The Pennsylvania State University
The Graduate School
Eberly College of Science

NEW STATISTICAL ANALYTIC TOOLS FOR HIGH DIMENSIONAL DATA

A Dissertation in
Statistics
by
Songshan Yang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2018

The dissertation of Songshan Yang was reviewed and approved* by the following:

Runze Li
Eberly Family Chair Professor of Statistics and Professor of Public Health Sciences
Dissertation Advisor, Chair of Committee

Bing Li
Professor of Statistics

Matthew Reimherr
Assistant Professor of Statistics

Jingzhi Huang
Professor of Finance, David H. McKinley Professor of Business , Professor of Mathematics

Naomi S. Altman
Professor of Statistics
Associate Head for Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

This dissertation studies the feature screening and two-sample mean testing procedures for high-dimensional data. Firstly, a new feature screening procedure based on the joint quasi-likelihood is proposed for generalized varying coefficient models. Secondly, we propose a new testing method considering the correlation structure for high-dimensional mean vectors.

Generalized varying coefficient models are particularly useful for examining dynamic effects of covariates on a continuous, binary or count response. This dissertation is concerned with feature screening for generalized varying coefficient models with ultrahigh dimensional covariates. The proposed screening procedure is based on joint quasi-likelihood of all predictors, and therefore is distinguished from marginal screening procedures proposed in the literature. In particular, the proposed procedure can effectively identify active predictors that are jointly dependent but marginal independent of the response. In order to carry out the proposed procedure, we propose an effective algorithm and establish the ascent property of the proposed algorithm. We further prove that the proposed procedure possesses the sure screening property. That is, with probability tending to one, the selected variable set includes the actual active predictors. We examine the finite sample performance of the proposed procedure and compare it with existing ones via Monte Carlo simulations, and illustrate the proposed procedure by a real data example.

Testing the population mean is fundamental in statistical inference. The traditional Hotelling's T^2 test becomes practically infeasible due to the singularity of sample covariance matrix when the dimensionality of the data is larger than the sample size. For a symmetric positive definite W matrix, we consider $T = (\mathbf{x}_1 - \mathbf{x}_2)^T W (\mathbf{x}_1 - \mathbf{x}_2)$ for the two sample problem. We first prove that in order to maximize the asymptotic power of T, $W = \lambda \Sigma^{-1}$ for some positive constant λ. The goal is to model correlation matrix and use the correlation to improve the

power of a test. We consider linear structure models for the inverse of correlation matrix $\Omega \hat{=} R^{-1}$: $\Omega(\boldsymbol{\theta}) = \theta_1 G_1 + \sum_{l=2}^{L} \theta_l G_l$. An estimation procedure for $\boldsymbol{\theta}$ is proposed and the asymptotic power of the proposed test by incorporating correlation information is demonstrated. We compare the performances of the proposed test and the existing methods via Monte Carlo simulations, and a real data example is also given.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to express my sincere gratitude to my advisor Professor Runze Li for the continuous support of my Ph.D study and related research, for his expert guidance, motivation, and immense knowledge, but also for the hard question which incented me to widen my research from various perspectives. Without his incredibly patience and timely help, my thesis work would have been a frustrating and overwhelming pursuit.

I would also like to thank the rest of the members of my doctoral committee: Professor Bing Li, Professor Jinzhi Huang and Professor Reimherr for having served on my committee. Without their motivation and valuable advice, the thesis work would not have been successful. In addition, I want to thank all the people who helped me throughout this academic exploration.

Last but not the least, I would like to thank my parents for their unconditional love and support during my whole life.

This dissertation research was supported by National Institute on Drug Abuse (NIDA) grants P50 DA039838 and P50 DA036107, and National Science Foundation grants DMS 1512422 and DMS 1820702. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, the NIDA, the NIH, or the NNSFC

Chapter 1

Introduction

1.1 An Overview of Variable Selection and Feature Selection

High-dimensional data analysis problems have arisen in the areas such as genomics, proteomics, finance, biomedical imaging, tomography and tumor classifications. The classical statistical methods are challenged when the number of features can be much greater than the sample size, which motivates the statisticians to develop new methodologies for the analysis of high-dimensional data. Fan and Li (2007) gave a comprehensive overview of statistical challenges of high dimensionality and Fan (2014a) introduced the challenges in the analysis of big data problems. Variable selection and feature selection have been the most popular topics in high-dimensional data analysis in the last two decades.

Traditional variable selection methods such as AIC, BIC and Mallow's C_p are not applicable to high-dimensional data due to their tremendous computational cost. Penalized regression methods such as nonnegative garrote (Breiman, 1995), least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and minimax concave penalty MCP (Zhang, 2010) can select significant variables and estimate regression coefficients simultaneously and have been widely used in high-dimensional analysis. However, since modern applications in some areas such as genomics and proteomics generate ultrahigh-dimensional data whose number of predictors grows

exponentially with sample size, the aforementioned variable selection techniques may fail due to the computational complexity.

The difficulty in the analysis of ultrahigh-dimensional data motivates researchers to create new statistical methods. Donoho (2005; 2006) proved the individual equivalence of the minimal L_1-norm and the minimal L_0-norm solutions. Candes and Tao (2007) proposed the Dantzig selector for a linear model with much more predictors than observations. On the other hand, several feature screening methods based on marginal utilities have also been proposed. Fan and Lv (2008) first introduced the concept of sure screening property in ultrahigh-dimensional data analysis and proposed the sure independence screening (SIS) and the iterated sure independence screening (ISIS) for linear regression models. Hall and Miller (2009) proposed a feature ranking method using a generalized empirical correlation learning and extended the feature selection method to nonlinear models. Fan et al. (2009) and Fan and Song (2010) further extended SIS and ISIS from linear regression model to generalized linear regression models. Fan et al. (2011) developed the nonparametric feature screening technique based on B-spline expansion for the ultrahigh-dimensional additive model. Zhu et al. (2011) proposed a sure independence ranking and screening (SIRS) procedure to select important predictors in the multi-index model; Li et al. (2012b) proposed a model-free sure independence screening procedure based on the distance correlation (DC-SIS). However, the marginal screening methods fail to identify the active predictors which are marginally independent but jointly dependent of response. Sometimes the marginal screening methods select the inactive predictors which are highly correlated with the important predictors.

Varying coefficient models with ultrahigh-dimensional covariates (features) could be very useful for analyzing genetic study data to examine varying gene effects. The collected data set frequently has an ultrahigh dimensionality p that is allowed to diverge at a nonpolynomial (NP) rate with the sample size n, namely $\log(p) = O(n^a)$ for some $a > 0$. Traditional statistical methods confront significant challenges in dealing with such high-dimensional data sets. Fan et al. (2014b) extended the nonparametric B-spline method for varying coefficient models and proposed a marginal sure screening procedure. Liu et al. (2014) proposed another marginal sure screening procedure based on the conditional correlation co-

efficient for varying coefficient models. But those two methods also have the same drawbacks as other marginal screening methods. Wang (2009) proposed a forward regression approach to feature screening in ultrahigh dimensional linear models, Xu and Chen (2014) proposed a feature screening procedure for generalized linear models via the sparsity-restricted maximum likelihood estimator and Yang et al. (2016) proposed sure joint screening for the Cox's model. As demonstrated in Wang (2009), Xu and Chen (2014) and Yang et al. (2016), their approaches can perform better than the marginal screening procedures, and can effectively identify predictors that are jointly dependent but marginally independent of the response.

In this thesis, we propose a new feature screening procedure for ultrahigh-dimensional generalized varying coefficient linear models. The proposed procedure is distinguished from the existing sure independence screening (SIS) procedures (Fan and Song, 2010, Fan, Ma and Dai, 2014b, Liu et al., 2014) in that the proposed procedure is based on joint likelihood of potential active predictors, and therefore is not a marginal screening procedure. We also demonstrate that the newly proposed procedure can outperform the marginal screening procedure for the ultrahigh-dimensional generalized varying coefficient linear models. This thesis makes the following major contributions to the literature.

(a) We propose a sure joint screening (SJS) procedure for ultrahigh dimensional generalized varying-coefficient models. We further propose an effective algorithm to carry out the proposed screening procedure, and demonstrate the ascent property of the proposed algorithm.

(b) We establish the screening property for the proposed joint screening procedure.

The proposed procedure can effectively identify active predictors that are jointly dependent but marginally independent of the response without performing an iterative procedure. We develop a computationally effective algorithm to carry out the proposed procedure and establish the ascent property of the proposed algorithm. We further prove that the proposed procedure possesses the sure screening property. That is, with probability tending to one, the selected variable set includes the actual active predictors.

1.2 A Brief Introduction of Two Sample Mean Testing Problems in High-dimensional Data Analysis

The research of testing the equivalence of two-sample mean vectors has been well developed in classical multivariate analysis, but it confronts the new challenge in high dimensional data analysis. Suppose that for $i = 1, 2$, $\{\mathbf{x}_{ij}, j = 1, \cdots, N_i\}$ is a random sample from a population $\mathbf{x}_i$ with finite mean vectors $\boldsymbol{\mu}_i$ and finite positive definite covariance matrix Σ. The two sample mean testing problem is to test

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ vs. } H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{1.2.1}$$

The classical Hotelling's T^2 test is used in the two-sample mean testing problem when $n = (N_1 + N_2 - 2) > p$ and $\mathbf{x}_{ij} \sim N_p(\boldsymbol{\mu}_i, \Sigma), i = 1, 2$. The test statistic is defined by

$$T^2 = \frac{N_1 N_2}{N_1 + N_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{1.2.2}$$

where $\bar{\mathbf{x}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{x}_{ij}, i = 1, 2$, and $S = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{N_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T$.

Under the null hypothesis H_0,

$$\frac{n - p + 1}{np} T^2 \sim F_{p,n+1-p}, \tag{1.2.3}$$

hence we reject the null hypothesis when

$$T^2 > F_{p,n+1-p}(\alpha), \tag{1.2.4}$$

where $F_{p,n+1-p}(\alpha)$ is the $1 - \alpha$ quantile of the distribution $F_{p,n+1-p}$.

Testing the hypothesis in (1.2.1) becomes challenging for high-dimensional data and attracts lots of researchers to create new testing methods. The traditional Hotelling's T^2 test given by (1.2.2) is not well defined because of the singularity of S when $p > N_1 + N_2$. Bai and Saranadasa (1996) and Pan and Zhou (2011) demonstrated that the power of the Hotelling's T^2 test can be adversely affected even when p is close to $N_1 + N_2$, since S is nearly not invertible. Bai and Saranadasa

(1996), Srivastava and Du (2008), Srivastava (2009) developed several new theories and methodologies in the two-sample mean testing problems in a large dimensional setting with $p/N \to \kappa \in (0,1)$. Lee et al. (2012), Srivastava et al. (2013), Chen and Qin (2010) and Thulin (2014) extended their research into a high-dimensional setting without imposing condition $\kappa \to (0,1)$. Chen et al. (2011) proposed to use a ridge-type covariance matrix estimator $S + \lambda I_p$ to replace S in (1.2.2) and introduced regularized Hotelling's T^2 test. Some researchers considered projecting the high-dimensional samples to a low-dimensional space and then processing the classical Hotelling's T^2 test. Lopes et al. (2011a; 2011b) constructed the random projection test and suggested projecting the high-dimensional sample to a $[n/2]$(the integer part of $n/2$)-dimensional space. Li et al. (2015) derived the theoretical optimal direction with which the projection test possesses the best power under alternatives and used a sample-splitting strategy to construct an exact t-test.

The aforementioned test methods included approximations of the covariance matrix Σ. Bai and Saranadasa 1996, Srivastava and Du (2008), and Chen and Qin (2010) replaced the covariance matrix by diagonal estimators that make no essential use of correlation structure. Chen et al. (2011) and Li et al. (2015) used a ridge-type covariance matrix estimator $S + \lambda I_p$. However, those estimates may not be accurate enough and may affect the power of corresponding tests. This thesis proposes to model the correlation matrix (R) and toimprove the power of a test that involves the correlation matrix. This thesis assumes the inverse of the correlation matrix R^{-1} can be represented as a linear combination of a set of matrix bases:

$$\mathbf{R}^{-1} = \theta_1 \mathbf{A}_1 + \cdots + \theta_K \mathbf{A}_K.$$

We propose estimating $\boldsymbol{\theta}$ by minimizing the following quadratic loss

$$\min_{\boldsymbol{\theta}} \operatorname{tr}[\hat{\mathbf{R}}(\theta_1 \mathbf{A}_1 + \cdots \theta_K \mathbf{A}_K) - \mathbf{I}_p]^2, \tag{1.2.5}$$

where $\hat{R}$ is the sample correlation matrix.

The thesis shows that as the sample size goes to infinity, the ratio between the minimizer of (1.2.5) and the true θ goes to a constant and the asymptotic joint distribution of $\hat{\theta}_k, k = 1, 2, \cdots, K$ is a normal distribution.

In practical implementation, we may introduce a relative large number of $\mathbf{A}_k$s

into the model (1.2.5) to reduce approximation error. Thus, we introduce regularization method to reduce model complexity of model (1.2.5). The contribution of this project can be summarized as follows:

(1) We propose a new hypothesis testing method for two sample mean problem of high dimensional by considering the linear structure of the precision matrix;
(2) We derive the limiting null distribution of the new test statistic under both the null hypothesis and the alternative hypothesis;
(3) We also propose the idea of using regularization method to select the matrix bases;
(4) The numerical studies show the outstanding performance of the new method when there exists strong correlations among variables.

