neq l} h_3(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l).$$

Thus

$$\begin{aligned} n^4\widehat{S}_3 =& n(n-1)(n-2)(n-3)\widehat{S}_3^* + n(n-1)(n-2)\widehat{S}_1^*, \\ \widehat{S}_3 =& \widehat{S}_3^*(n-1)(n-2)(n-3)/n^3 + \widehat{S}_1^*(n-1)(n-2)/n^3. \end{aligned}$$

Take $n \geq 6\pi^2/\varepsilon$, then $\widehat{S}_1^*(n-1)(n-2)/n^3 \leq \varepsilon$ and $S_3(6n^2-11n+6)/n^3\} \leq \varepsilon$. We have

$$\begin{aligned} &\Pr(|\widehat{S}_3 - S_3| \geq 3\varepsilon) \\ =&\Pr\{|\widehat{S}_3^*(n-1)(n-2)(n-3)/n^3 + \widehat{S}_1^*(n-1)(n-2)/n^3 - S_3| \geq 3\varepsilon\} \\ \leq&\Pr\{|\widehat{S}_3^*(n-1)(n-2)(n-3)/n^3 - S_3| \geq 2\varepsilon\} \\ \leq&\Pr\{|\widehat{S}_3^* - S_3|(n-1)(n-2)(n-3)/n^3 \geq 2\varepsilon - S_3(6n^2-11n+6)/n^3\} \\ \leq&\Pr\{|\widehat{S}_3^* - S_3| \geq \varepsilon\}. \end{aligned}$$

Since $0 \leq h_3^* \leq \pi^2$, applying Lemma 5.2 again, we have

$$\Pr(|\widehat{S}_3 - S_3| \geq 3\varepsilon) \leq 2\exp\{-2[n/4]\varepsilon^2/\pi^4\}. \tag{5.10}$$

It remains to deal with the second term $\widehat{S}_2$. Note that

$$\begin{aligned} \sum_{i,j,k,l,r} h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r) = & \sum_{i\neq j\neq l\neq r\neq k} h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r) + \\ & \sum_{i=j\neq l\neq r\neq k} h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r) + \\ & \sum_{i=r\neq j\neq l\neq k} h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r) + \\ & \sum_{l=j\neq i\neq r\neq k} h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r) + \\ & \sum_{l=r\neq i\neq j\neq k} h_2(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k, \mathbf{z}_l, \mathbf{z}_r). \end{aligned}$$

Thus

$$\begin{aligned} n^5\widehat{S}_2 =& n(n-1)(n-2)(n-3)(n-4)\widehat{S}_2^* + 4n(n-1)(n-2)(n-3)\widehat{S}_3^*, \\ \widehat{S}_2 =& (n-1)(n-2)(n-3)(n-4)/n^4\widehat{S}_2^* + 4(n-1)(n-2)(n-3)\widehat{S}_3^*/n^4. \end{aligned}$$

Take $n \geq 10\pi^2/\varepsilon$, we have

$$4(n-1)(n-2)(n-3)\widehat{S}_3^*/n^4 \leq \varepsilon \text{ and}$$

$$(1-(n-1)(n-2)(n-3)(n-4)/n^4)S_3 \leq \varepsilon.$$

As a result,

$$\begin{aligned} &\Pr(|\widehat{S}_2 - S_2| \geq 3\varepsilon) \\ =&\Pr\{|\widehat{S}_2^*(n-1)(n-2)(n-3)(n-4)/n^4 + 4(n-1)(n-2)(n-3)\widehat{S}_3^*/n^4 - S_2| \geq 3\varepsilon\} \\ \leq&\Pr\{|\widehat{S}_2^*(n-1)(n-2)(n-3)(n-4)/n^4 - S_2| \geq 2\varepsilon\} \\ \leq&\Pr\{|\widehat{S}_3^* - S_3|(n-1)(n-2)(n-3)(n-4)/n^4 \geq \varepsilon\} \\ \leq&\Pr\{|\widehat{S}_3^* - S_3| \geq \varepsilon\}. \end{aligned}$$

Since $0 \leq h_2^* \leq \pi^2$, by Lemma 5.2, we have

$$\Pr(|\widehat{S}_2 - S_2| \geq 3\varepsilon) \leq 2\exp\{-2\lfloor n/5\rfloor\varepsilon^2/\pi^4\}. \tag{5.11}$$

Combining (5.9), (5.10) and (5.11), we have

$$\begin{aligned} &\Pr\{|(\widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3) - (S_1 + S_2 - 2S_3)| \geq 11\varepsilon\} \\ \leq&\Pr(|\widehat{S}_1 - S_1| \geq 2\varepsilon) + \Pr(|\widehat{S}_2 - S_2| \geq 3\varepsilon) + \Pr(|\widehat{S}_3 - S_3| \geq 3\varepsilon) \\ \leq&6\exp\{-2\lfloor n/5\rfloor\varepsilon^2/\pi^4\}. \end{aligned}$$

Therefore

$$\Pr\{|(\widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3) - (S_1 + S_2 - 2S_3)| \geq \varepsilon\} \leq c_1\exp\{-c_2 n\varepsilon^2\}, \tag{5.12}$$

with choices $c_1 = 6$ and $c_2 = 1/(320\pi^4)$. For the second part, we apply Lemma 5.3

to (5.12) with the choice $M = (2\pi)^2$, we obtain

$$\Pr\{|\widehat{\mathrm{PC}}(\mathbf{X}, \mathbf{Y})^2 - \mathrm{PC}(\mathbf{x}, \mathbf{y})^2| \geq \varepsilon\} \leq 5c_1 \exp\{-c_2 n\varepsilon^2 \gamma\}.$$

where $\gamma = \min\{\gamma_x^3\gamma_y^3/4M^4, \gamma_x^2\gamma_y^2/4M^4, \gamma_x\gamma_y/4\}$.

Proof of Theorem 5.2

Recall that

$$\omega_k = \mathrm{PC}(X_k, \mathbf{y}) \text{ and } \widehat{\omega}_k = \mathrm{PC}(\mathbf{X}_k, \mathbf{Y}),$$

where $\mathbf{X}_k$ is the kth column of $\mathbf{X}$. By Theorem 5.1 and Condition 1 (a), we have

$$\Pr(|\widehat{\omega}_k - \omega_k| \geq \delta) \leq O(\exp\{-c_4 n\delta^2\}),$$

where c_4 is some constant. We consider the compliment $\mathcal{A} \nsubseteq \widehat{\mathcal{A}}(\delta)$, meaning that there is at least one $k \in \mathcal{A}$ such that $k \notin \widehat{\mathcal{A}}(\delta)$. We have

$$\begin{aligned}
\Pr(\mathcal{A} \nsubseteq \widehat{\mathcal{A}}(\delta)) &= \Pr(\cup_{k\in\mathcal{A}}\{k \notin \widehat{\mathcal{A}}(\delta)\}) \\
&\leq \sum_{k\in\mathcal{A}} \Pr(\widehat{\omega}_k \leq \delta) \\
&\leq \sum_{k\in\mathcal{A}} \Pr(|\widehat{\omega}_k - \omega_k| \geq \delta) \\
&\leq O(s\exp\{-c_4 n\delta^2\}).
\end{aligned}$$

Thus $\Pr(\mathcal{A} \subseteq \widehat{\mathcal{A}}(\delta)) \geq 1 - O(s\exp\{-c_4 n\delta^2\})$.