1.3 Organization of this dissertation

The rest of the dissertation is organized as follows. Chapter 2 provides a literature review of feature screening methods and existing two-sample mean testing methods in high-dimensional data analysis. Chapter 3 proposes a new feature screening method for the ultrahigh-dimensional generalized varying-coefficient linear models and further demonstrates the ascent property of the proposed algorithm carrying out the proposed feature screening procedure. It further studies the sampling property of the proposed procedure and establish its sure screening property. In addition, it presents numerical comparisons, an empirical analysis of a real data example, and some discussions. Chapter 4 proposes the theoretical properties of the regularization method for precision matrix estimation and provides the application of such an estimate to test the two-sample mean problem. It then demonstrates the gain in power by incorporating correlation information. A real data example is shown to compare our test to other existing testing methods. Chapter 5 summarizes the research in this thesis and discusses the possible applications of the proposed methods in the future.

Chapter 2

Literature Review

In this chapter, we give a brief literature review on three topics: (1) variable selection via penalized regression methods; (2) feature screening procedures for ultrahigh-dimensional data and (3) two-sample mean testing techniques for high-dimensional data. First, we review some penalized regression methods and related algorithms. Second, we briefly review some feature screening procedures based on different models such as linear models, generalized linear models and time varying coefficient models, and also introduce some feature screening procedures considering the joint effect among predictors; and finally, we review some statistical methods for the two-sample mean testing problems in the high-dimensional data analysis.

2.1 Variable Selection via Penalized Regression Methods

Although the subset selection procedures based on the classical criteria admit nice sampling properties (Barron, Birge and Massart, 1999), they are infeasible when the number of predictors is large due to the heavy computational cost. To address this issue, researchers advocate using penalized regression approaches to select the important variables and estimate the regression coefficients simultaneously. In this section, we mainly introduce the penalized methods and numerical algorithms used in the thesis.

Consider linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{2.1.1}$$

where $\mathbf{Y} = (Y_1, ..., Y_n)^T$ is an n-element response vector, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ is an $n \times p$ design matrix and $\mathbf{x}_1, ..., \mathbf{x}_n$ are independently and identically distributed (IID), $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ is a p-vector and $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ is a random error vector with n IID elements. The penalized least square function is defined by

$$Q(\boldsymbol{\beta}) = \frac{1}{2n}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.1.2}$$

where $p_\lambda(\cdot)$ is the penalty function and $\lambda \geqslant 0$ is a tuning parameter controlling the amount of shrinkage applied to the estimate. The basic idea of penalized least squares methods is to minimize (2.1.2).

Fan and Li (2001) advocated three properties of the penalized least squares method:

1. Unbiasedness: The estimator is nearly unbiased for the truly large coefficients to reduce model bias.

2. Sparsity: The estimator automatically sets small estimated coefficients to zero, to reduce model complexity.

3. Continuity: The estimator is continuous in the data, in order to guarantee the model prediction to be stable.

Furthermore, Antoniadis and Fan (2001) proposed that the penalized least squares estimator possesses the following three properties:

1. Approximate unbiasedness if $p_\lambda^{'}(t) = 0$ for large t;

2. Sparsity if $\min_{t \geqslant 0}\{t + p_\lambda^{'}(t)\} > 0$;

3. Continuity if and only if $\text{argmin}_{t \geqslant 0}\{t + p_\lambda^{'}(t)\} = 0$.

2.1.1 Least Absolute Shrinkage and Selection Operator

Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) for variable selection in linear models. The penalty function corresponding to LASSO is

$$p_\lambda(|t|) = \lambda|t|. \tag{2.1.3}$$

With this penalty function, the penalized least squares function

$$Q(\boldsymbol{\beta}) = \frac{1}{2}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.1.4}$$

This is equivalent to minimizing the residual sum of squared errors subject to the constraint $\sum_{j=1}^{p} |\beta_j| < s$, by which the model size is controlled and the sparsity is guaranteed. Provided that λ is sufficiently large, a portion of the values that make up $\boldsymbol{\beta}$ will be exactly 0 for the LASSO penalty function. Thus, the LASSO provides a continuous subset selection procedure. LASSO is also consistent for estimating $\boldsymbol{\beta}$ under appropriate conditions, which was investigated in Knight and Fu (2000).

On the other hand, LASSO has some drawbacks at the same time. First of all, LASSO cannot handle collinearity problem since it tends to select only one variable from the group and ignore the rest when the pairwise correlations exist among a group of variables. Besides, LASSO is not suitable for general factor selection since it can only select individual input variables. Another drawback of LASSO estimator is that it equally penalizes all the coefficients, resulting in the biasness of large coefficients.

2.1.2 Smoothly Clipped Absolute Deviation (SCAD) Penalty

Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty function to select variables and estimate coefficients simultaneously. The SCAD penalty function is defined by

$$p_\lambda^{'}(|t|) = \lambda I(|t| \leqslant \lambda) + \frac{(a\lambda - |t|)_+ I(|t| > \lambda)}{a - 1}. \tag{2.1.5}$$

The penalty function above is continuous and symmetric, leaving large values of the parameter λ not excessively penalized. The penalty function also satisfies all conditions for the aforementioned advocated properties, and Fan and Li (2001) suggested a practical choice for $a > 2$, and often $a = 3.7$ in SCAD from the view of Bayes risks. Furthermore, Fan and Li (2001) proved the SCAD penalized estimator possesses the oracle property. In other word, the non-zero component is estimated as well as it would have been if the correct model were known in advance. In addition, when a component of the true parameter is 0, it is estimated as 0 with probability tending to one.

2.1.3 Minimax Concave Penalty (MCP)

Zhang (2010) proposed the minimax concave penalty (MCP) and it is defined by

$$p_\lambda(|t|) = \lambda(|t| - |t|^2/2a\lambda)I(|t| < a\lambda) + \frac{a\lambda^2}{2}I(|t| \geqslant a\lambda) \tag{2.1.6}$$

where $a > 0$. MCP is motivated by and rather similar to SCAD. The MCP enjoys the aforementioned three desired properties and the oracle property. Zhang (2010) discussed the issue of choosing a in depth; $a = 3$ is suggested for penalized linear regression and $a = 30$ is suggested for penalized logistic regression.

2.1.4 Coordinate Descent Algorithms

It is hard to optimize aforementioned penalized least squares function due to the nonconvexity. However, some convex functions can be used to approximate them, thus, the nonconvex problem can be solved via convex optimization algorithms. Fan and Li (2001) proposed a unified local quadratic approximation (LQA) algorithm for optimizing nonconvex penalized least squares, the idea of which is to locally and iteratively approximate $Q(\boldsymbol{\beta})$ in (2.1.2) by a quadratic function. Zou and Li (2008) introduced the local linear approximation to the nonconvex penalty functions. Efron et al. (2004) proposed a fast and efficient Least Angle Regression (LARS) algorithm. In this section, we mainly focus on coordinate descent algorithm which successively optimizes on coordinate at a time. The procedure of coordinate descent algorithm can be summarized as:

(1) Set the initial value $\boldsymbol{\beta}_0$.

(2a) Successively optimizes $Q(\boldsymbol{\beta})$ in (2.1.2) from the first, the second,....., and the p-th coordinate while keeps other coordinates fixed.

(2b) Repeat 2a until some convergence criterion is satisfied.

Denote $\mathbf{X}_{-j}$ and $\widehat{\boldsymbol{\beta}}_{-j,0}$ as $\mathbf{X}$ and $\widehat{\boldsymbol{\beta}}_0$ with the j-th column and j-th component removed, respectively. When we are optimizing the j-th component β_j fixing other components at their current value $\boldsymbol{\beta}_0$ in (2a), we update the component by

$$\beta_j = \arg\min_{\beta_j} Q_j(\beta_j) = \arg\min_{\beta_j}(\frac{1}{2n}||\mathbf{R}_j - \mathbf{x}_j\beta_j||^2 + p_\lambda(|\beta_j|) + c) \tag{2.1.7}$$

where $\mathbf{x}_j$ is the j-th component, $\mathbf{R}_j = \mathbf{Y} - \mathbf{X}_{-j}\widehat{\boldsymbol{\beta}}_{-j,0}$ and $c = ||p_\lambda(|\widehat{\boldsymbol{\beta}}_{-j,0})||_1$ is a constant. For an orthonormal design where $\mathbf{X}^T\mathbf{X} = nI_p$, $Q_j(\beta_j)$ can be simplified to

$$Q_j(\beta_j) = \frac{1}{2}(\beta_j - \widehat{c}_j)^2 + p_\lambda(|\beta_j|), \tag{2.1.8}$$

where $\widehat{c}_j = n^{-1}\mathbf{X}_j^T\mathbf{R}_j$. For LASSO, the solution to (2.1.8) is

$$\widehat{\theta}_{LASSO} = \text{sgn}(\beta_j)(|\beta_j| - \lambda)_+. \tag{2.1.9}$$

For SCAD, the solution to (2.1.8) is

$$\widehat{\theta}_{SCAD} = \begin{cases} \text{sgn}(\beta_j)(|\beta_j| - \lambda)_+ & \text{when } |\beta_j| \leqslant \lambda; \\ \text{sgn}(\beta_j)[(a-1)|\beta_j| - a\lambda]/(a-2) & \text{when } 2\lambda < |\beta_j| \leqslant a\lambda; \\ \beta_j & \text{when } |\beta_j| > a\lambda. \end{cases} \tag{2.1.10}$$

And for MCP, the solution to (2.1.8) is

$$\widehat{\theta}_{MCP} = \begin{cases} \text{sgn}(\beta_j)(|\beta_j| - \lambda)_+/(1 - 1/a) & \text{when } |\beta_j| \leqslant a\lambda; \\ \beta_j & \text{when } |\beta_j| > a\lambda. \end{cases} \tag{2.1.11}$$

Wu and Lange (2008) first applied coordinate descent algorithm to LASSO, and Friedman et al. (2007) showed that coordinate descent algorithm is very competitive with LARS algorithm for computing the solution path.

2.2 Feature Screening for Ultrahigh-dimensional Data

The penalized regression methods reviewed in last section have been successfully applied in high-dimensional data analysis, but when the dimensionality of data grows exponentially with the sample size, they are challenged in terms of statistical accuracy, algorithm stability and computational complexity. Such ultrahigh-dimensional data analysis has gained much popularity in the modern scientific fields such as genomics and proteomics, economics and finance. In this section, we briefly introduce feature selection procedures for different models, and both the strengths and weaknesses of each method are demonstrated.

2.2.1 Linear Models and Transformed Linear Models

We first review some feature selection methods for linear models based on Pearson correlation, and for transformed linear models based on generalized correlation and rank correlation.

2.2.1.1 Pearson Correlation

Fan and Lv (2008) introduced the concept of sure screening and proposed the sure independence screening(SIS) method based on the correlation learning. The SIS method can shrink the dimensionality from high to a moderate level.

Consider a linear model (2.1.1), if all the variables are standardized with mean 0 and standard deviation 1,

$$\omega = \mathbf{X^T Y} \tag{2.2.12}$$

is the marginal correlations of predictors with response variables. Fan and Lv (2008) proposed that by sorting the p componentwise magnitudes of ω in a decreasing order, a submodel

$$\mathcal{M}_\gamma = \{1 \leqslant i \leqslant p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all}\} \tag{2.2.13}$$

can be obtained, where $\gamma \in (0, 1)$ and $[\gamma n]$ is the integer part of γn. Hence

the full model is shrunken to the submodel $\mathcal{M}_\gamma$ with size $d = [\gamma n]$ according to the marginal correlations of the predictors with the response variable. Let $\mathcal{M}_* = \{1 \leqslant i \leqslant p : \beta_i \neq 0\}$ be the true model with size s, and suppose we have the following four conditions:

1. $p > n$ and $\log(p) = O(n^\xi)$ for some $\xi \in (0, 1-2\kappa)$, where $\kappa > 0$.

2. Denote $\mathbf{z} = \Sigma^{-1/2}\mathbf{X}$ and $\mathbf{Z} = \mathbf{X}\Sigma^{-1/2}$, where $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ and $\Sigma = \text{cov}(\mathbf{X})$, then $\mathbf{z}$ has a spherically symmetric distribution. If there are $c, c_1 > 1$ and $C_1 > 0$ such that the deviation inequality

$$P\{\lambda_{\max}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T) > c_1 \text{ or } \lambda_{\min}(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^T) > 1/c_1\} \leqslant \exp(-C_1 n)$$

holds for any $n \times \tilde{p}$ submatrix $\tilde{\mathbf{Z}}$ of $\mathbf{Z}$ with $cn < \tilde{p} \leqslant p$.

3. $\text{var}(Y) = O(1)$ and, for some $\kappa \geqslant 0$ and $c_2, c_3 > 0$,

$$\min_{i \in \mathcal{M}_*} |\beta_i| \geqslant \tfrac{c_2}{n^\kappa} \text{ and } \min_{i \in \mathcal{M}_*} |\text{cov}(\beta_i^{-1}Y, X_i)| \geqslant c_3.$$

4. There are some $\tau \geqslant 0$ and $c_4 \geqslant 0$ such that

$$\lambda_{\max}(\Sigma) \leqslant c_4 n^\tau$$

Then SIS is proved to have the sure screening property as follows:

Theorem 2.2.1. *If $2\kappa + \tau < 1$, then there is some $\theta < 1 - 2\kappa - \tau$ such that, when $\gamma \sim cn^{-\theta}$, we have, for some $C > 0$,*

$$P(\mathcal{M}_* \subset \mathcal{M}_\gamma) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n)\}]$$

it indicates that $P(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1$ as $n \to \infty$.

Since SIS method enjoys the sure screening property, Fan and Lv (2008) implemented a two-stage selection method. First, we use SIS to reduce the dimensionality to a moderate level, so the full model $\{1, ..., p\}$ is shrunken to a submodel $\mathcal{M}_\gamma$ with size $d < n$, and then apply a lower dimensional model selection method to the submodel $\mathcal{M}_\gamma$ such as SCAD, LASSO, adaptive LASSO and the Dantzig Selector. From the results of their numerical studies, SIS-SCAD outperforms other

combinations and generates smaller and more accurate models. They also showed that SIS-SCAD has the oracle properties.

However, SIS is not a perfect model selection method and has some drawbacks. First, some unimportant predictors which are highly correlated with the important predictors will be included, and other important predictors that are relatively weakly correlated with the response will be neglected; Second, the important predictors that are marginally independent of the responses but jointly dependent of the responses cannot be selected by the SIS; Another drawback is that the collinearity among the predictors make the variable selection more difficult. In order to overcome these drawbacks, Fan and Lv (2008) proposed an iterative SIS method, that is to apply SIS method iteratively and in each step the residual from the model selected in the previous step is treated as the response.

2.2.1.2 Generalized Correlation and Rank Correlation

The SIS method works well for the linear regression model with ultrahigh-dimensional predictors, however, the Pearson correlation cannot be directly used to do the feature selection in the nonlinear model. In order to capture both linearity and nonlinearity, Hall and Miller (2009) proposed a feature selection method based on the generalized correlations between the response and predictors.

Hall and Miller (2009) defined the generalized correlation between the j-th predictor and Y by:

$$\rho_g(X_j, Y) = \sup_{h \in \mathcal{H}} \frac{\text{cov}\{h(X_j), Y\}}{\sqrt{\text{var}\{h(X_j)\}\text{var}(Y)}} \tag{2.2.14}$$

where $\mathcal{H}$ is a vector space generated by any given set of functions h. If we restrict $\mathcal{H}$ to a space of constant and linear functions, $\rho_g(X_j, Y)$ is the absolute value of Pearson correlation between X_j and Y. Assume that $\mathbf{Y} = (Y_1, ..., Y_n)^T$ is an n-element response vector, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ is an $n \times p$ design matrix and $\mathbf{x}_1, ..., \mathbf{x}_n$ are IID, $\rho_j(X_j, Y)$ can be estimated by

$$\widehat{\rho}_g(X_j, Y) = \sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^n \{h(X_{ij}) - \bar{h}_j\}(Y_i - \bar{Y})}{\sum_{i=1}^n \{h^2(X_{ij}) - \bar{h}_j^2\} \sum_{i=1}^n (Y_i - \bar{Y})^2}, \tag{2.2.15}$$

where $\bar{h}_j = n^{-1} \sum_{i=1}^n h(X_{ij})$ and $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$. In addition, $\sum_{i=1}^n (Y_i - \bar{Y})^2$ does

not depend on j, so

$$\widehat{\phi}_g(X_j, Y) = \sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^n \{h(X_{ij}) - \bar{h}_j\}(Y_i - \bar{Y})}{\sum_{i=1}^n \{h^2(X_{ij}) - \bar{h}_j^2\}} \tag{2.2.16}$$

can be used instead.

The generalized correlation reflects both the linear and nonlinear relationships between the predictors and the response. Hall and Miller (2009) proposed a new feature screening method using generalized correlations as the marginal utility, which is ranking all the predictors based on the magnitude of the estimated generalized correlation of each predictor. Hall and Miller (2009) also introduced a bootstrap method to assess the authority of ranking all predictors. Let $\widehat{r}_j$ denotes the rank of the j-th predictors based on the magnitude of the estimated generalized correlation of the j-th predictor. Compute the generalized correlation $\widehat{\rho}_g^*(X_j, Y)$ of j-th predictor in the bootstrapped sample, and calculate the rank $\widehat{r}^*(j)$ of the bootstrapped sample. Given a level α, a nominal $(1-\alpha)$ two-sided, equal tailed prediction interval $[\widehat{r}_-(j), \widehat{r}_+(j)]$ is computed based on the distribution of $\widehat{r}^*(j)$'s of the bootstrapped samples. Hall and Miller (2009) suggested the j-th predictor is considered as influential if $\widehat{r}_+(j) < \frac{1}{2}p$. The rule assumes that the total number of important predictors is less than p, and more than half of all the predictors are rejected by this rule. Hall and Miller (2009) also indicated that there may exist high rate of false positives under this rule. Thus $\frac{1}{2}p$ can be replaced by some smaller fraction of p.

Hall and Miller's method is based on making transformation on the covariates, an alternative way to characterize the nonlinearity is to make transformations on the response. Li et al. (2012a) defined a marginal rank correlation

$$\omega_j = \frac{1}{n(n-1)} \sum_{i \neq l} I(X_{ij} < X_{lj}) I(Y_i < Y_l) - \frac{1}{4}, \tag{2.2.17}$$

to measure the importance of the j-th predictor X_j. According to the magnitude of ω_j's, a feature selection procedure selects a submodel $\widehat{\mathcal{M}}_{\gamma_n} = \{1 \leqslant j \leqslant p : |\omega_j| > \gamma_n\}$, where γ_n is a threshold value.

Li et al. (2012a) proposed that the feature screening procedure based on the rank correlation is robust against heavy-tailed distribution and invariant under

monotonic transformations and enjoys sure screening property under some technical conditions. The feature selection method is also robust to the outliers and influence points in the observations.

2.2.2 Generalized Linear Models

Generalized linear models have been widely used in statistical research and applications. Thus, it is necessary to extend the feature selection procedures from linear models to generalized linear models. In recent years, researchers have developed some feature selection methods for generalized linear models.

2.2.2.1 Marginal Likelihood Screening

Fan et al. (2009) extended the SIS method to generalized linear models(GLIM) by ranking the marginal likelihood of each predictor. Assume the conditional density function of y with the canonical form is given by:

$$f(y|\mathbf{X}) = \exp\{y\theta(\mathbf{X}) - b(\theta(\mathbf{X})) + c(y)\} \tag{2.2.18}$$

where $b(\cdot)$ and $c(\cdot)$ are known functions and $\theta(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$. Assume that $E(Y|\mathbf{X}) = b'(\theta(\mathbf{X})) = g^{-1}(\beta_0 + \mathbf{X}^T\boldsymbol{\beta})$. $\mathbf{X}$ is a $n \times p$ matrix $(\mathbf{x}_1, ..., \mathbf{x}_p)$, and each column is standardized with mean zero and standard deviation one. Denote the negative likelihood function of the i-th observation is $l(\beta_0 + \mathbf{x}_i^T\beta, Y_i)$, so the marginal likelihood of j-th feature is

$$L_j = \min_{\beta_0,\beta_j} \sum_{i=1}^{n} l(\beta_0 + x_{ij}\beta_j, Y_i) \tag{2.2.19}$$

The SIS idea in this situation is first to compute the marginal likelihood $L = (L_1, ..., L_p)^T$ and then select the important predictors by ranking the marginal likelihood. A submodel

$$\mathcal{M}_\gamma = \{1 \leqslant i \leqslant p : |L_i| \text{ is among the first } [\gamma n] \text{ smallest of all}\} \tag{2.2.20}$$

is obtained, where $\gamma \in (0, 1)$ and $[\gamma n]$ is the integer part of γn. After implementing this method, the full model $\{1, ..., p\}$ is reduced to a moderate level $d = [\gamma n]$, then

some well-developed variable selection methods such as LASSO, adaptive LASSO, SCAD and Dantzig Selector can be used.

If we partition the sample in two parts and apply SIS to each partition, and denote the two active indices by $\tilde{\mathcal{M}}_1$ and $\tilde{\mathcal{M}}_2$, which satisfies $|\tilde{\mathcal{M}} = \tilde{\mathcal{M}}_1 \cap \tilde{\mathcal{M}}_2| = d$. The method also possesses the sure screening property.

Theorem 2.2.2. *If $2\kappa + \tau < 1$, where $\kappa > 0$ and $\tau > 0$, then for some $C > 0$, it follows*

$$P(\mathcal{M}_* \subset \tilde{\mathcal{M}}) = 1 - O[\exp\{-Cn^{1-2\kappa}/\log(n) + \log(p)\}]$$

As shown in Section 2.2.1.1, the marginal feature screening methods fail when the predictor is marginally uncorrelated but jointly related to the response, or jointly uncorrelated with the response but highly correlated with some important predictors. Fan et al. (2009) proposed an iterative feature screening method under generalized linear model which follows 4 steps:

1. Calculate the marginal likelihood vector $(L_1, ..., L_p)$ and obtain a submodel $\widehat{\mathcal{M}}_1 = \{1 \leqslant j \leqslant p : L_j$ is among the first k_1 smallest of all$\}$. Then apply some well developed variable selection methods such as LASSO and SCAD to select a new subset $\widehat{\mathcal{M}}$.

2. Then apply the SIS method to the model $\{1, ..., p\}/\widehat{\mathcal{M}}$, which is to compute:

$$L_j^{(2)} = \min_{\beta_0, \boldsymbol{\beta}_{\widehat{\mathcal{M}}}, \beta_j} \sum_{i=1}^{n} l(\beta_0 + \mathbf{X}_{i,\widehat{\mathcal{M}}}^T \boldsymbol{\beta}_{\widehat{\mathcal{M}}} + X_{ij}\beta_j, Y_i) \tag{2.2.21}$$

   where $\mathbf{X}_{i,\widehat{\mathcal{M}}}^T$ is the sub-vector of $\mathbf{X}_i$ containing the elements in $\mathcal{M}$. Select a subset $\widehat{\mathcal{M}}_2 = \{j \in \{1, ..., p\}/\widehat{\mathcal{M}} : L_j^{(2)}$ is among the first k_2 smallest of all$\}$.

3. Some large dimensional variable selection methods such as LASSO and SCAD is applied to the set $\widehat{\mathcal{M}} \cup \widehat{\mathcal{M}}_2$ to update the subset $\widehat{\mathcal{M}}$.

4. Repeat step 2 and 3 until $|\widehat{\mathcal{M}}| \geqslant d$. So the final selected model is $\widehat{\mathcal{M}}$.

Fan et al. (2009) also proved that the iterative method possesses the sure screening property.

2.2.2.2 Maximum Marginal Likelihood Estimator

Fan and Song (2010) proposed a screening method by ranking the magnitude of the maximum marginal likelihood estimator (MMLE). The generalized linear model and the negative log-likelihood function are same as those in section 2.2.2.1. When $p > n$, the minimizer of the negative log-likelihood is not well defined.

Assume that the predictors are standardized with mean zero and deviation one, the MMLE $\widehat{\boldsymbol{\beta}}_j^M$ of the j-th predictor is defined as

$$\widehat{\boldsymbol{\beta}}_j^M = (\widehat{\beta}_0^M, \widehat{\beta}_1^M) = \arg\min_{\beta_{j0},\beta_{j1}} \sum_{i=1}^{n} l(\beta_{j0} + X_{ij}\beta_{j1}, Y_i). \tag{2.2.22}$$

This could be computed quickly and the implementation is robust. Then Fan and Song (2010) also gave the definition of the population version of MMLE:

$$\widehat{\boldsymbol{\beta}}_j^M = (\widehat{\beta}_0^M, \widehat{\beta}_1^M) = \arg\min_{\beta_{j0},\beta_{j1}} El(\beta_0 + X_{ij}\beta_j, Y), \tag{2.2.23}$$

where E is the expectation under the true model.

Based on the magnitude of MMLE, a submodel could be selected:

$$\widehat{\mathcal{M}}_\gamma = \{1 \leqslant j \leqslant p : |\widehat{\beta}_{j1}^M| \geqslant \gamma_n\} \tag{2.2.24}$$

where γ_n is a predefined value. Fan and Song (2010) also established the theoretical properties of MMLE. Define the true model as $\mathcal{M}_* = \{1 \leqslant j \leqslant p : \beta_j \geqslant 0\}$ with size d. Fan and Song (2010) first proved that the marginal regression parameter $\widehat{\beta}_{j1}^M = 0$ if and only if $\text{cov}(Y, X_j) = 0$, which shows that the marginal regression parameter is in fact a measurement of the correlation between the covariate and the response. In order to prove sure screening property of MMLE, Fan and Song (2010) first established the following result:

Theorem 2.2.3. *If* $|cov(Y, X_j)| \geqslant c_1 n^{-\kappa}$ *for* j *in*$\mathcal{M}_*$ *and a positive constant* $c_1 > 0$*, then there exists a positive constant* c_2 *such that* $\min j \in \mathcal{M}_* |\beta_{j1}| \geqslant c_2 n^{-\kappa}$ *if* $b''(\cdot)$ *is bounded or* $EG(a|X_j|)|X_j|I(|X_j| \geqslant n^\eta) < dn^{-\kappa}$*, where* $0 < \eta < \kappa$*, and some small positive constants* a *and* d*, where* $G(|x|) = \sup_{|u|\leqslant|x|} |b'(u)|$.

Theorem 3 reveals that the marginal signals are stronger than the stochastic

noise when X_j's are correlated with the response. Then Fan and Song (2010) established the uniform convergence and sure screening property of MMLE under some technical conditions.