Proof of Theorem 5.3

Let $k_1 = \arg\min_{k\in\mathcal{A}} \widehat{\omega}_k$ and $k_2 = \arg\max_{k\in\mathcal{A}^c} \widehat{\omega}_k$. For any $0 \leq \kappa < 1/2$, we have

$$
\begin{aligned}
&\Pr\{\min_{k\in\mathcal{A}} \widehat{\omega}_k - \max_{k\in\mathcal{A}^c} \widehat{\omega}_k \le 0\} \\
\le&\Pr\{\min_{k\in\mathcal{A}} \widehat{\omega}_k - \max_{k\in\mathcal{A}^c} \widehat{\omega}_k \le \min_{k\in\mathcal{A}} \omega_k - \max_{k\in\mathcal{A}^c} \omega_k - 2c_3 n^{-\kappa}\} \\
=&\Pr\{(\min_{k\in\mathcal{A}} \omega_k - \min_{k\in\mathcal{A}} \widehat{\omega}_k) + (\max_{k\in\mathcal{A}^c} \widehat{\omega}_k - \min_{k\in\mathcal{A}^c} \omega_k) \ge 2c_3 n^{-\kappa}\} \\
\le&\Pr\{(\omega_{k_1} - \widehat{\omega}_{k_1}) + (\widehat{\omega}_{k_2} - \omega_{k_2}) \ge 2c_3 n^{-\kappa}\} \\
\le&\Pr\{|\widehat{\omega}_{k_1} - \omega_{k_1}| \ge c_3 n^{-\kappa}\} + \Pr\{|\widehat{\omega}_{k_2} - \omega_{k_2}| \ge c_3 n^{-\kappa}\} \\
\le&2\Pr\{\max_{1\le k\le p} |\widehat{\omega}_k - \omega_k| \ge c_3 n^{-\kappa}\} \\
=&c_5' p \exp\{-c_5 n^{1-2\kappa}\},
\end{aligned}
$$

where $c_5, c_5' > 0$ are positive constants. The first inequality follows Condition 1 (b) and the last equality is implied by Theorem 5.1. Hence we have

$$
\Pr\{\min_{k\in\mathcal{A}} \widehat{\omega}_k - \max_{k\in\mathcal{A}^c} \widehat{\omega}_k > 0\} \ge 1 - O(p\exp\{-c_5 n^{1-2\kappa}\}).
$$

If we further assume $\log p = o(n^{1-2\kappa})$, we know $p < \exp\{c_5 n^{1-2\kappa}/2\}$ for large n. Then we have

$$
p\exp\{-c_5 n^{1-2\kappa}\} \le \exp\{-c_5 n^{1-2\kappa}/2\} \le \exp\{-2\log n\} \le n^{-2},
$$

for large n. Thus for some n_0, we have

$$
\sum_{n=n_0}^{\infty} c_5' p \exp\{-c_5 n\varepsilon^2\} \le c_5' \sum_{n=n_0}^{\infty} n^{-2} \le \infty.
$$

Therefore, by Borel-Contelli Lemma, we obtain

$$
\Pr(\limsup_{n\to\infty}\{\min_{k\in\mathcal{A}} \widehat{\omega}_k - \max_{k\in\mathcal{I}} \widehat{\omega}_k \le 0\}) = 0,
$$

As a result

$$
\begin{aligned}
&\Pr(\liminf_{n\to\infty}\{\min_{k\in\mathcal{A}} \widehat{\omega}_k - \max_{k\in\mathcal{I}} \widehat{\omega}_k > 0\}) \\
=&\Pr([\limsup_{n\to\infty}\{\min_{k\in\mathcal{A}} \widehat{\omega}_k - \max_{k\in\mathcal{I}} \widehat{\omega}_k \le 0\}]^c) \\
=&1.
\end{aligned}
$$

Chapter 6
Summary and Future Work

6.1 Summary of the dissertation

In this dissertation, we study two fundamental problems for high-dimensional statistics: mean vector test of high-dimensional data and feature screening for ultra-high dimensional data. In the first part, consisting of chapter 3 and chapter 4, we focus on the estimation of linear functional and its application to projection test for high-dimensional mean vector. In chapter 3, we study the regularized quadratic programming with nonconvex penalty and linear constraint. Under the assumption that the quadratic form satisfies the restrict strong convexity (RSC) condition, we establish the L_1 and L_2 deterministic error bounds for any stationary point that satisfies the necessary first order condition to be a local minimum. Further assuming the strict dual feasibility, we show that the stationary point is unique and establish the support recovery and L_∞ error bound. We propose an ADMM algorithm with local linear approximation to solve the nonconvex quadratic programming, which is guaranteed to converge to a local minimum. In particular, we study three applications of the regularized quadratic programming: (1) estimation of linear functional, (2) F-type test for regression coefficients and (3) sparse linear discriminant analysis. Convergence rates in terms of L_1 and L_2 norms are established for the proposed estimator. In chapter 4, we apply the linear functional, also known as the optimal projection direction, to perform projection test for high-dimensional data assuming that the linear functional is sparse. We propose two sparse projection tests. The first test is the sparse projection test with data splitting. The entire dataset is partitioned into two sets. We use the first

set to estimate the optimal projection direction and perform the test only using the data in the second set. This data splitting projection test achieves an exact t-test under normality assumption. The second test is the sparse online projection test, which updates the estimation of optimal projection direction whenever a new observation arrives. We derive the asymptotic normal distribution of the test statistic under both null hypothesis and alternative hypothesis. This sparse online projection test improves the power of data splitting projection test. A mini-batch version of the online projection test is also proposed. This test updates the estimation when a batch of new observations arrive and thus reduces the computational burden a lot. In addition, we conduct numerical studies to compare the finite sample performance of the proposed projection tests with several existing tests. The numerical results show that the proposed projection tests can control the type I error rate well and are much more powerful than other existing tests. The second part of this dissertation, consisting of chapter 5, focuses on the model free feature screening for ultra-high dimensional data. The proposed method is based on ranking the projection correlations between features and response. This projection correlation based screening procedure is model free in the sense that it does not require specifying a regression model and can be applied to grouped variables as well. This procedure is also insensitive to outliers since no moment conditions are needed for the data. Theoretical analysis demonstrates the proposed method enjoys not only the sure screening property but also a stronger result called rank consistency property.

6.2 Future work

There are a lot of interesting topics we can continue to work on. In chapter 3, we consider the regularized quadratic programming with linear constraint $\mathbf{C}\boldsymbol{\beta} \leq \mathbf{b}$ and achieve the following convergence rates (error bounds) under proper assumptions,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 = O(s\lambda/\kappa), \text{ and } \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O(\sqrt{s}\lambda/\kappa).$$

These convergence rates coincide with the results in Negahban et al. (2012) and Loh and Wainwright (2015) where no such linear constraint is imposed. These convergence rates are known to be optimal when no constraint exists (Raskutti

et al. 2011, Negahban et al. 2012, Wang et al. 2014). Are they also optimal when we have additional constraint? Suppose the rank of matrix $\mathbf{C}$ is m and assume $m < s$, one interesting question is can we improve the convergence rate by taking advantage of the linear constraint. For example, we may expect to achieve the following convergence rate,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 = O((s-m)\lambda/\kappa), \text{ and } \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O(\sqrt{s-m}\lambda/\kappa).$$

In chapter 4, we propose the online projection test for high-dimensional mean vector. Projection test under the online framework can improve the performance of projection test using the data splitting technique. On one hand, the online projection test keeps updating the estimated projection direction when new observations arrive, thus we can obtain a more and more accurate estimation of the optimal projection direction. On the other hand, only the first $k_n = n^{-\tau} = o(n)$ observations are discarded when performing the test while the data splitting procedure discards about half of the data. Hence the online projection test is more powerful than the data splitting procedure. Though discarding the first k_n data points does not affect the asymptotic results, of interest is to develop a new projection test without throwing away any data points. One possible approach is to pretend the discarded k_n observations come from the future and reuse them. Another approach is to use the entire dataset to obtain an estimate of the projection direction and then perform the projection test with the entire dataset. In both cases, we expect to see some improvement on the test power since no observation is discarded. The challenge is it can be very completed to derive the limiting null distribution since the estimated projection direction and the data are no longer independent.