Theorem 2.2.4. *(1). If $n^{1-2\kappa}/(k_n^2 K_n^2) \to \infty$, then for any $c_3 > 0$, there exits a positve constant c_4 such that*

$$P(\max_{1\leqslant j\leqslant p} |\widehat{\beta}_{j1}^M - \beta_{j1}^M| \geqslant c_3 n^{-\kappa}) \leqslant p\{\exp(-c_4 n^{1-2\kappa}/(k_n^2 K_n^2)) + nm_1 \exp(-m_0 K_n^{\alpha})\}. \tag{2.2.25}$$

(2). By taking $\gamma_n = c_5 n^{-\kappa}$ with $c_5 \leqslant c_2/2$, we have

$$P(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\gamma_n}) \geqslant 1 - d\{\exp(-c_4 n^{1-2\kappa}/(k_n^2 K_n^2)) + nm_1 \exp(-m_0 K_n^{\alpha})\},$$

where $d = |\mathcal{M}_*|$, the size of the true model; $k_n = b'(K_n B + B) + m_0 K_n^{\alpha}/s_0$ with $s_0, m_0 > 0$, B is the upper bound of the true value of β_{j1}^M and K_n is the supremum norm of $\mathbf{X}$.

Fan and Song (2010) also indicated that the marginal screening method by ranking the magnitude of MMLE can handle the NP-dimensionality $\log(p) = o(n^{(1-2\kappa)\alpha/(\alpha+2)})$, with $\alpha = \infty$ for the case of bounded covariates. This is weaker than SIS proposed by Fan and Lv (2008) when the covariates are normal, but the method allows nonnormal covariates and other error distributions. Fan and Song (2010) showed that the number of selected variables $|\widehat{\mathcal{M}}_{\gamma_n}|$ is bounded by $O\{n^{2\kappa}\lambda_{\max}(\Sigma)\}$, where $\lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ with probability approaching one under some regularity conditions, so the value of κ determines the threshold γ_n.

Fan and Song (2010) also proposed that the MMLE screening method is equivalent to the method stated in Section 2.1 since both of them have sure screening property and the number of selected variables are of the same order of magnitude.

2.2.3 Varying Coefficient Model

Ultrahigh dimensional varying coefficient models have become more and more important in statistical research as a useful extension of linear models. In this model, the regression coefficients are changed over different subjects featured by certain

covariates. Since the number of predictors is much larger than the sample size, feature selection is fundamental for the analysis of ultrahigh dimensional varying coefficient models. In this section, we review two existing statistical feature selection methods for the varying coefficient models.

2.2.3.1 Nonparametric Independence Screening

The varying coefficient model is an important class of nonparametric regression model. It is defined as

$$Y = \sum_{j=1}^{p} \beta_j(u) X_j + \epsilon, \tag{2.2.26}$$

where $\beta_i(u)$'s are coefficient functions.

Feature screening methods are needed when the number of covariates is large. Fan et al. (2014b) proposed a nonparametric independence screening (NIS) by ranking the nonparametric marginal utility of each covariate given u.

Denote the true model $\mathcal{M}_* = \{j : 1 \leqslant j \leqslant p, E[\beta_j^2(u)] > 0\}$ with the size $d = |\mathcal{M}_*|$. Fan et al. (2014b) first fitted the marginal regression of each covariate given u, then find $a_j(u)$ and $b_j(u)$ which minimize $E\{(Y - a_j(u) - b_j(u)X_j)^2|u\}$. The expressions of the minimizers are $b_j(u) = \frac{\text{cov}[X_j, Y|u]}{\text{var}[X_j|u]}$ and $a_j(u) = E[Y|u] - b_j(u)E[X_j|u]$. Let $a_0(u) = E[Y|u]$, the nonparametric marginal utility of each covariate

$$u_j = E(a_j(u) + b_j(u)X_j)^2 - E(a_0(u))^2 = E[\frac{(\text{cov}[X_j, Y|u])^2}{\text{var}[X_j|u]}]. \tag{2.2.27}$$

Let $a_j(u)$ and $b_j(u)$ be approximated by splines method. Define

$$\mathbf{B}(u) = \{B_1(u), ..., B_{l_n}(u)\}^T$$

be a normalized B-spline basis, $\widehat{a}_j(u) = \mathbf{B}(u)^T\widehat{\eta}_j, \widehat{b}_j(u) = \mathbf{B}(u)^T\widehat{\theta}_j$, and $\widehat{a}_0(u) = \mathbf{B}(u)^T\widehat{\eta}_0$. It can be shown that $(\widehat{\eta}_j^T, \widehat{\theta}_j^T)^T = (\mathbf{Q}_{nj}^T\mathbf{Q}_{nj})^{-1}\mathbf{Q}_{nj}^T\mathbf{Y}$, $\widehat{\eta}_0 = (\mathbf{B}_n^T\mathbf{B}_n)^{-1}\mathbf{B}_n^T\mathbf{Y}$, where $\mathbf{Q}_{nj} = (\mathbf{B}_n, \Phi_{nj}) = \begin{pmatrix} \mathbf{B}(u_1)^T & X_{j1}\mathbf{B}(u_1)^T \\ \vdots & \vdots \\ \mathbf{B}(u_n)^T & X_{jn}\mathbf{B}(u_n)^T \end{pmatrix}_{n \times 2l_n}$, $\mathbf{B}_n = \begin{pmatrix} \mathbf{B}(u_1)^T \\ \vdots \\ \mathbf{B}(u_n)^T \end{pmatrix}_{n \times l_n}$

and $\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}_{n\times 1}$. Thus, the estimate of the marginal utility

$$\begin{aligned} \widehat{u}_j =& ||\widehat{a}_j(u) + \widehat{b}_j(u)\mathbf{X}_j||_n^2 - ||\widehat{a}_0(u)||_n^2 \\ =& \frac{1}{n}\sum_{i=1}^{n}(\widehat{a}_j(u_i) + \widehat{b}_j(u_i)X_{ji})^2 - \frac{1}{n}\sum_{i=1}^{n}(\widehat{a}_0(u_i))^2. \end{aligned} \tag{2.2.28}$$

A submodel $\mathcal{M}_{\gamma_n} = \{1 \leqslant j \leqslant p : \widehat{u}_j \geqslant \gamma_n\}$ can be selected, where γ_n is a predefined threshold. Fan et al. (2014b) also showed that it is equivalent to ranking the covariates by the residual sum of squares of marginal nonparametric regression, which is

$$\widehat{v}_j = ||\mathbf{Y} - \widehat{a}_j(u) - \widehat{b}_j(u)\mathbf{X}_j||_n^2, \tag{2.2.29}$$

and a submodel $\mathcal{M}_{\nu_n} = \{1 \leqslant j \leqslant p : \widehat{v}_j \leqslant \nu_n\}$, where ν_n is a predefined threshold.

Fan et al. (2014b) proved the sure screening property of the proposed method under some regularity condition. However, the proposed method also suffers the weakness of SIS such as failure of selecting important variables which are marginally independent of Y and select unimportant variables which are highly correlated with important variables. Thus, Fan et al. (2014b) adopted two iterative methods, conditional-INIS and greedy-INIS to improve the performance of the proposed method. A group penalty is needed because an estimated coefficient function vanishes if and only if all of the coefficient in the corresponding spline expansion are zero. Fan et al. (2014b) implement group-SCAD in the paper.

2.2.3.2 Conditional Correlation Learning

Liu et al. (2014) proposed a new feature selection method for the varying coefficient models based on conditional correlation coefficient (CC-SIS). The conditional correlation between the response and the j-th predictor X_j given u is defined to be

$$\rho(X_j, Y|u) = \frac{\mathrm{cov}(X_j, Y|u)}{\sqrt{\mathrm{cov}(X_j, X_j|u)\mathrm{cov}(Y, Y|u)}}, \tag{2.2.30}$$

and the marginal utility is

$$\rho_{j0}^{*} = E\{\rho^2(X_j, Y|u)\}. \tag{2.2.31}$$

In order to estimate $\rho(X_j, Y|u)$, we need to estimate five conditional means such as $E(Y|u)$, $E(Y^2|u)$, $E(X_j|u)$, $E(X_j^2|u)$ and $E(X_jY|u)$. Liu et al. (2014) used the kernel smoothing method to estimate those conditional means. The estimation of $E(Y|u)$ is

$$\widehat{E}(Y|u) = \sum_{i=1}^{n} \frac{K_h(u_i - u)y_i}{\sum_{i=1}^{n} K_h(u_i - u)}, \tag{2.2.32}$$

where $K(t)$ is a kernel function, h is a bandwidth and $K_h(t) = h^{-1}K(t/h)$. The kernel regression estimates of the other four conditional means have the similar definitions. The estimation of the conditional covariance $\widehat{\text{cov}}(X_j, y|u) = \widehat{E}(X_jy|u) - \widehat{E}(X_j|u)\widehat{E}(y|u)$, and the estimation of the conditional correlation is

$$\widehat{\rho}(X_j, Y|u) = \frac{\widehat{\text{cov}}(X_j, Y|u)}{\sqrt{\widehat{\text{cov}}(X_j, X_j|u)\widehat{\text{cov}}(Y, Y|u)}}. \tag{2.2.33}$$

The kernel regression can guarantee $\widehat{\text{cov}}(X_j, X_j|u) \geqslant 0$ and $\widehat{\text{cov}}(Y, Y|u) \geqslant 0$, and the bandwidth h of the five conditional means are required to be same.

The plug-in estimate of ρ_{j0}^{*} is

$$\widehat{\rho}_j^{*} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\rho}^2(X_j, Y|u_i). \tag{2.2.34}$$

Based on the magnitude of $\widehat{\rho}_j^{*}$'s, the screened submodel is defined as

$$\widehat{\mathcal{M}} = \{j : 1 \leqslant j \leqslant p : \widehat{\rho}_j^{*} \text{ is among the first } d \text{ largest}\}, \tag{2.2.35}$$

where $|\widehat{\mathcal{M}}| = d$ is taken to be smaller than the sample size n. Thus, the full model is reduced to a moderate scale. Liu et al. (2014) indicated we can set $d = [n^{4/5}/\log(n^{4/5})]$ for ultrahigh-dimensional varying coefficient models.

Denote the true model and its complement by $\mathcal{M}_*$ and $\mathcal{M}_*^c$. To establish the

ranking consistency property, some regularity conditions are needed:

1. The population level unconditioned-squared correlation cannot be too small.
2. u, $\mathbf{X}$ and ε are independent given $\mathbf{X}^T_{\mathcal{M}_*}\boldsymbol{\beta}_{\mathcal{M}_*}(u)$ and the linearity condition is satisfied:

$$E\{\mathbf{X}|X^T_{\mathcal{M}_*}\boldsymbol{\beta}_{\mathcal{M}_*}(u), u\} = \text{cov}(\mathbf{X}, \mathbf{X}^T_{\mathcal{M}_*}|u)\boldsymbol{\beta}_{\mathcal{M}_*}(u)\{\text{cov}(\mathbf{X}^T_{\mathcal{M}_*}|u)\}^{-1} \times \boldsymbol{\beta}^T_{\mathcal{M}_*}(u)\mathbf{X}_{\mathcal{M}_*}. \tag{2.2.36}$$

3. The density function of u has continuous second-order derivatives.
4. The kernel function $K(\cdot)$ is symmetric and uniformly bounded on its finite support.
5. X_j and Y satisfy the sub exponential tail probability uniformly in p.
6. All conditional means and their corresponding first and second derivatives are finite and the conditional variances are significantly greater than zero.

Under Conditions 1-6, Liu et al. (2014) proved the ranking consistency property:

$$\lim_{n\to\infty} \inf\{\min_{j\in\mathcal{M}_*} \widehat{\rho}^*_j \ - \max_{j\in\mathcal{M}^c_*} \widehat{\rho}^*_j\} > 0 \text{ in probability}, \tag{2.2.37}$$

which states all the true predictors have larger $\widehat{\rho}^*$'s than the unimportant ones.

Liu et al. (2014) also established the Sure Screening Property under conditions 3-6, that is $P(\mathcal{M}_* \subset \widehat{\mathcal{M}}) \to 1$ as $n \to \infty$.

CC-SIS is based on a marginal utility. Thus it may fail to identify the important variables which are marginally uncorrelated to the response but jointly correlated to the response. Liu et al. (2014) proposed an iterative CC-SIS that can overcome this weakness.

2.2.4 Joint Effects

The aforementioned feature screening procedures are based on the marginal utilities between response and predictors. They may fail to select the active variables which are marginally uncorrelated with the responses but jointly correlated with

the responses and some inactive variables may be selected if they are highly correlated with the active ones. The screening methods considering the joint effects between predictors overcome the aforementioned weakness. In this section, two feature selection procedures accounting for the joint effect of features are reviewed.

2.2.4.1 Sparse MLE

The marginal screening methods fail to select the important variables which are marginally independent of the response and remove the unimportant variables which are highly correlated with the important ones. Iterative SIS methods improve the performance of the marginal screening method, however, they have higher computational cost and increased complexity. Xu and Chen (2014) proposed a new method via the sparsity-restricted maximum likelihood estimator (SMLE) for the generalized linear models, which considers the joint effects of features in the screening process. The new method overcomes the weakness of marginal screening methods and also enjoys lower computational cost.