In chapter 5, we propose a feature screening procedures based on projection correlation. Let $\widehat{\omega}_j$ be the sample projection correlation between the response and the jth feature, the selected features are given by

$$\widehat{\mathcal{M}} = \{j : \widehat{\omega}_j \geq \delta, 1 \leq j \leq p\},$$

or equivalently,

$$\widehat{\mathcal{M}} = \{j : \widehat{\omega}_j \text{ are among the top } d \text{ ones}, 1 \leq j \leq p\}.$$

We show that with a proper choice of δ or d, our feature screening procedure enjoys the sure screening property, that is, all important features are selected with probability approaching to 1. However, the proper choice of threshold δ or d depends on unknown parameters. From a practical point of view, one may be conservative and use a relatively large value of d such that all important features are included. In that case, many unimportant features may enter the selected model as well. Zhu et al. (2011) proposed a threshold rule by introducing a series of artificial auxiliary variables. Based on our simulation study, this kind of threshold rule is also conservative and allows a lot of unimportant features entering the selected model. We are also interested in developing some data-driven method to find a proper threshold δ or d such that not only the sure screening property is satisfied but the number of unimportant features that are selected is well controlled.

Appendix A
Properties of Nonconvex Regularizers

We state several properties of the nonconvex penalty function $P_\lambda(t)$.

Lemma A.1. *Assume penalty function $P_\lambda(t)$ satisfies conditions (i)-(v) in Section 3.2.1, then*

(1) $|P_\lambda(t_1) - P_\lambda(t_2)| \leq \lambda|t_1 - t_2|$ *for any* $t_1, t_2 \in \mathbb{R}$.

(2) For any $\boldsymbol{\beta} \in \mathbb{R}^p$*, we have* $\lambda\|\boldsymbol{\beta}\|_1 \leq P_\lambda(\boldsymbol{\beta}) + \frac{\mu}{2}\|\boldsymbol{\beta}\|_2^2$.

(3) Suppose $\|\boldsymbol{\beta}^\star\|_0 = s$*, then for any* $\boldsymbol{\beta} \in \mathbb{R}^p$ *such that* $cP_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\boldsymbol{\beta}) \geq 0$ *with* $c \geq 1$*, we have* $cP_\lambda(\boldsymbol{\beta}^\star) - P_\lambda(\boldsymbol{\beta}) \leq \lambda(c\|\boldsymbol{\delta}_{\mathcal{A}}\|_1 - \|\boldsymbol{\delta}_{\mathcal{A}^c}\|_1)$*, where* $\boldsymbol{\delta} = \boldsymbol{\beta} - \boldsymbol{\beta}^\star$ *and* $\mathcal{A}$ *is the index set of the* s *largest elements of* $\boldsymbol{\delta}$ *in magnitude.*

Proof. The proof can be found in Loh and Wainwright (2015). □

Lemma A.2. *The function* $J_\lambda(t) - \frac{\mu}{2}t^2 = \lambda|t| - P_\lambda(t) - \frac{\mu}{2}t^2$ *is concave and differentiable.*

Proof. When $t > 0$, $J_\lambda(t) - \frac{\mu}{2}t^2 = \lambda t - P_\lambda(t) - \frac{\mu}{2}t^2$. By assumption, we know $P_\lambda(t) + \frac{\mu}{2}t^2$ is convex and thus $J_\lambda(t) - \frac{\mu}{2}t^2$ is concave when $t > 0$. Similarly, $J_\lambda(t) - \frac{\mu}{2}t^2$ is concave when $t < 0$. As a result, the derivative of $J_\lambda(t) - \frac{\mu}{2}t^2$ is decreasing on $(-\infty, 0)$ and $(0, \infty)$. At $t = 0$, we have $J'(0) = 0$. Therefore, $J_\lambda(t) - \frac{\mu}{2}t^2$ is differential and its derivative is monotonically decreasing. Hence $J_\lambda(t) - \frac{\mu}{2}t^2$ is a concave function. □

Appendix B
Sub-Gaussian Random Variable

The following lemmas show some nice properties of sub-Gaussian random variable.

Lemma B.1. *Let $X_1, X_2, \ldots, X_n$ be sub-Gaussian random variables with variance proxy σ^2, then*

$$E\left(\max_{1\le i\le n} X_i\right) \le \sigma\sqrt{2\log n} \text{ and } E\left(\max_{1\le i\le n} |X_i|\right) \le \sigma\sqrt{2\log(2n)}.$$

Proof. Let $Z = \max_{1\le i\le n} X_i$. By Jensen's inequality, for any $t > 0$

$$\exp\{t\mathrm{E}[Z]\} \le \mathrm{E}\exp\{tZ\} \le \sum_{i=1}^{n} \mathrm{E}\exp\{tX_i\} \le n\exp\{t^2\sigma^2/2\}.$$

Taking the logarithm of both sides, we get

$$\mathrm{E}[Z] \le \frac{\log n}{t} + \frac{t\sigma^2}{2}.$$

Taking $t = \sqrt{2\log n}/\sigma$ yields the first inequality. The second inequality follows trivially by noting that

$$\max_{1\le i\le n} |X_i| = \max_{1\le i\le 2n} X_i,$$

with $X_{n+i} = -X_i, i = 1, \ldots, n$. □

Lemma B.2. *Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent and identically distributed sub-Gaussian random vectors with $\|\mathbf{x}_i\|_{\psi_2} = K$. Let $\widehat{\mathbf{\Sigma}} = (\widehat{\sigma}_{ij})_{p\times p}$ be the sample*

covariance matrix. If $\log p < n$*, then there exists a constant* M *only depending on* K *such that* $\max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}| \leq M\sqrt{\log p/n}$ *with probability at least* $1 - 2p^{-1}$*.*

Proof. Let $\mathbf{x} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_i$ be the sample mean and $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^\top$ be the sample covariance matrix. The $\widehat{\mathbf{\Sigma}}$ can be decomposed as

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^\top - \bar{\mathbf{x}}\bar{\mathbf{x}}^\top.$$

Without loss of generality, we assume $\mathrm{E}(\mathbf{x}_i) = \mathbf{0}$. Therefore,

$$\max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}| \leq \max_{i,j} |\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_{ki}\mathbf{x}_{kj} - \sigma_{ij}| + \max_{i,j} |\bar{\mathbf{x}}_i\bar{\mathbf{x}}_j|.$$

By the property of sub-Gaussian random variables, we know

$$\|\mathbf{x}_{ki}\mathbf{x}_{kj}\|_{\psi_1} \leq 2\|\mathbf{x}_{ki}\|_{\psi_2}\|\mathbf{x}_{kj}\|_{\psi_2} \leq 2K^2,$$

and hence $\|\mathbf{x}_{ki}\mathbf{x}_{kj} - \sigma_{ij}\|_{\psi_1} \leq 4K^2$. As a result,

$$\mathrm{P}\left(\max_{i,j}\left|\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_{ki}\mathbf{x}_{kj} - \sigma_{ij}\right| > t\right) \leq \max\left(2p^2\exp\left\{-c_1 n\frac{t^2}{16K^4}\right\}, 2p^2\exp\left\{-c_1 n\frac{t}{4K^2}\right\}\right).$$

It is easy to verify that $\|\bar{\mathbf{x}}_i\|_{\psi_2} \leq \sqrt{c/n}K$ and $\|\bar{\mathbf{x}}_i\bar{\mathbf{x}}_j\|_{\psi_1} \leq 2\|\bar{\mathbf{x}}_i\|_{\psi_2}\|\bar{\mathbf{x}}_j\|_{\psi_2} \leq 2cK^2/n$, we have

$$\mathrm{P}(\max_{i,j} |\bar{\mathbf{x}}_i\bar{\mathbf{x}}_j| > t) \leq 2p^2\exp\left\{-\frac{nt}{2cK^2}\right\}.$$

By the choice of $t = \frac{M}{2}\sqrt{\frac{\log p}{n}}$, we have $\max_{i,j} |\widehat{\sigma}_{ij} - \sigma_{ij}| \leq M\sqrt{\log p/n}$ with probability at least $1 - 2p^{-1}$ whenever $M > C$, where C is a constant only depending on K.

□

Appendix C
Error Bounds Under the RE Condition

Instead of imposing the restricted strong convexity (RSC) condition, here we state some results on the error bounds and strong oracle property for the regularized quadratic programming (3.7) under the restricted eigenvalue (RE) condition. We say a matrix $\mathbf{W}$ satisfies the RE(q, κ) condition with $r > 1$ if

$$\min_{\boldsymbol{\beta} \neq \mathbf{0}, \|\boldsymbol{\beta}_{\mathcal{A}}\|_1 \leq q, \|\boldsymbol{\beta}_{\mathcal{A}^c}\|_1} \frac{\boldsymbol{\beta}^\top \mathbf{W} \boldsymbol{\beta}}{\|\boldsymbol{\beta}\|_2^2} = \kappa > 0, \tag{C.1}$$

where $\mathcal{A}$ is the support of true parameter $\boldsymbol{\beta}^\star$. By the definition, we know that if $q' > q$, then RE(q', κ) implies RE(q, κ). This type of restricted eigenvalue condition was proposed in Bickel et al. (2009) when they studied the connection between the Lasso and the Dantzig selector. Wang et al. (2013) also adopted this type of condition to study the SCAD-CCCP estimator under a high-dimensional regression setting.

It is natural to ask under what conditions a matrix $\mathbf{W}$ would satisfy the RE condition. In many applications, $\mathbf{W}$ is of the form of $\mathbf{W} = \mathbf{X}^\top \mathbf{X}/n$ where $\mathbf{X}$ is a $n \times p$ design matrix. If $\mathbf{X}$ has i.i.d. entries following a sub-Gaussian distribution, then known results from random matrix theory implies that this RE condition holds with a high probability. In statistical applications, we are more interested in design matrix $\mathbf{X}$ with substantial dependency. Let $\mathbf{x}_i$ denote the i-th row of $\mathbf{X}$ and if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d. random sub-Gaussian vectors, Rudelson and Zhou (2012) showed that the RE condition holds with high probability when the sample size n is larger than some constant depending on dimension p and sparsity level s. The RE

condition is known to be a relatively mild condition for high-dimensional estimation and is much weaker than the irrepresentable condition (Zhao and Yu 2006) which is a necessary condition for Lasso estimator to be model selection consistent.

We consider the penalty function $p_\lambda(t)$ from the family $\mathcal{P}_\lambda(a_0, a_1, a_2, a_3)$ which is a set of concave folded penalty functions and each element $p_\lambda(t)$ in the set satisfies the following conditions:

(1) $p_\lambda(t)$ is symmetric on the real line and is non-decreasing and concave in $t \in [0, \infty)$ with $p_\lambda(0) = 0$.

(2) $p_\lambda(t)$ is differentiable in $t \in (0, \infty)$ with $p'_\lambda(0+) = a_0\lambda$ for some $a_0 > 0$.

(3) $p'_\lambda(t) \geq a_1\lambda$ in $t \in (0, a_2\lambda]$ for some $a_1 \leq a_0$.

(4) $p'_\lambda(t) = 0$ if $t \in [a_3\lambda, \infty)$ for $a_3 > a_2$.

This class of folded concave penalty functions $\mathcal{P}_\lambda(a_0, a_1, a_2, a_3)$ includes the two most widely used nonconvex penalty functions: the SCAD (Fan and Li 2001) and the MCP (Zhang 2010). In particular, the SCAD belongs to $\mathcal{P}_\lambda(1, 1, 1, a_3)$ with $a_3 > 2$ and the MCP belongs to $\mathcal{P}_\lambda(1, 1/2, a_3/2, a_3)$ with some $a_3 > 0$.

To establish the deterministic error bound for our proposed LLA estimator, we specify a set of regularity conditions. Recall that a_0, a_1, a_2, a_3 are the parameters in the folded concave penalty.

(C.A.1) The true parameter $\boldsymbol{\beta}^\star$ satisfies the linear equality constraint $\mathbf{C}\boldsymbol{\beta}^\star = \mathbf{b}$.

(C.A.2) $\tau < \min\{1, \kappa a_2/(3a_0\sqrt{s_0})\}$.

(C.A.3) There exists a vector $\mathbf{w}$ such that $2\|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{C}^\top\mathbf{w} - \mathbf{q}\|_\infty \leq \tau\lambda a_1$.

Theorem C.1 (Deterministic error bounds)**.** *Let $\widehat{\boldsymbol{\beta}}$ be the LLA estimator to (3.7) with $p_\lambda(\cdot) \in \mathcal{P}(a_0, a_1, a_2, a_3)$. Assume that $\mathbf{W}$ satisfies RE$(\max\{1 + 2a_0/a_1, 3\}, \kappa)$ and conditions (C.A.1) - (C.A.3) hold, then we have*

(1) $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \leq a\lambda s_0/\kappa$, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \leq (2a_0 + a_1)\lambda\sqrt{s_0}/\kappa$,

(2) $\sqrt{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)^\top\mathbf{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)} \leq (2a_0 + a_1)\lambda\sqrt{s_0/\kappa}$,

where $a = 6a_0 + 2a_1 + 4a_0^2/a_1$.

Proof. Let $\widehat{\boldsymbol{\beta}}$ be some estimator of $\boldsymbol{\beta}^\star$ and denote $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star$. Let

$$Q(\boldsymbol{\beta}|\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top \boldsymbol{\beta} + \boldsymbol{\lambda}^\top |\boldsymbol{\beta}|, \tag{C.2}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)^\top$ satisfying $\lambda_j \le a_0\lambda$ for all j and $\lambda_j \ge a_1\lambda$ for $j \in \mathcal{A}^c$.