Consider a random sample of size n from a generalized linear model, the log-likelihood function of $\boldsymbol{\beta}$ is given by

$$\ell_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{(\mathbf{X}_i^T \boldsymbol{\beta}) Y_i - b(\mathbf{X}_i \boldsymbol{\beta})\} \tag{2.2.38}$$

under the canonical link. The SMLE is defined by

$$\widehat{\boldsymbol{\beta}}_{[k]} = \arg\max_{\boldsymbol{\beta}} \ell_n(\boldsymbol{\beta}) \text{ subject to } ||\boldsymbol{\beta}||_0 \leqslant k \tag{2.2.39}$$

where $||\cdot||_0$ denotes the number of nonzero components of a vector and k is larger than the cardinality of the true model $|\mathcal{M}_*|$. Let $\widehat{\mathcal{M}} = \{1 \leqslant j \leqslant p : \widehat{\boldsymbol{\beta}}_{[k]j} \neq 0\}$ be the nonzero components of $\widehat{\boldsymbol{\beta}}_{[k]}$. The SMLE method is a joint-likelihood-supported screening method that accounts for the joint effects between features.

In order to obtain the SMLE with a low computational cost, Xu and Chen (2014) developed an iterative hard-thresholding algorithm (IHT) to estimate the SMLE. For a given $\boldsymbol{\beta}$, the log-likelihood function $\ell_n(\boldsymbol{\beta})$ is approximated by

$$h_n(\boldsymbol{\gamma}; \boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T S_n(\boldsymbol{\beta}) - (u/2)||\boldsymbol{\gamma} - \boldsymbol{\beta}||_2^2 \tag{2.2.40}$$

for some scaling parameter $u > 0$, where $||\cdot||_2$ is the L_2 norm and $S_n(\boldsymbol{\beta})$ is the score function. The first two terms are given by Taylor's expansion and the third term is a regularization term. It can be seen that $h_n(\boldsymbol{\gamma};\boldsymbol{\beta})$ approximates $\ell_n(\boldsymbol{\beta})$ very well when $\boldsymbol{\gamma}$ is close to $\boldsymbol{\beta}$ and in fact $h_n(\boldsymbol{\beta};\boldsymbol{\beta}) = \ell_n(\boldsymbol{\beta})$.

In the iterative process, $\boldsymbol{\beta}^{(t)}$ is updated by

$$\boldsymbol{\beta}^{(t+1)} = \arg\max_{\boldsymbol{\gamma}} h_n(\boldsymbol{\gamma};\boldsymbol{\beta}^{(t)}) \text{ subject to } ||\boldsymbol{\gamma}||_0 \leqslant k. \tag{2.2.41}$$

The regularization term in (2.2.41) prevents $\boldsymbol{\beta}^{(t+1)}$ far away from $\boldsymbol{\beta}^{(t)}$ which makes $\boldsymbol{\beta}^{(t+1)}$ a nongreedy update for obtaining $\widehat{\boldsymbol{\beta}}_{[k]}$. Thus, the iteration is started with an initial $\boldsymbol{\beta}^{(0)}$ and stopped until $||\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}||_2$ falls below some tolerance level. The solution to (2.2.41) is

$$\boldsymbol{\beta}^{(t+1)} = \mathbf{H}(\boldsymbol{\beta}^{(t)} + u^{-1}\mathbf{X}^T\{\mathbf{y} - b^{'}(\mathbf{X}\boldsymbol{\beta}^{(t)})\}; k), \tag{2.2.42}$$

where $\mathbf{H}(\boldsymbol{\gamma}; k) = [H(\gamma_1; r), ..., H(\gamma_p; r)]$ and $H(\gamma; r) = \gamma I(|\gamma| > r)$ with r as the k-th largest component of $\boldsymbol{\gamma}$. Xu and Chen (2014) showed the increment property of SMLE method, so the sequence $\boldsymbol{\beta}^{(t+1)}$ based on IHT algorithm increases the value of $\ell_n(\cdot)$ and necessarily converges to a local maximum of $\ell_n(\cdot)$. The increment property guarantees that the SMLE method is a promising method for feature screening. Xu and Chen (2014) also proved the sure screening property of SMLE method under some technical conditions, which means that $p(\mathcal{M}_* \subset \widehat{\mathcal{M}}) \to 1$, as $n \to \infty$.

2.2.4.2 Sure Joint Screening

Yang et al. (2016) proposed a feature screening method based on the joint partial likelihood for the Cox's model:

$$h(t|\mathbf{X}) = h_0(t)\exp(\mathbf{X}^T\boldsymbol{\beta}), \tag{2.2.43}$$

where $h_0(t)$ is an unspecified baseline hazard function and T is the survival time. Denote the observed time by $Z = \min(T, C)$ and the event indicator by $\delta = I(T < C)$, where C is the censoring time. Suppose that $\{(\mathbf{X}_i, Z_i, \delta_i)\}$ is an IID random

sample from model (2.2.43) and $t_1^0 < \cdots < t_N^0$ are the ordered failure times. Denote the risk set right before the time t_j^0 by R_j:

$$R_j = \{i : Z_i \geqslant t_j^0\}. \tag{2.2.44}$$

Let N failures at time $t_1^0 < \cdots < t_N^0$ be $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(N)}$, the partial likelihood function of the random sample is

$$\ell_p(\boldsymbol{\beta}) = \sum_{j=1}^{N} [\mathbf{X}_{(j)}^T \boldsymbol{\beta} - \log\{\sum_{i \in R_j} \exp(\mathbf{X}_i^T \boldsymbol{\beta})\}]. \tag{2.2.45}$$

Denote the true model by $\mathcal{M}_*$. From here, it is for survival data. Yang et al. (2016) proposed a screening method for the Cox model by maximize the constrained partial likelihood

$$\widehat{\boldsymbol{\beta}}_m = \arg\max_{\boldsymbol{\beta}} \ell_p(\boldsymbol{\beta}) \text{ subject to } ||\boldsymbol{\beta}_m||_0 \leqslant m, \tag{2.2.46}$$

where $m > |\mathcal{M}_*|$. Since it is impossible to maximize the constrained partial likelihood in the high dimensional setting, Yang et al. (2016) considered a proxy of the partial likelihood function

$$\ell_p(\boldsymbol{\gamma}) \approx \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p'(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p''(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}), \tag{2.2.47}$$

where $\ell_p'(\boldsymbol{\beta}) = \partial \ell_p(\boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$ and $\ell_p''(\boldsymbol{\beta}) = \partial^2 \ell_p(\boldsymbol{\gamma}) / \partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$. For the setting of large p and small n, $\ell_p''(\boldsymbol{\beta})$ is not invertible. Thus, the authors proposed to use the following approximation for $\ell_p''(\boldsymbol{\gamma})$

$$g(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell_p(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell_p'(\boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T W (\boldsymbol{\gamma} - \boldsymbol{\beta}), \tag{2.2.48}$$

where u is a scaling constant and $W = \text{diag}\{-\ell_p''(\boldsymbol{\beta})\}$. The authors approximated $\ell_p''(\boldsymbol{\beta})$ by $u\text{diag}\{\ell_p''(\boldsymbol{\beta})\}$.

It can be seen that $g(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is an additive function of γ_j for any given β since W

is a diagonal matrix. Thus, the maximizer of the following maximization problem

$$\max_{\boldsymbol{\gamma}} g(\boldsymbol{\gamma}|\boldsymbol{\beta}) \text{ subject to } ||\boldsymbol{\gamma}||_0 \leqslant m \tag{2.2.49}$$

is $\widehat{\gamma}_j = \tilde{\gamma}_j I\{|\tilde{\gamma}_j| > |\tilde{\gamma}_{(m+1)}|\} := H(\tilde{\gamma}_j; m)$, where $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\beta} + u^{-1}W^{-1}\ell_p'(\boldsymbol{\beta})$ and $|\tilde{\gamma}_{(m+1)}|$ is the $m+1$-th largest among $\{\tilde{\gamma}_1, \cdots, \tilde{\gamma}_p\}$.

The algorithm for the feature screening procedure is:

1. Set the initial $\boldsymbol{\beta}^{(0)} = \mathbf{0}$.

2. Set $t = 0, 1, 2, \cdots$ and iteratively conduct 2a and 2b until some convergence criterion is satisfied.

Step 2a. Compute $\tilde{\boldsymbol{\gamma}}^{(t)} = (\tilde{\gamma}_1^{(t)}, \cdots, \tilde{\gamma}_p^{(t)})^T = \boldsymbol{\beta}^{(t)} + u_t^{-1}W^{-1}(\boldsymbol{\beta}^{(t)})\ell_p'(\boldsymbol{\beta}^{(t)})$ and $\tilde{\boldsymbol{\beta}}^{(t)} = (H(\tilde{\gamma}_1^{(t)}; m), \cdots, H(\tilde{\gamma}_p^{(t)}; m)) := \mathbf{H}(\tilde{\boldsymbol{\gamma}}^{(\mathbf{t})}; \mathbf{m})$.

Set $S_t = \{j : \tilde{\beta}_j \neq 0\}$, the nonzero elements of $\tilde{\boldsymbol{\beta}}^{(t)}$.

Step 2b. Update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(t+1)}$ as follows. If $j \notin S_t$, set $\beta_j^{(t+1)} = 0$; otherwise, set $\{\beta_j^{(t+1)} : j \in S_t\}$ be the maximum likelihood estimate of the submodel S_t.

The proposed method can select the important variables that are marginally independent but jointly dependent of the survival time and is considered to perform better than the marginal screening method. In addition, it can be carried out with low computational cost. Yang et al. (2016) also demonstrated the increment property and the sure screening property of the proposed method under some certain conditions.

2.3 Review of Two Sample Mean Testing in High-dimensional setting

In this section, we focus on two-sample mean testing methods. These methods can be directly applied to the one-sample mean testing problems. Suppose that $\mathbf{x}_1 = \{\mathbf{x}_{11}, \cdots, \mathbf{x}_{1N_1}\}$ and $\mathbf{x}_2 = \{\mathbf{x}_{21}, \cdots, \mathbf{x}_{2N_2}\}$ are two independent p-dimensional random samples with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices Σ_1 and Σ_2, respectively. We assume that $\Sigma_1 = \Sigma_2 = \Sigma$ and the two sample mean testing problem is

to test the null hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \text{ vs. } H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{2.3.50}$$

All the testing methods reviewed in this section can be extended to the one-sample testing problem.

2.3.1 Classical Hotelling's T^2 Test

The classical Hotelling's T^2 test is used in the two sample mean testing problems when $n = (N_1 + N_2 - 2) > p$ and $\mathbf{x}_{ij} \sim N_p(\boldsymbol{\mu}_i, \Sigma), i = 1, 2$. The test statistic is defined by

$$T^2 = \frac{N_1 N_2}{N_1 + N_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{2.3.51}$$

where $\bar{\mathbf{x}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}, i = 1, 2$, and $S = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{N_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T$.

Under the null hypothesis H_0,

$$\frac{n - p + 1}{np} T^2 \sim F_{p,n+1-p}, \tag{2.3.52}$$

hence we reject the null hypothesis when

$$T^2 > F_{p,n+1-p}(\alpha), \tag{2.3.53}$$

where $F_{p,n+1-p}(\alpha)$ is the $1 - \alpha$ quantile of the distribution $F_{p,n+1-p}$.

Bai and Sranadasa (1996) derived the approximation of the power function of Hotelling's T^2 test for the two sample problem.

Theorem 2.3.1. *If $y_n = p/n \to y \in (0, 1)$, $N_1/(N_1 + N_2) \to \kappa \in (0, 1)$ and $||\delta|| = o(1)$, then*

$$\beta_H(\delta) - \Phi(-\xi_\alpha + \sqrt{\tfrac{n(1-y)}{2y}}\kappa(1 - \kappa)||\delta||^2) \to 0,$$

where $\delta = \Sigma^{-1/2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\beta_H(\delta)$ is the power function of Hotelling's T^2 test.

Denote $n||\delta||^2 \to a > 0$, and Theorem 5 shows that the limiting power of Hotelling's T^2 test is slowly increasing for y close to 1 as a increases. However, the Hotelling's T^2 test cannot be used when $p > n$, since the matrix S is not invertible.

2.3.2 Dempster's Test

Under the normality assumption $\mathbf{x}_i = \{\mathbf{x}_{ij} : \mathbf{x}_{ij} \sim N_p(\boldsymbol{\mu}_i, \Sigma), i = 1, 2, j = 1, \cdots, N_i\}$ and $\mathbf{X}^T = (\mathbf{x}_{11}, \cdots, \mathbf{x}_{1N_1}, \mathbf{x}_{21}, \cdots, \mathbf{x}_{2N_2})$ and $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent, Dempster (1958; 1960) proposed a non-exact test for the two-sample mean testing problems with $p > n$. Consider a matrix with $N_1 + N_2$ orthogonal columns $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_{N_1+N_2})$ in a Euclidean space, where

$$\begin{aligned} \mathbf{b}_1 &= \sqrt{\frac{1}{N_1+N_2}}\mathbf{1}_{N_1+N_2}, \\ \mathbf{b}_2 &= (\sqrt{\frac{N_2}{N_1(N_1+N_2)}}\mathbf{1}_{N_1}, -\sqrt{\frac{N_1}{N_1(N_1+N_2)}}\mathbf{1}_{N_2}) \end{aligned} \tag{2.3.54}$$

and $\mathbf{b}_i, i = 3, \cdots, N_1 + N_2$ are chosen to ensure the orthogonality of $\mathbf{B}$. Denote $\mathbf{Y} = \mathbf{B}^T\mathbf{X} = (\mathbf{y}_1, \cdots, \mathbf{y}_{N_1+N_2})^T$. It is seen that $E(\mathbf{y}_1) = (N_1\boldsymbol{\mu}_1 + N_2\boldsymbol{\mu}_2)/(N_1+N_2)$, $E(\mathbf{y}_2) = \sqrt{(N_1N_2)/(N_1+N_2)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\{\mathbf{y}_i, i = 3, \cdots, N_1+N_2\}$ are distributed as $N(\mathbf{0}, \Sigma)$. Thus, $\mathbf{y}_2$ is the difference between the sample means, and $\mathbf{y}_2 \sim N(\mathbf{0}, \Sigma)$ under H_0.