We first claim that if λ is chosen such that $a_1\lambda \ge 2\|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{C}^\top\mathbf{w} - \mathbf{q}\|_\infty$ and $Q(\widehat{\boldsymbol{\beta}}|\boldsymbol{\lambda}) \le Q(\boldsymbol{\beta}^\star|\boldsymbol{\lambda})$, then we have $\|\boldsymbol{\Delta}\|_2 \le (2a_0 + a_1)\lambda\sqrt{s_0}/\kappa$ and $\|\boldsymbol{\Delta}\|_1 \le a\lambda s_0/\kappa$, where $a = 6a_0 + 2a_1 + 4a_0^2/a_1$, To see that, notice

$$\frac{1}{2}\widehat{\boldsymbol{\beta}}^\top \mathbf{W}\widehat{\boldsymbol{\beta}} - \mathbf{q}^\top\widehat{\boldsymbol{\beta}} + \boldsymbol{\lambda}^\top|\widehat{\boldsymbol{\beta}}| \le \frac{1}{2}\boldsymbol{\beta}^{\star\top}\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{q}^\top\boldsymbol{\beta}^\star + \boldsymbol{\lambda}^\top|\boldsymbol{\beta}^\star|.$$

Therefore

$$\frac{1}{2}\widehat{\boldsymbol{\beta}}^\top \mathbf{W}\widehat{\boldsymbol{\beta}} - \mathbf{w}^\top\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{q}^\top\widehat{\boldsymbol{\beta}} + \boldsymbol{\lambda}^\top|\widehat{\boldsymbol{\beta}}| \le \frac{1}{2}\boldsymbol{\beta}^{\star\top}\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{w}^\top\mathbf{C}\boldsymbol{\beta}^\star - \mathbf{q}^\top\boldsymbol{\beta}^\star + \boldsymbol{\lambda}^\top|\boldsymbol{\beta}^\star|,$$

since $\mathbf{C}\widehat{\boldsymbol{\beta}} = \mathbf{C}\boldsymbol{\beta} = \mathbf{b}$. By rearrangement, we have

$$\begin{aligned}
\frac{1}{2}\boldsymbol{\Delta}^\top\mathbf{W}\boldsymbol{\Delta} + \boldsymbol{\lambda}^\top|\widehat{\boldsymbol{\beta}}| &\le \boldsymbol{\Delta}^\top(\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{C}^\top\mathbf{w} - \mathbf{q}) + \boldsymbol{\lambda}^\top|\boldsymbol{\beta}^\star| \\
&\le \|\boldsymbol{\Delta}\|_1 \cdot \|\mathbf{W}\boldsymbol{\beta}^\star - \mathbf{C}^\top\mathbf{w} - \mathbf{q}\|_\infty + \boldsymbol{\lambda}^\top|\boldsymbol{\beta}^\star| \\
&\le \frac{a_1}{2}\lambda\|\boldsymbol{\Delta}\|_1 + \boldsymbol{\lambda}^\top|\boldsymbol{\beta}^\star|.
\end{aligned}$$

Now breaking $\boldsymbol{\beta}^\star$ and $\widehat{\boldsymbol{\beta}}$ into $\boldsymbol{\beta}^\star = (\boldsymbol{\beta}^\star_{\mathcal{A}}, \boldsymbol{\beta}^\star_{\mathcal{A}^c})$ and $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_{\mathcal{A}}, \widehat{\boldsymbol{\beta}}_{\mathcal{A}^c})$, we have

$$\begin{aligned}
\frac{1}{2}\boldsymbol{\Delta}^\top\mathbf{W}\boldsymbol{\Delta} + \boldsymbol{\lambda}_{\mathcal{A}}^\top|\widehat{\boldsymbol{\beta}}_{\mathcal{A}}| + \boldsymbol{\lambda}_{\mathcal{A}^c}^\top|\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c}| &\le \frac{a_1}{2}\lambda\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 + \frac{a_1}{2}\lambda\|\boldsymbol{\Delta}_{\mathcal{A}^c}\|_1 + \boldsymbol{\lambda}^\top|\boldsymbol{\beta}^\star| \\
\frac{1}{2}\boldsymbol{\Delta}^\top\mathbf{W}\boldsymbol{\Delta} + \boldsymbol{\lambda}_{\mathcal{A}}^\top|\boldsymbol{\beta}^\star_{\mathcal{A}}| - \boldsymbol{\lambda}_{\mathcal{A}}^\top|\boldsymbol{\Delta}_{\mathcal{A}}| + a_1\lambda\|\widehat{\boldsymbol{\beta}}_{\mathcal{A}^c}\|_1 &\le \frac{a_1}{2}\lambda\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 + \frac{a_1}{2}\lambda\|\boldsymbol{\Delta}_{\mathcal{A}^c}\|_1 + \boldsymbol{\lambda}^\top|\boldsymbol{\beta}^\star| \\
\frac{1}{2}\boldsymbol{\Delta}^\top\mathbf{W}\boldsymbol{\Delta} &\le (a_0 + \frac{a_1}{2})\lambda\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 - \frac{a_1}{2}\lambda\|\boldsymbol{\Delta}_{\mathcal{A}^c}\|_1.
\end{aligned}$$

Using the fact that $\frac{1}{2}\boldsymbol{\Delta}^\top\mathbf{W}\boldsymbol{\Delta} \ge 0$, we know $\|\boldsymbol{\Delta}_{\mathcal{A}^c}\|_1 \le (1 + 2a_0/a_1)\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1$. Under the RE condition of the theorem, we know $\kappa\|\boldsymbol{\Delta}\|_2^2 \le (2a_0 + a_1)\lambda\|\boldsymbol{\Delta}\|_1 \le (2a_0 + a_1)\lambda\sqrt{s_0}\|\boldsymbol{\Delta}\|_2$, hence $\|\boldsymbol{\Delta}\|_2 \le (2a_0 + a_1)\lambda\sqrt{s_0}/\kappa$. Therefore, $\|\boldsymbol{\Delta}\|_1 = \|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 + \|\boldsymbol{\Delta}_{\mathcal{A}^c}\|_1 \le (2 + 2a_0/a_1)\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 \le (2 + 2a_0/a_1)\sqrt{s_0}\|\boldsymbol{\Delta}_{\mathcal{A}}\|_2 \le a\lambda s_0/\kappa$, where $a = 6a_0 + 2a_1 + 4a_0^2/a_1$. We prove the claim.

By the definition of $\widehat{\boldsymbol{\beta}}^{(1)}$ and applying the claim with $\boldsymbol{\lambda}_1 = (\tau a_0\lambda, \ldots, \tau a_0\lambda)^\top$,

we have $\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^\star\|_2 \leq (2a_0 + a_1)\tau\lambda\sqrt{s_0}/\kappa \leq 3a_0\tau\lambda\sqrt{s_0}/\kappa$, which implies that $\max_{j\in\mathcal{A}^c} |\widehat{\beta}_j^{(1)}| \leq a_2\lambda$ since $\tau \leq \kappa a_2/(3a_0\sqrt{s_0})$. As a result, we have $p_\lambda(|\widehat{\beta}_j^{(1)}|) \leq a_0\lambda$ for all $j \in \mathcal{A}$ and $p_\lambda(|\widehat{\beta}_j^{(1)}|) \geq a_1\lambda$ for $j \in \mathcal{A}^c$. Let $\boldsymbol{\lambda}_2 = (p_\lambda(|\widehat{\beta}_1^{(1)}|), \ldots, p_\lambda(\widehat{\beta}_p^{(1)}))^\top$ and $\boldsymbol{\lambda}_2$ satisfies the condition of the claim. Let $\widehat{\boldsymbol{\beta}}$ be the LLA estimator. By the claim, we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 \leq (2a_0 + a_1)\lambda\sqrt{s_0}/\kappa$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_1 \leq a\lambda s_0/\kappa$, which completes part (1). For part (2), note that

$$\begin{aligned}
\boldsymbol{\Delta}^\top \mathbf{W}\boldsymbol{\Delta} &\leq (2a_0 + a_1)\lambda\|\boldsymbol{\Delta}_{\mathcal{A}}\|_1 \\
&\leq (2a_0 + a_1)\lambda\sqrt{s_0}\|\boldsymbol{\Delta}_{\mathcal{A}}\|_2 \\
&\leq (2a_0 + a_1)^2\lambda^2 s_0/\kappa.
\end{aligned}$$

Therefore $\sqrt{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)^\top \mathbf{W}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)} \leq (2a_0 + a_1)\lambda\sqrt{s_0/\kappa}$. □

Remark. Negahban et al. (2012) established a similar error bound for a more general loss function $\mathcal{L}(\boldsymbol{\beta})$ and a certain type of penalty function $\mathcal{R}(\boldsymbol{\beta})$ without the linear constraint. They assume the loss function $\mathcal{L}(\boldsymbol{\beta})$ is convex and differentiable, and satisfies the restricted strong convexity (RSC) condition. The penalty function $\mathcal{R}(\boldsymbol{\beta})$ satisfies the decomposable condition. In our case, the quadratic loss $\frac{1}{2}\boldsymbol{\beta}^\top\mathbf{W}\boldsymbol{\beta} - \mathbf{q}^\top\boldsymbol{\beta}$ satisfies the RSC condition. However, the folded-concave penalty functions do not satisfy the decomposable condition. By local linear approximation, we replace the folded concave penalty by weighted L_1 penalty, which satisfies the decomposable condition. Furthermore, we allow linear equality constraint imposed on $\boldsymbol{\beta}$.

Next, we establish the strong oracle property of the LLA estimator. Let

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{\mathcal{A}\mathcal{A}} & \mathbf{W}_{\mathcal{A}\mathcal{A}^c} \\ \mathbf{W}_{\mathcal{A}^c\mathcal{A}} & \mathbf{W}_{\mathcal{A}^c\mathcal{A}^c} \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} \mathbf{q}_{\mathcal{A}} \\ \mathbf{q}_{\mathcal{A}^c} \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \mathbf{C}_{\mathcal{A}} & \mathbf{C}_{\mathcal{A}^c} \end{pmatrix}. \tag{C.3}$$

The oracle estimator is the estimator that we know the true support of $\boldsymbol{\beta}^\star$ in advance. Define the oracle estimator as $\widehat{\boldsymbol{\beta}}^{(o)} = (\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(o)}, \mathbf{0})$, where $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(o)}$ is the solution to the following problem

$$\begin{aligned}
&\arg\min_{\boldsymbol{\beta}_{\mathcal{A}}} \frac{1}{2}\boldsymbol{\beta}_{\mathcal{A}}^\top \mathbf{W}_{\mathcal{A}\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} - \mathbf{q}_{\mathcal{A}}^\top\boldsymbol{\beta}_{\mathcal{A}} \\
&\quad \text{s.t. } \mathbf{C}_{\mathcal{A}}\boldsymbol{\beta}_{\mathcal{A}} = \mathbf{b}.
\end{aligned} \tag{C.4}$$

If $\mathbf{W}_{\mathcal{A}\mathcal{A}}$ is non-singular, we can derive the closed form solution of the oracle

estimator. As a result, we have $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(o)} = \mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{q}_{\mathcal{A}} - \mathbf{C}_{\mathcal{A}}^{\top}\boldsymbol{\gamma})$, where

$$\boldsymbol{\gamma} = (\mathbf{C}_{\mathcal{A}}\mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{C}_{\mathcal{A}}^{\top})^{-1}(\mathbf{C}_{\mathcal{A}}\mathbf{W}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{q}_{\mathcal{A}} - \mathbf{b}). \tag{C.5}$$

To establish the strong oracle property, we further specify the following conditions

(C.B.1) (minimal signal condition) The minimal signal of $\boldsymbol{\beta}^{\star}$ satisfies

$$\min_{j\in\mathcal{A}}|\beta_j^{\star}| > (a_2 + a_3)\lambda. \tag{C.6}$$

(C.B.2) There exists a $r \times 1$ vector $\mathbf{u}$ such that

$$\|\mathbf{W}_{\mathcal{A}^c}\widehat{\boldsymbol{\beta}}^{(o)} + \mathbf{C}_{\mathcal{A}^c}^{\top}\mathbf{u}\|_{\infty} \leq a_0\lambda, \tag{C.7}$$

where $\mathbf{W}_{\mathcal{A}^c} = (\mathbf{W}_{\mathcal{A}^c\mathcal{A}}\ \mathbf{W}_{\mathcal{A}^c\mathcal{A}^c})$.

Theorem C.2 (Strong oracle property)**.** *Let $\widehat{\boldsymbol{\beta}}(\lambda)$ be the LLA estimator to (3.7) with parameter λ, under the conditions of Theorem C.1 and conditions (C.B.1) - (C.B.2) hold, then the LLA estimator $\widehat{\boldsymbol{\beta}}(\lambda)$ equals to the oracle estimator*

$$\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{\beta}}^{(o)}. \tag{C.8}$$

Proof. By definition, $\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}:\mathbf{C}\boldsymbol{\beta}=\mathbf{b}}\ Q_{\lambda}(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^{(1)})$, where

$$Q_{\lambda}(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^{(1)}) = \frac{1}{2}\boldsymbol{\beta}^{\top}\mathbf{W}\boldsymbol{\beta} - \mathbf{q}^{\top}\boldsymbol{\beta} + \sum_{j=1}^{p} p_{\lambda}'(|\widehat{\beta}_j^{(1)}|)|\beta_j| \tag{C.9}$$

Since the objective function $Q_{\lambda}(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^{(1)})$ is a convex function of $\boldsymbol{\beta}$, the KKT condition is necessary and sufficient for characterizing the minimum. To verify that $\widehat{\boldsymbol{\beta}}^{(o)}$ is the minimizer of $Q_{\lambda}(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}^{(1)})$, it is sufficient to show that there exists a vector $\mathbf{u}$ such that

$$\mathbf{w}_{(j)}^{\top}\widehat{\boldsymbol{\beta}}^{(o)} - q_j + \mathrm{sgn}(\beta_j^{(o)})p_{\lambda}'(|\widehat{\beta}_j^{(1)}|) + \mathbf{u}^{\top}\mathbf{c}_{(j)} = 0,\ j \in \mathcal{A}, \tag{C.10}$$

and

$$|\mathbf{w}_{(j)}^{\top}\widehat{\boldsymbol{\beta}}^{(o)} - q_j + \mathbf{u}^{\top}\mathbf{c}_{(j)}| \leq a_0\lambda,\ j \in \mathcal{A}^c. \tag{C.11}$$

where $\mathbf{w}_{(j)}^\top$ is the j-th column of $\mathbf{W}$, $\mathbf{c}_{(j)}$ is the j-th column of $\mathbf{C}$ and q_j is the j-th element of $\mathbf{q}$.