Dempster (1958; 1960) defined the nonexact test statistic by

$$F = \frac{Q_2}{\sum_{i=1}^{N_1+N_2} Q_i/n}, \tag{2.3.55}$$

where $Q_i = \mathbf{y}_i^T\mathbf{y}_i$. Dempster (1958; 1960) assumed that Q_i is distributed as $m\chi_r^2$ and m and r can be solved by the method of moments, the distribution of F is $F_{r,nr}$.

Dempster (1960) provided two methods for estimating r as follows:

1. Assume that $Q_i \sim m\chi_r^2$, denote a sufficient statistic

$$t = n[\log(\frac{1}{n}\sum_{i=3}^{N_1+N_2} Q_i)] - \sum_{i=3}^{N_1+N_2} Q_i \tag{2.3.56}$$

that depends only on r and when r is small, its distribution can be approximated by

$$t \sim [\frac{1}{r} + \frac{1+1/n}{3r^2}]\chi^2_{n-1}. \tag{2.3.57}$$

Thus, the first estimate of r is defined by the solution of the following equation

$$t = [\frac{1}{\widehat{r}_1} + \frac{1+1/n}{3\widehat{r}_1^2}](n-1). \tag{2.3.58}$$

2. The second estimate of r is constructed by using the angles between $\mathbf{y}_i$ and $\mathbf{y}_j$, which is denoted by θ_{ij} with $3 \leqslant i < j \leqslant (N_1 + N_2)$. Let

$$u_{ij} = -\log(\sin^2(\theta_{ij})) \sim (\frac{1}{r} + \frac{3}{2r^2})\chi^2_1, \tag{2.3.59}$$

and the second estimate of r can be obtained by solving

$$t + \sum_{3\leqslant i<j\leqslant(N_1+N_2)} u_{ij} = (\frac{1}{\widehat{r}_2} + \frac{3}{2\widehat{r}_2^2}) \begin{pmatrix} n \\ 2 \end{pmatrix} + [\frac{1}{\widehat{r}_2} + \frac{1+1/n}{3\widehat{r}_2^2}](n-1) \tag{2.3.60}$$

Bai and Saranadasa (1996) suggested that

$$r = \frac{(\mathrm{tr}\Sigma)^2}{\mathrm{tr}\Sigma^2} \text{ and } m = \frac{\mathrm{tr}\Sigma^2}{\mathrm{tr}\Sigma}, \tag{2.3.61}$$

and if $nS \sim W_p(\Sigma, n)$,

$$\frac{n^2}{(n+2)(n-1)}[\mathrm{tr}S^2 - (\mathrm{tr}S)^2/n] \tag{2.3.62}$$

is an unbiased and ratio-consistent estimator of $\mathrm{tr}\Sigma^2$. Thus, we can obtain another estimate of r

$$\widehat{r} = \frac{(\mathrm{tr}S)^2}{\frac{n^2}{(n+2)(n-1)}[\mathrm{tr}S^2 - (\mathrm{tr}S)^2/n]}. \tag{2.3.63}$$

Bai and Saranadasa (1996) also gave the approximation of the power function of

Dempster's test.

Theorem 2.3.2. *If $y_n = p/n \to y \in (0,1)$, $N_1/(N_1+N_2) \to \kappa \in (0,1)$ and r is known, when $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\Sigma(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = o((1/N_1+1/N_2)tr\Sigma^2)$ and $\lambda_{\max} = o(\sqrt{tr\Sigma^2})$, we have*

$$\beta_{T_D}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) - \Phi(-\xi_\alpha + \frac{\kappa(1-\kappa)||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||^2}{\sqrt{2tr(\Sigma^2)}}) \to 0,$$

where $\lambda_{\max}$ is the maximum eigenvalue of Σ and $\beta_{T_D}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ is the power function of the Dempster's test.

Theorem 6 reveals that if y is close to 1, the asymptotic power of the Dempster's test increases much faster than that of the Hotelling's Test as the non-central parameter increases.

2.3.3 Testing Methods Using Diagonal Estimators for Σ

In this section, we review some testing methods using diagonal estimators of covariance matrix Σ. These methods have been shown to be more powerful than the classical Hotelling's T^2 test when the dimension is close to the sample size. In addition, they can also be used for the high-dimensional data.

2.3.3.1 Bai-Saranadasa Test (BS test)

The Hotelling's T^2 test and Dempster's test depend on the normality assumption and Dempster's test involves a complicated estimation of r. Bai and Saranadasa (1996) proposed a new test for H_0 without the normality assumption to simplify the testing procedure. The new test is based on the following assumptions:

1. $\mathbf{x}_{ij} = \Gamma\mathbf{z}_{ij}$, $i = 1,2$, $j = 1,\cdots,N_i$, where Γ is a $p\times m$ matrix satisfying $\gamma\Gamma^T = \Sigma$, and $\mathbf{z}_{ij}$ are IID random m-vectors satisfying $E(\mathbf{z}_{ij}) = \mathbf{0}$, $\mathrm{Var}(\mathbf{z}_{ij}) = I_m$, $E(z_{ijk}) < \infty$, $k = 1,\cdots,m$, $E(z_{ijk}^4) = 3+\Delta < \infty$ and whenever $\sum_{k=1}^{m}\nu_k = 4$, $E\prod_{k=1}^{m} z_{ijk}^{\nu_k}$ equals 0 when at least one ν_k equals 1 and equals 1 when there are two ν_k's equal to 2.

2. $y_n = p/n \to y \in (0,1)$, $N_1/(N_1+N_2) \to \kappa \in (0,1)$.

3. $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o((1/N_1 + 1/N_2)\mathrm{tr}\Sigma^2)$ and $\lambda_{\max} = o(\sqrt{\mathrm{tr}\Sigma^2})$.

Bai and Saranadasa (1996) first constructed a statistic

$$T_n = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - (\frac{1}{N_1} + \frac{1}{N_2})\mathrm{tr}S. \tag{2.3.64}$$

Under the null hypothesis H_0, $E(T_n) = 0$ and $\mathrm{Var}(T_n) = \sigma^2_{T_n} = 2(\frac{1}{N_1} + \frac{1}{N_2})^2(1 + \frac{1}{n})\mathrm{tr}\Sigma^2$ under the normality assumption and $\mathrm{Var}(T_n) = \sigma^2_{T_n}(1 + o(1))$ without the normality assumption but assumptions 1-3 are satisfied.

The authors proved that under Assumption 1-3,

$$\frac{T_n}{\sqrt{\mathrm{Var}T_n}} \sim N(0, 1), \text{ as } n \to \infty. \tag{2.3.65}$$

They also showed that (2.3.62) is an unbiased and ratio-consistent estimator of $\mathrm{tr}\Sigma^2$. Hence, the test statistic is defined by

$$T_{BS} = \frac{T_n}{\sqrt{\mathrm{Var}T_n}} = \frac{(\frac{1}{N_1} + \frac{1}{N_2})(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \mathrm{tr}S}{\sqrt{\frac{2(n+1)n}{(n+2)(n-1)}[\mathrm{tr}S^2 - (\mathrm{tr}S)^2/n]}}, \tag{2.3.66}$$

and $T_{BS} \sim N(0, 1)$.

Bai and Saranadasa (1996) derived the asymptotic power of their test,

$$\beta_{T_{BS}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \Phi(-\xi_\alpha + \frac{\kappa(1-\kappa)||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2}{\sqrt{2\mathrm{tr}\Sigma^2}}) \to 0, \tag{2.3.67}$$

under the assumptions 1-3, which is same as the asymptotic power of Dempster's test.

BS test is also more powerful than Hotelling's T^2 test when y is close to 1, and it simplifies the Dempster's test by avoiding estimating r. BS test is sightly more powerful than Dempster's test because an error may be caused by the estimation of r in Dempster's test.

2.3.3.2 Chen and Qin Test(CQ test)

BS test reveals the restriction on n, p and the largest eigenvalue λ_p of Σ are needed to control the terms $\sum_{j=1}^{N_i} \mathbf{x}_{ij}^T\mathbf{x}_{ij}, i = 1, 2$ in $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$. Chen and Qin (2010)

proposed a test (CQ test) that avoid the effect of $\sum_{j=1}^{N_i} \mathbf{x}_{ij}^T \mathbf{x}_{ij}, i = 1, 2$. The authors assumed the following factor-like model structure :

$$\mathbf{x}_{ij} = \Gamma_i \mathbf{z}_{ij}, i = 1, 2, j = 1, \cdots, N_i \tag{2.3.68}$$

where each Γ_i is a $p \times m$ matrix for some $m > p$ such that $\Gamma_i \Gamma_i^T = \Sigma_i$, and $\mathbf{z}_{ij}$ are IID random m-vectors satisfying $E(\mathbf{z}_{ij}) = \mathbf{0}$, $\mathrm{Var}(\mathbf{z}_{ij}) = I_m$. In addition, it is required that $E(z_{ijk}^4) = 3 + \Delta < \infty$ and

$$E(z_{ijl_1}^{\alpha_1} z_{ijl_2}^{\alpha_2} \cdots z_{ijl_q}^{\alpha_q}) = E(z_{ijl_1}^{\alpha_1}) E(z_{ijl_2}^{\alpha_2}) \cdots E(z_{ijl_q}^{\alpha_q}) \tag{2.3.69}$$

for some $q > 0$ such that $\sum_{l=1}^{q} \alpha_l \leqslant 8$ and $l_1 \neq l_2 \neq \cdots \neq l_q$. We can see that CQ test does not require $\Sigma_1 = \Sigma_2$.

Chen and Qin (2010) proposed to use a test statistic

$$T_{n_{CQ}} = \frac{\sum_{i \neq j}^{N_1} \mathbf{x}_{1i}^T \mathbf{x}_{1j}}{N_1(N_1 - 1)} + \frac{\sum_{i \neq j}^{N_2} \mathbf{x}_{2i}^T \mathbf{x}_{2j}}{N_2(N_2 - 1)} - 2\frac{\sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \mathbf{x}_{1i}^T \mathbf{x}_{2j}}{N_1 N_2}, \tag{2.3.70}$$

with $E(T_{n_{CQ}}) = 0$, and

$$\mathrm{Var}(T_{n_{CQ}}) \to \frac{2}{N_1(N_1 - 1)} \mathrm{tr}(\Sigma_1^2) + \frac{2}{N_2(N_2 - 1)} \mathrm{tr}(\Sigma_2^2) + \frac{4}{N_1 N_2} \mathrm{tr}(\Sigma_1 \Sigma_2) \tag{2.3.71}$$

under H_0. If the following conditions

$$\begin{aligned} &N_1/(N_1 + N_2)\kappa \in (0, 1), \text{ as } n \to \infty, \\ &(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o[n^{-1} \mathrm{tr}(\Sigma_1 + \Sigma_2)^2] \text{ for } i = 1, 2, \\ &\mathrm{tr}(\Sigma_i \Sigma_j \Sigma_k \Sigma_l) = o[\mathrm{tr}^2((\Sigma_1 + \Sigma_2)^2)] \end{aligned} \tag{2.3.72}$$

are satisfied, then

$$T_{CQ} = \frac{T_{n_{CQ}}}{\sqrt{\mathrm{Var}(T_{n_{CQ}})}} \to N(0, 1) \tag{2.3.73}$$

under H_0. The power function under condition $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma_i (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o[n^{-1} \mathrm{tr}(\Sigma_1 +$

$\Sigma_2)^2]$ is

$$\beta_{CQ}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \Phi(-\xi_\alpha + \frac{n\kappa(1-\kappa)||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2}{\sqrt{2\text{tr}(\tilde{\Sigma}_\kappa^2)}}), \tag{2.3.74}$$

where $\tilde{\Sigma}_\kappa = (1-\kappa)\Sigma_1 + \kappa\Sigma_2$.

Chen and Qin (2010) also proposed the ratio-consistent estimators for $\text{tr}(\Sigma_i^2)$ and $\text{tr}(\Sigma_1\Sigma_2)$:

$$\widehat{\text{tr}(\Sigma_i^2)} = [N_i(N_i-1)]^{-1}\text{tr}\{\sum_{j\neq k}^{N_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i(j,k)})\mathbf{x}_{ij}^T(\mathbf{x}_{ik} - \bar{\mathbf{x}}_{i(j,k)})\mathbf{x}_{ik}^T\} \tag{2.3.75}$$

and

$$\widehat{\text{tr}(\Sigma_1\Sigma_2)} = (N_1N_2)^{-1}\text{tr}\{\sum_{l=1}^{N_1}\sum_{k=1}^{N_2}(\mathbf{x}_{1l} - \bar{\mathbf{x}}_{1(l)})\mathbf{x}_{1l}^T(\mathbf{x}_{2k} - \bar{\mathbf{x}}_{2(k)})\mathbf{x}_{2k}^T\}, \tag{2.3.76}$$

where $\bar{\mathbf{x}}_{i(j,k)}$ is the ith sample mean after excluding $\mathbf{x}_{ij}$ and $\mathbf{x}_{ik}$ and $\mathbf{x}_{i(l)}$ is the ith sample mean without $\mathbf{x}_{il}$.