We first verify (C.10). Note that with the initial value $\mathbf{0}$, we have $\widehat{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}:\mathbf{C}\boldsymbol{\beta}=\mathbf{b}} Q_{\tau\lambda}(\boldsymbol{\beta}|\mathbf{0})$. From the proof of Theorem C.1, we know $\|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^\star\|_\infty \leq \|\widehat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}^\star\|_2 \leq a_2\lambda$ and thus $|\widehat{\beta}_j - \beta_j^\star|_\infty \leq a_2\lambda$ for all $j \in \mathcal{A}$. By the minimal signal condition, we have $|\widehat{\beta}_j^{(1)}| > a_3\lambda$ for $j \in \mathcal{A}$ and therefore $p'_\lambda(|\widehat{\beta}_j^{(1)}|) = 0$ for $j \in \mathcal{A}$. Note that $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(o)}$ is the solution to (C.4), then $\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(o)}$ satisfies its KKT condition, that is, there exists a vector $\mathbf{u}$ such that

$$\mathbf{W}_{\mathcal{A}\mathcal{A}}\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(o)} - \mathbf{q}_{\mathcal{A}} + \mathbf{C}_{\mathcal{A}}^\top\mathbf{u} = \mathbf{0}. \tag{C.12}$$

(C.12) implies that for all $j \in \mathcal{A}$, we have

$$\mathbf{w}_{(j)}^\top\widehat{\boldsymbol{\beta}}^{(o)} - q_j + \mathbf{v}^\top\mathbf{c}_{(j)} = 0. \tag{C.13}$$

Therefore (C.10) holds. (C.11) is implied by condition (C.7), which completes the proof.

□

Appendix D
Central Limit Theorem for Martingale Difference

Definition D.1. *A stochastic process* $\{X_n, n \geq 1\}$ *is a martingale if*

(i) $E(|X_n|) \leq \infty$ *for all* n *and*

(ii) $E(X_{n+1}|X_1, \ldots, X_n) = X_n$.

Definition D.2. *A stochastic process* $\{X_n, n \geq 1\}$ *is a martingale difference if*

(i) $E(|X_n|) \leq \infty$ *for all* n *and*

(ii) $E(X_{n+1}|X_1, \ldots, X_n) = 0$.

From the definitions above, if $X_1, X_2, \ldots$ is a martingale difference, then the partial sum $S_n = \sum_{i=1}^n X_i$ is a martingale. Next we state a lemma on the central limit theorem for martingale difference. This lemma is a special case of Theorem 3.2 in Hall and Heyde (2014).

Lemma D.1. *Let* $X_1, \ldots, X_n$ *be a sequence of martingale difference and*

$$E(\max_i |X_i|) \to 0 \text{ and } \sum_{i=1}^n X_i^2 \xrightarrow{p} \sigma^2,$$

then $S_n \xrightarrow{d} N(0, \sigma^2)$, *where* $S_n = \sum_{i=1}^n X_i$.

Proof. The proof can be found in Hall and Heyde (2014). □

Bibliography

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

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

Bickel, P. J. and Levina, E. (2008), 'Covariance regularization by thresholding', *The Annals of Statistics* **36**(6), 2577–2604.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009), 'Simultaneous analysis of Lasso and Dantzig selector', *The Annals of Statistics* **37**(4), 1705–1732.

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J. (2011), 'Distributed optimization and statistical learning via the alternating direction method of multipliers', *Foundations and Trends® in Machine Learning* **3**(1), 1–122.

Cai, T. and Liu, W. (2011*a*), 'Adaptive thresholding for sparse covariance matrix estimation', *Journal of the American Statistical Association* **106**(494), 672–684.

Cai, T. and Liu, W. (2011*b*), 'A direct estimation approach to sparse linear discriminant analysis', *Journal of the American Statistical Association* **106**(496), 1566–1577.

Cai, T., Liu, W. and Luo, X. (2011), 'A constrained ℓ_1 minimization approach to sparse precision matrix estimation', *Journal of the American Statistical Association* **106**(494), 594–607.

Cai, T., Liu, W. and Xia, Y. (2014), 'Two-sample test of high dimensional means under dependence', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(2), 349–372.

Candes, E. and Tao, T. (2007), 'The Dantzig selector: Statistical estimation when p is much larger than n', *The Annals of Statistics* **35**(6), 2313–2351.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997), 'Generalized partially linear single-index models', *Journal of the American Statistical Association* **92**(438), 477–489.

Caruana, R. (1997), 'Multitask learning', *Machine Learning* **28**(1), 41–75.

Chen, S. X., Li, J. and Zhong, P.-S. (2014), 'Two-sample tests for high dimensional means with thresholding and data transformation', *arXiv preprint arXiv:1410.2848* .

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**(2), 808–835.

Chen, X., Xu, M. and Wu, W. B. (2015), 'Regularized estimation of linear functionals for high-dimensional time series', *arXiv preprint arXiv:1506.03832* .

Consortium, W. T. C. C. (2007), 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature* **447**(7145), 661.

Cui, H., Li, R. and Zhong, W. (2015), 'Model-free feature screening for ultrahigh dimensional discriminant analysis', *Journal of the American Statistical Association* **110**(510), 630–641.

Dempster, A. (1958), 'A high dimensional two sample significance test', *The Annals of Mathematical Statistics* **29**(4), 995–1010.

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

Fan, J., Feng, Y. and Song, R. (2011), 'Nonparametric independence screening in sparse ultra-high-dimensional additive models', *Journal of the American Statistical Association* **106**(494), 544–557.

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

Fan, J., Ke, Y. and Wang, K. (2016), 'Decorrelation of covariates for high dimensional sparse regression', *arXiv preprint arXiv:1612.08490* .

Fan, J. and Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Fan, J., Liao, Y. and Liu, H. (2016), 'An overview of the estimation of large covariance and precision matrices', *The Econometrics Journal* **19**(1).

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

Fan, J. and Lv, J. (2010), 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica* **20**(1), 101.

Fan, J., Ma, Y. and Dai, W. (2014), 'Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models', *Journal of the American Statistical Association* **109**(507), 1270–1284.

Fan, J., Samworth, R. and Wu, Y. (2009), 'Ultrahigh dimensional feature selection: beyond the linear model', *Journal of Machine Learning Research* **10**(Sep), 2013–2038.

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

Fan, J., Xue, L. and Zou, H. (2014), 'Strong oracle optimality of folded concave penalized estimation', *The Annals of Statistics* **42**(3), 819.

Fang, E. X., He, B., Liu, H. and Yuan, X. (2015), 'Generalized alternating direction method of multipliers: new theoretical insights and applications', *Mathematical Programming Computation* **7**(2), 149–187.

Frank, L. E. and Friedman, J. H. (1993), 'A statistical view of some chemometrics regression tools', *Technometrics* **35**(2), 109–135.

Freund, Y. and Schapire, R. E. (1997), 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and System Sciences* **55**(1), 119–139.

Hall, P. and Heyde, C. C. (2014), *Martingale limit theory and its application*, Academic Press.

Hall, P. and Miller, H. (2009), 'Using generalized correlation to effect variable selection in very high dimensional problems', *Journal of Computational and Graphical Statistics* **18**(3), 533–550.

Hardle, W., Hall, P. and Ichimura, H. (1993), 'Optimal smoothing in single-index models', *The Annals of Statistics* **21**(1), 157–178.

Hardle, W., Liang, H. and Gao, J. (2012), *Partially linear models*, Springer Science & Business Media.

He, X., Wang, L. and Hong, H. G. (2013), 'Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data', *The Annals of Statistics* **41**(1), 342–369.

Hoeffding, W. (1948), 'A non-parametric test of independence', *The Annals of Mathematical Statistics* **19**(4), 546–557.

Hotelling, H. (1931), 'The generalization of student's ratio', *The Annals of Mathematical Statistics* **2**(3), 360–378.

Huber, P. J. (1964), 'Robust estimation of a location parameter', *The Annals of Mathematical Statistics* **35**(1), 73–101.

Knight, K. and Fu, W. (2000), 'Asymptotics for lasso-type estimators', *The Annals of Statistics* **28**(5), 1356–1378.

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

Lauter, J. (1996), 'Exact t and F tests for analyzing studies with multiple endpoints', *Biometrics* **52**(3), 964–970.