2.3.3.3 Srivastava and Du Test (SD test)

Srivastava and Du (2008) proposed another test for two-sample mean problems under the normality assumption: $\mathbf{x}_i = \{\mathbf{x}_{ij} : \mathbf{x}_{ij} \sim N_p(\boldsymbol{\mu}_i, \Sigma), i = 1, 2, j = 1, \cdots, N_i\}$, and $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent.

Define the diagonal matrix of $\Sigma = (\sigma_{ij})$ and sample covariances by

$$D_\sigma = diag(\sigma_{11}, \cdots, \sigma_{pp}) \tag{2.3.77}$$

$$D_S = diag(s_{11}, \cdots, s_{pp}), \tag{2.3.78}$$

where $\{s_{ii}, i = 1, \cdots, p\}$ are the diagonal elements of sample covariance matrix S. Then the sample correlation matrix is defined by

$$R = D_S^{-\frac{1}{2}} S D_S^{-\frac{1}{2}}. \tag{2.3.79}$$

and denote $\lambda_{ip}, i = 1, \cdots, p$ are the eigen values of sample correlation matrix R.

Srivastava and Du (2008) constructed a test statistic

$$T_{SD} = \frac{\frac{N_1N_2}{N_1+N_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T D_S^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \frac{np}{n-2}}{[2(\mathrm{tr}R^2 - \frac{p^2}{n})c_{p,n}]^{\frac{1}{2}}}, \tag{2.3.80}$$

where $c_{p,n} \to 1$ in probability as $(n,p) \to \infty$, and

$$c_{p,n} = 1 + \frac{\mathrm{tr}R^2}{p^{3/2}}. \tag{2.3.81}$$

The authors showed that $T_{SD} \sim N(0,1)$ and derived the asymptotic function of SD test

$$\beta_{T_{SD}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \to \Phi(-\xi_\alpha + \frac{N_1N_2}{N_1+N_2}\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T D_\sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\sqrt{2\mathrm{tr}R^2}}), \text{ as } n,p \to \infty, \tag{2.3.82}$$

when

$$\begin{aligned} &n = O(p^\zeta) \text{ with } \frac{1}{2} < \zeta \leqslant 1 \text{ and } N_1/(N_1+N_2) \to \kappa \in (0,1); \\ &0 < \lim_{p\to\infty} \frac{\mathrm{tr}R^i}{p} < \infty, i = 1,...,4 \text{ and } \lim_{p\to\infty} \max_{1\leqslant i\leqslant p} \frac{\lambda_{ip}}{\sqrt{p}} = 0; \\ &(\frac{N_1+N_2}{nN_1N_2})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \leqslant M, \text{ M does not depend on } p. \end{aligned} \tag{2.3.83}$$

Srivastava and Du (2008) also showed that under the conditions (2.3.83), $\beta_{T_{SD}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geqslant \beta_{T_{BS}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \simeq \beta_{T_D}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ with strictly inequality unless $\sigma_{11} = \cdots = \sigma_{pp}$.

2.3.4 Projection Methods

The test methods using diagonal estimators of the covariance matrix Σ may lose power because of the limited use of the covariance structure. We can use the covariance structure more effectively by projecting the high-dimensional data to a lower dimensional space and then use the classical Hotelling's T^2 test. We give a brief overview of some projection testing methods in this part.

2.3.4.1 Lopes, Jacob and Wainwright Test (LJW test)

The classical Hotelling's T^2 test is not well defined when $p > n$ since S is not invertible. BS, SD and CQ test formed estimates of Σ by some diagonal matrices which are easily invertible. However, the limited use of covariance structure sacrifices power when non-trivial correlation exists. Lopes et al. (2011a) proposed a testing method with projected samples in a space of lower dimension that utilizes the covariance structure more effectively.

LJW test is processed under the normality assumptions with $\Sigma_1 = \Sigma_2$ when $p \geqslant n/2$. Let P_k^T be a $k \times p$ projection matrix with IID $N(0,1)$ entries, where k is suggested to take $[n/2]$. The projected samples $\{P_k^T \mathbf{x}_{11}, \cdots, P_k^T \mathbf{x}_{1N_1}\}$ and $\{P_k^T \mathbf{x}_{21}, \cdots, P_k^T \mathbf{x}_{2N_2}\}$ are IID $N(P_k^T \boldsymbol{\mu}_i, P_k^T \Sigma P_k)$ respectively, with $i = 1, 2$. The Hotelling's T^2 test can be processed to the two projected samples since $n > k$:

$$H_{p0} : P_k^T \boldsymbol{\mu}_1 = P_k^T \boldsymbol{\mu}_2 \text{ vs. } H_{p1} : P_k^T \boldsymbol{\mu}_1 \neq P_k^T \boldsymbol{\mu}_2 \tag{2.3.84}$$

The corresponding Hotelling's T^2 test statistic

$$T_{Hp}^2 = \frac{N_1 N_2}{N_1 + N_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T P_k (P_k^T S P_k)^{-1} P_k^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{2.3.85}$$

and $\frac{n-k+1}{kn} T_{Hp}^2 \sim F_{k,n-k+1}$. The formula of T_{Hp}^2 shows that its distribution is the same under both H_{p0} and H_0, so we can reject H_0 if $T_{Hp}^2 > \frac{kn}{n-k+1} F_{k,n-k+1}(\alpha)$, where $F_{k,n-k+1}(\alpha)$ is the $1-\alpha$ quantile of $F_{k,n-k+1}$.

Suppose that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o(1)$ and $\frac{N_1}{N_1+N_2} \to \kappa \in (0,1)$, the authors showed that under all sequences of projections P_k^T, the asymptotic power function of LJW test

$$\beta((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \Sigma, P_k^T) - \Phi(-\xi_\alpha + \frac{\kappa(1-\kappa)}{\sqrt{2}} \sqrt{n} \Delta_k^2) \to 0 \text{ as } n \to \infty, \tag{2.3.86}$$

where $\Delta_k^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T P_k (P_k^T S P_k)^{-1} P_k^T (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

Lopes et al. (2011b) refined their first projection test method by computing the average of the matrix $P_k (P_k^T S P_k)^{-1} P_k^T$ over the ensemble P_k. The test statistic

is defined by

$$\bar{T}^2_{Hp} = \frac{N_1 N_2}{N_1 + N_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T E_{P_K}[P_k(P_k^T S P_k)^{-1} P_k^T](\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \tag{2.3.87}$$

For choosing the number of independent copies of P_k, Lopes et al. (2011b) suggested $\bar{T}^2_{Hp}$ statistic stabilizes after averaging 30 projections.

Let $y_n = k/n$, where $k \in \{1, \cdots, \min(n, p)\}$ and the authors suggested $k = [n/2]$, $\bar{\mu}_n = \frac{y_n}{1-y_n} n$ and $\bar{\sigma}_n = \sqrt{\frac{2y_n}{(1-y_n)^3}}\sqrt{n}$, the authors proved that if $y_n = y + o(\frac{1}{\sqrt{n}})$ for a constant $y \in (0, 1)$,

$$\frac{\bar{T}^2_{Hp} - \bar{\mu}_n}{\bar{\sigma}_n} \to N(0, 1) \text{ as } (n, p) \to \infty, \tag{2.3.88}$$

under H_0. If the two conditions $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = o(1)$ and $\frac{N_1}{N_1+N_2} \to \kappa \in (0, 1)$ are also satisfied, as $(n, p) \to \infty$, the asymptotic power function of the test is

$$\beta_{Hp}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \Sigma) = \Phi(-\xi_\alpha + \kappa(1-\kappa)\sqrt{\frac{1-y}{2y}}\sqrt{n}\bar{\Delta}_k) + o(1), \tag{2.3.89}$$

where $\bar{\Delta}_k = E_{P_k}[\Delta_k]$.

2.3.4.2 Optimal Direction (OD test)

Li et al. (2015) proposed another projection test and also derived the optimal projection direction with the best power under the alternatives. Suppose that $\mathbf{x}_{ij}, i = 1, 2$ and $j = 1, \cdots, N_i$ is a random sample from population $\mathbf{x}_i$ with mean $\boldsymbol{\mu}_i$ and covariance Σ. Let A be a $p \times k$ nonzero constant projection matrix with $k \ll p$, and $A^T \bar{\mathbf{x}}_i \to N(A^T \boldsymbol{\mu}_i, A^T \Sigma A)$ in distribution and $A^T S A - A^T \Sigma A \to \mathbf{0}$ in probability. Define the projection Hotelling's T^2 test to be

$$\bar{T}^2_A = \frac{N_1 N_2}{N_1 + N_2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T A(A^T S A)^{-1} A^T (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \tag{2.3.90}$$

which is a two sample Hotelling's T^2 test based on $\mathbf{y}_{ij} = A^T \mathbf{x}_{ij}$. Thus, we have

$$\frac{N_1 + N_2 - k - 1}{kn} \bar{T}^2_A \sim F_{k, N_1+N_2-k-1} \tag{2.3.91}$$

under H_0. The authors proved that the projection test $\bar{T}_A^2$ reaches its best power at $k = 1$ and $A = a = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

In order to construct $\bar{T}_a^2$ in practice, a sample-splitting strategy is used. The random sample $\mathbf{x}_{ij}, i = 1, 2$ and $j = 1, \cdots, N_i$ is partitioned into two separate sets: $\mathcal{S}_{i1} = \mathbf{x}_{i1}, \cdots, \mathbf{x}_{iN_{i1}}$ and $\mathcal{S}_{i2} = \mathbf{x}_{i1}, \cdots, \mathbf{x}_{iN_{i2}}$, where $N_{i1} + N_{i2} = N_i$. $\mathcal{S}_{i1}, i = 1, 2$ is used to estimate a and $\mathcal{S}_{i2}$ is used to construct $\bar{T}_a^2$. Let $\bar{\mathbf{x}}_{11} - \bar{\mathbf{x}}_{21}$ and S_1 be the sample mean difference and pooled sample covariance matrix obtained from $\mathcal{S}_{i1}, i = 1, 2$, respectively. Since S_1 is not invertible when $p > n$, the authors estimate a by $\widehat{a} = (S_1 + \lambda D)^{-1}(\bar{\mathbf{x}}_{11} - \bar{\mathbf{x}}_{21})$, where $D = \text{diag}(S_1)$ and λ is a parameter. Thus, the projection test with the optimal direction is

$$\bar{T}_{\widehat{a}}^2 = \frac{N_{12}N_{22}}{N_{12} + N_{22}}(\bar{\mathbf{x}}_{12} - \bar{\mathbf{x}}_{22})^T \widehat{a}(\widehat{a}^T S_2 \widehat{a})^{-1}\widehat{a}^T(\bar{\mathbf{x}}_{12} - \bar{\mathbf{x}}_{22}), \tag{2.3.92}$$

where $\bar{\mathbf{x}}_{11} - \bar{\mathbf{x}}_{21}$ and S_1 are the sample mean difference and pooled sample covariance matrix obtained from $\mathcal{S}_{i2}, i = 1, 2$. Since $\widehat{a}$ is independent of $\mathcal{S}_{i2}$ and $\bar{T}_{\widehat{a}}^2$ follows a central $F_{1,N_{12}+N_{22}-1}$ distribution, $\bar{T}_{\widehat{a}}^2$ is equivalent to an exact t-test based on $A^T\mathbf{x}_{ij}$, where $\mathbf{x}_{ij} \in \mathcal{S}_{i2}, i = 1, 2$. Li et al. (2015) suggested that we may choose $N_{i2} = 0.6N_i$ and $\lambda = (N_{11} + N_{21})^{-\tau}$ with $\tau = 0.5$ in practice.

The authors also demonstrated that under the local alternative:

$$H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}\sqrt{\frac{1}{N_1} + \frac{1}{N_2}} \tag{2.3.93}$$

where $\eta = \boldsymbol{\delta}^T\Sigma^{-1}\boldsymbol{\delta}$, the asymptotic power of the OD test is no less than those of BS, DS and CQ test under certain conditions.

Chapter 3

Group Feature Selection in Ultrahigh Dimensional Generalized Varying-coefficient Linear Models

3.1 Background

Let Y be the response variable and $\{\mathbf{x}, U\}$ its associated covariates, where $\mathbf{x} = (X_1, \cdots, X_p)$ and U be p-dimensional and univariate covariates respectively. Further, let $\mu(\mathbf{x}, U) = E(Y|\mathbf{x}, U)$. The GVCM assumes that

$$\eta(\mathbf{x}, U) \widehat{=} g\{\mu(\mathbf{x}, U)\} = \mathbf{x}^T \boldsymbol{\alpha}(U), \tag{3.1.1}$$

where $g(\cdot)$ is a known link function and $\boldsymbol{\alpha}(\cdot)$ is a vector consisting of unspecified smooth regression coefficient functions. Here it is assumed that all $\alpha_j(\cdot)$'s are nonparametric functions and the support of U is finite and denoted by $[a, b]$.