Li, D. and Xue, L. (2015), 'Joint limiting laws for high-dimensional independence tests', *arXiv preprint arXiv:1512.08819* .

Li, D., Xue, L. and Zou, H. (2018), 'Applications of Peter Hall's martingale limit theory to estimating and testing high dimensional covariance matrices', *Statistica Sinica* **28**(4), 2657–2670.

Li, R., Huang, Y., Wang, L. and Xu, C. (2015), 'Projection test for high-dimensional mean vectors with optimal direction'.

Li, R., Zhong, W. and Zhu, L. (2012), 'Feature screening via distance correlation learning', *Journal of the American Statistical Association* **107**(499), 1129–1139.

Lin, L., Sun, J. and Zhu, L. (2013), 'Nonparametric feature screening', *Computational Statistics & Data Analysis* **67**, 162–174.

Liu, H., Yao, T. and Li, R. (2016), 'Global solutions to folded concave penalized nonconvex learning', *Annals of Statistics* **44**(2), 629.

Liu, J., Li, R. and Wu, R. (2014), 'Feature selection for varying coefficient models with ultrahigh-dimensional covariates', *Journal of the American Statistical Association* **109**(505), 266–274.

Liu, W. and Luo, X. (2012), 'High-dimensional sparse precision matrix estimation via sparse column inverse operator', *arXiv preprint arXiv:1203.3896* **38**.

Loh, P.-L. and Wainwright, M. J. (2015), 'Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima', *Journal of Machine Learning Research* **16**(1), 559–616.

Loh, P.-L. and Wainwright, M. J. (2017), 'Support recovery without incoherence: A case for nonconvex regularization', *The Annals of Statistics* **45**(6), 2455–2482.

Lopes, M., Jacob, L. and Wainwright, M. J. (2011), A more powerful two-sample test in high dimensions using random projection, *in* 'Advances in Neural Information Processing Systems', pp. 1206–1214.

Ma, S., Li, R. and Tsai, C.-L. (2017), 'Variable screening via quantile partial correlation', *Journal of the American Statistical Association* **112**(518), 650–663.

Mai, Q. and Zou, H. (2012), 'The Kolmogorov filter for variable screening in high-dimensional binary classification', *Biometrika* **100**(1), 229–234.

Mai, Q. and Zou, H. (2015), 'The fused Kolmogorov filter: a nonparametric model-free screening method', *The Annals of Statistics* **43**(4), 1471–1497.

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

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

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012), 'A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers', *Statistical Science* **27**(4), 538–557.

Pan, G. and Zhou, W. (2011), 'Central limit theorem for hotelling's T_2 statistic under large dimension.', *The Annals of Applied Probability* **21**(5), 1860–1910.

Pan, R., Wang, H. and Li, R. (2016), 'Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening', *Journal of the American Statistical Association* **111**(513), 169–179.

Percival, C. J., Huang, Y., Jabs, E. W., Li, R. and Richtsmeier, J. T. (2014), 'Embryonic craniofacial bone volume and bone mineral density in fgfr2+/p253r and nonmutant mice', *Developmental Dynamics* **243**(4), 541–551.

Raskutti, G., Wainwright, M. J. and Yu, B. (2011), 'Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls', *IEEE transactions on information theory* **57**(10), 6976–6994.

Rothman, A. J., Levina, E. and Zhu, J. (2009), 'Generalized thresholding of large covariance matrices', *Journal of the American Statistical Association* **104**(485), 177–186.

Rudelson, M. and Zhou, S. (2012), Reconstruction from anisotropic random measurements, *in* 'Conference on Learning Theory', pp. 10–1.

Segal, M. R., Dahlquist, K. D. and Conklin, B. R. (2003), 'Regression approaches for microarray data analysis', *Journal of Computational Biology* **10**(6), 961–980.

Serfling, R. J. (1980), *Approximation theorems of mathematical statistics*, Vol. 162, John Wiley & Sons.

Srivastava, M. S. (2009), 'A test for the mean vector with fewer observations than the dimension under non-normality', *Journal of Multivariate Analysis* **100**(3), 518–532.

Srivastava, M. S. and Du, M. (2008), 'A test for the mean vector with fewer observations than the dimension', *Journal of Multivariate Analysis* **99**(3), 386–402.

Székely, G. J. and Rizzo, M. L. (2010), 'Brownian distance covariance', *The Annals of Applied Statistics* **3**(4), 1236–1265.

Székely, G. J. and Rizzo, M. L. (2014), 'Partial distance correlation with methods for dissimilarities', *The Annals of Statistics* **42**(6), 2382–2412.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), 'Measuring and testing dependence by correlation of distances', *The Annals of Statistics* **35**(6), 2769–2794.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1), 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2003), 'Class prediction by nearest shrunken centroids, with applications to DNA microarrays', *Statistical Science* **18**(1), 104–117.

Vapnik, V. (2013), *The Nature of Statistical Learning Theory*, Springer Science & Business Media.

Wang, H. (2012), 'Factor profiled sure independence screening', *Biometrika* **99**(1), 15–28.

Wang, L., Kim, Y. and Li, R. (2013), 'Calibrating non-convex penalized regression in ultra-high dimension', *The Annals of Statistics* **41**(5), 2505.

Wang, Z., Liu, H. and Zhang, T. (2014), 'Optimal computational and statistical rates of convergence for sparse nonconvex learning problems', *The Annals of statistics* **42**(6), 2164.

Wu, Y. and Yin, G. (2015), 'Conditional quantile screening in ultrahigh-dimensional heterogeneous data', *Biometrika* **102**(1), 65–76.

Xu, C. and Chen, J. (2014), 'The sparse mle for ultrahigh-dimensional feature screening', *Journal of the American Statistical Association* **109**(507), 1257–1269.

Xu, G., Lin, L., Wei, P. and Pan, W. (2016), 'An adaptive two-sample test for high-dimensional means', *Biometrika* **103**(3), 609–624.

Xue, L. and Zou, H. (2011), 'Sure independence screening and compressed random sensing', *Biometrika* **98**(2), 371–380.

Xue, L., Zou, H. and Cai, T. (2012), 'Nonconcave penalized composite conditional likelihood estimation of sparse ising models', *The Annals of Statistics* **40**(3), 1403–1429.

Yang, G., Yu, Y., Li, R. and Buu, A. (2016), 'Feature screening in ultrahigh dimensional Cox's model', *Statistica Sinica* **26**, 881.

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

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

Zhang, C.-H. (2010), 'Nearly unbiased variable selection under minimax concave penalty', *The Annals of Statistics* **38**(2), 894–942.

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

Zhong, W. and Zhu, L. (2015), 'An iterative approach to distance correlation-based sure independence screening', *Journal of Statistical Computation and Simulation* **85**(11), 2331–2345.

Zhu, L., Li, L., Li, R. and Zhu, L. (2011), 'Model-free feature screening for ultrahigh-dimensional data', *Journal of the American Statistical Association* **106**(496), 1464–1475.

Zhu, L., Xu, K., Li, R. and Zhong, W. (2017), 'Projection correlation between two random vectors', *Biometrika* **104**(4), 829–843.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**(476), 1418–1429.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.

Zou, H. and Li, R. (2008), 'One-step sparse estimates in nonconcave penalized likelihood models', *The Annals of Statistics* **36**(4), 1509.

Vita

Wanjun Liu

Education

Ph.D. in Statistics, The Pennsylvania State University (2019)

B.S. in Statistics, University of Science and Technology of China (2014)