Suppose that $\{U_i, \mathbf{x}_i, Y_i\}$, $i = 1, \ldots, n$, constitute an independent and identically distributed sample and that conditionally on $\{U_i, \mathbf{x}_i\}$, the conditional quasi-likelihood of Y_i is $Q\{\mu(U_i, \mathbf{x}_i), Y_i\}$, where the quasi-likelihood function is defined by $Q(\mu, y) = \int_\mu^y \frac{s-y}{V(s)} ds$, or equivalently $\frac{\partial Q(\mu, y)}{\partial \mu} = \frac{y-\mu}{V(\mu)}$, for a specific variance function $V(s)$. Denote by $\ell\{\boldsymbol{\alpha}(\cdot)\}$ the quasi-likelihood (McCullagh and Nelder, 1989) of the

collected data $\{(U_i, \mathbf{x}_i, Y_i), i = 1, \ldots, n\}$. That is

$$\ell\{\boldsymbol{\alpha}(\cdot)\} = \sum_{i=1}^{n} Q[g^{-1}\{\mathbf{x}_i^T \boldsymbol{\alpha}(U_i)\}; Y_i]. \tag{3.1.2}$$

To estimate the nonparametric regression coefficient, we use B-spline regression method. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $l \geqslant 1$ and $\{\psi_{jk}, k = 1, \ldots, d_{n_j}\}$ denote a normalized B-spline basis with $\|\psi_{jk}\|_\infty \leqslant 1$ and $d_{nj} = O(n^{1/5})$, where $\|\cdot\|_\infty$ is the sup norm. For any $\alpha_{nj} \in \mathcal{S}_n$, we have

$$\alpha_{nj}(U) = \sum_{k=1}^{d_{n_j}} \beta_{jk}\psi_{jk}(U) = \boldsymbol{\beta}_j^T \boldsymbol{\psi}_j(U), \quad j = 1, \cdots, p, \tag{3.1.3}$$

for some coefficients $\{\beta_{jk}\}_{k=1}^{d_{n_j}}$. Here d_{n_j} increases with n. We allow d_{n_j} to be different for different j since different coefficient functions may have different smoothness. Under some conditions, each nonparametric coefficient function $\alpha_j(U), j = 1, \cdots, p$ can be well approximated by functions in $\mathcal{S}_n$.

Substituting (3.1.3) into (4.1.4), the maximum quasi-likelihood estimate of (4.1.4) is to maximize

$$\ell(\boldsymbol{\beta}) \widehat{=} \sum_{i=1}^{n} Q\left[g^{-1}\left\{\sum_{j=1}^{p} \boldsymbol{\beta}_j^T \boldsymbol{\psi}_j(U_i) X_{ij}\right\}; Y_i\right] = \sum_{i=1}^{n} Q[g^{-1}(\mathbf{z}_i^T \boldsymbol{\beta}); Y_i], \tag{3.1.4}$$

with respect to $\boldsymbol{\beta}$, where $\mathbf{z}_i = (X_{i1}\boldsymbol{\psi}_1(U_i)^T, \cdots, X_{ip}\boldsymbol{\psi}_p(U_i)^T)^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_p^T)^T$. With slight abuse notation, we use $\ell\{\boldsymbol{\alpha}(\cdot)\}$ in (4.1.4) and $\ell(\boldsymbol{\beta})$ in (4.1.5). However, the notation will be clear in the context. In the presence of ultrahigh dimensional covariate $\mathbf{x}$, the corresponding optimization problem becomes ill-posed. It is typical to assume sparsity. That is, only a few x-covariates are significant, and the others do not have impact on the response. We next propose a feature screening procedure for model (4.1.4).

3.2 A New Feature Screening Procedure

Denote $\|\alpha_j(U)\|_2 = [E\alpha_j^2(U)]^{1/2}$, the L_2-norm of $\alpha_j(U)$. For ease of presentation, s denotes an arbitrary subset of $\{1, \ldots, p\}$, $\mathbf{x}_s = \{x_j, j \in s\}$ and $\boldsymbol{\alpha}_s(U) = \{\alpha_j(U), j \in s\}$. For a set s, $\tau(s)$ stands for the cardinality of s. Suppose the effect of $\mathbf{x}$ is sparse, and the true value of $\boldsymbol{\alpha}(U)$ is $\boldsymbol{\alpha}^*(U)$, so $\boldsymbol{\beta}$ is corresponding to $\boldsymbol{\beta}^*$. Denote $s^* = \{j : \|\alpha_j(U)\|_2 > 0\}$. By sparsity, we means that $\tau(s^*)$ is much less than p. The goal of feature screening is to identify a subset s such that $s^* \subset s$ with overwhelming probability and $\tau(s)$ is also much less than p. Theoretically we may formulate this problem to be an optimization problem as below:

$$\max_{\boldsymbol{\alpha}(\cdot)} \ell\{\boldsymbol{\alpha}(\cdot)\} \quad \text{subject to } \tau(\{j : \|\alpha_j(\cdot)\|_2^2 > 0\}) \leqslant m, \tag{3.2.5}$$

for a pre-specified m, which is presumed to be much less than p.

When the approximation error is negligible, we construct a feature screening procedure by considering the following maximization problem:

$$\max_{\boldsymbol{\alpha}_n(\cdot)} \ell\{\boldsymbol{\alpha}_n(\cdot)\} \quad \text{subject to } \tau(\{j : \|\alpha_{nj}(\cdot)\|_2^2 > 0\}) \leqslant m. \tag{3.2.6}$$

Note that $\|\alpha_{nj}(U)\|_2^2 = \boldsymbol{\beta}_j^T E\{\boldsymbol{\psi}_j(U)\boldsymbol{\psi}_j(U)^T\}\boldsymbol{\beta}_j$. Under the assumption that $E\{\boldsymbol{\psi}_j(U)\boldsymbol{\psi}_j(U)^T\}$ is finite positive definite for all $j = 1, \cdots, p$, the maximization problem in (4.1.6) is equivalent to

$$\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}) \quad \text{subject to } \tau(\{j : \|\boldsymbol{\beta}_j\|_2^2 > 0\}) \leqslant m. \tag{3.2.7}$$

For high dimensional problems, it becomes almost impossible to solve the constrained maximization problem (3.2.7) directly. Alternatively, we consider a proxy of the quasi-likelihood function. It follows by the Taylor expansion for the quasi-likelihood function $\ell(\boldsymbol{\gamma})$ at $\boldsymbol{\beta}$ lying within a neighbor of $\boldsymbol{\gamma}$ that

$$\ell(\boldsymbol{\gamma}) \approx \ell(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell'(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell''(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}),$$

where $\ell'(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$ and $\ell''(\boldsymbol{\beta}) = \partial^2 \ell(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T|_{\boldsymbol{\gamma}=\boldsymbol{\beta}}$. Denote $P_t = \sum_{j=1}^p d_{nj}$. If $\ell''(\boldsymbol{\beta})$ is invertible, the computational complexity of calculating the

inverse of $\ell''(\boldsymbol{\beta})$ is $O(P_t^3)$. For large P_t, small n problems (i.e. $P_t \gg n$), $\ell''(\boldsymbol{\beta})$ becomes not invertible. Low computational cost is always desirable for feature screening. To cope with singularity of the Hessian matrix and save computational cost, we propose using the following approximation for $\ell''(\boldsymbol{\gamma})$

$$h(\boldsymbol{\gamma}|\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + (\boldsymbol{\gamma}-\boldsymbol{\beta})^T\ell'(\boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma}-\boldsymbol{\beta})^T W(\boldsymbol{\beta})(\boldsymbol{\gamma}-\boldsymbol{\beta}), \tag{3.2.8}$$

where u is a scaling constant to be specified and $W(\boldsymbol{\beta}) = \text{diag}(W_1(\boldsymbol{\beta}), \cdots, W_p(\boldsymbol{\beta}))$, a block diagonal matrix with $W_j(\boldsymbol{\beta})$ being a $d_{nj} \times d_{nj}$ matrix. Here we allow $W(\boldsymbol{\beta})$ to depend on $\boldsymbol{\beta}$. This implies that we approximate $\ell''(\boldsymbol{\beta})$ by $-uW(\boldsymbol{\beta})$. Throughout this paper, we will use $W_j(\boldsymbol{\beta}) = -\partial^2\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_j\partial\boldsymbol{\beta}_j^T$.

It can be seen that $h(\boldsymbol{\beta}|\boldsymbol{\beta}) = \ell(\boldsymbol{\beta})$, and under some conditions, $h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \leqslant \ell(\boldsymbol{\beta})$ for all $\boldsymbol{\gamma}$. This ensures the ascent property. See Theorem 1 below for more details. Since $W(\boldsymbol{\beta})$ is a block diagonal matrix, $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ is an additive function of γ_j for any given $\boldsymbol{\beta}$. The additivity enables us to have a closed form solution for the following maximization problem

$$\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\boldsymbol{\beta}) \qquad \text{subject to } \tau(\{j : \|\boldsymbol{\gamma}_j\|_2^2 > 0\}) \leqslant m, \tag{3.2.9}$$

for given $\boldsymbol{\beta}$ and m. Define $\tilde{\boldsymbol{\gamma}}_j = \boldsymbol{\beta}_j + u^{-1}W_j^{-1}(\boldsymbol{\beta}_j)\partial\ell(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_j$ for $j = 1, \cdots, p$, and $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\gamma}}_1^T, \cdots, \tilde{\boldsymbol{\gamma}}_p^T)^T = \boldsymbol{\beta} + u^{-1}W^{-1}(\boldsymbol{\beta})\ell'(\boldsymbol{\beta})$ is the maximizer of $h(\boldsymbol{\gamma}|\boldsymbol{\beta})$. Denote $g_j = \tilde{\boldsymbol{\gamma}}_j^T W_j(\boldsymbol{\beta}_j)\tilde{\boldsymbol{\gamma}}_j$ for $j = 1, \cdots, p$, and sort g_j so that $g_{(1)} \geqslant g_{(2)} \geqslant \cdots \geqslant g_{(p)}$. The solution of maximization problem (3.2.9) is the hard-thresholding rule defined below

$$\widehat{\boldsymbol{\gamma}}_j = \tilde{\boldsymbol{\gamma}}_j I\{g_j > g_{(m+1)}\}.$$

This enables us to effectively screen features by using the following algorithm.

Step 1. Set the initial value $\boldsymbol{\beta}_j^{(0)} = \mathbf{0}$, $j = 1, \cdots, p$.

Step 2. Set $t = 0, 1, 2, \cdots$, iteratively conduct Step 2a and Step 2b below until the algorithm converges.

Step 2a. Calculate $\tilde{\boldsymbol{\gamma}}_j^{(t)} = \boldsymbol{\beta}_j^{(t)} + u_t^{-1}W_j^{-1}(\boldsymbol{\beta}_j)\partial\ell(\boldsymbol{\beta}^{(t)})/\partial\boldsymbol{\beta}_j$, and $g_j^{(t)} = \{\tilde{\boldsymbol{\gamma}}_j^{(t)}\}^T W_j(\boldsymbol{\beta}^{(t)})\tilde{\boldsymbol{\gamma}}_j^{(t)}$. Let $g_{(1)}^{(t)} \geqslant g_{(2)}^{(t)} \geqslant \cdots \geqslant g_{(p)}^{(t)}$, the order statistics of $g_j^{(t)}$s. Set $S_t = \{j : g_j^{(t)} \geqslant g_{(m+1)}^{(t)}\}$, the nonzero index set.

Step 2b. Update $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{\beta}_1^{(t+1)}, \cdots, \boldsymbol{\beta}_p^{(t+1)})^T$ as follows. If $j \notin S_t$, set $\boldsymbol{\beta}_j^{(t+1)} = \mathbf{0}$, otherwise, set $\{\boldsymbol{\beta}_j^{(t+1)} : j \in S_t\}$ be the maximum likelihood estimate of the submodel S_t.

Remark: Unlike the screening procedures based on marginal partial likelihood methods, our proposed procedure is to iteratively update $\boldsymbol{\beta}$ using Step 2. This enables the proposed screening procedure to incorporate correlation information among the predictors through updating $\ell_p'(\boldsymbol{\beta})$ and $\ell_p''(\boldsymbol{\beta})$. Thus, the proposed procedure is expected to perform better than the marginal screening procedures when there are some predictors that are marginally independent. Meanwhile, since each iteration in Step 2 can avoid large-scale matrix inversion and, therefore, it can be carried out with low computational costs.

Conditioning on S_t containing all actively predictors, one may directly apply existing results in spline regression for $\boldsymbol{\beta}_j^{(t+1)}$ and $\widehat{\boldsymbol{\alpha}}_j$. Without conditioning on S_t, it is very challenging in establishing theoretical properties for $\boldsymbol{\beta}_j^{(t+1)}$s.

Theorem 3.2.1. *Let $\{\boldsymbol{\beta}^{(t)}\}$ be the sequence defined in Step 2b in the above algorithm. Denote*

$$\rho^{(t)} = \sup_{\boldsymbol{\beta}} \Big[\lambda_{\max}\{W^{-1/2}(\boldsymbol{\beta}^{(t)})\{-\ell''(\boldsymbol{\beta})\}W^{-1/2}(\boldsymbol{\beta}^{(t)})\}\Big].$$

Here and hereafter $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ stands for the maximal and the minimal eigenvalues of a matrix A, respectively. If $u_t \geqslant \rho^{(t)}$, then

$$\ell(\boldsymbol{\beta}^{(t+1)}) \geqslant \ell(\boldsymbol{\beta}^{(t)}),$$

where $\boldsymbol{\beta}^{(t+1)}$ is defined in Step 2b in the above algorithm.

Theorem 3.2.1 claims the ascent property of the proposed algorithm if u_t is appropriately chosen. That is, the proposed algorithm may improve the current estimate within the feasible region (i.e. $\tau(\{j : \|\alpha_j(U)\|_2 > 0\}) \leqslant m$), and the resulting estimate in the current step may serve as a refinement of the last step. This theorem also provides us some insights about choosing u_t in practical implementation. For varying coefficient models: $E(Y|U, \mathbf{x}) = \mathbf{x}^T\boldsymbol{\alpha}(U)$, we may set $\ell\{\boldsymbol{\alpha}(\cdot)\} = -2^{-1}\sum_{i=1}^n\{Y_i - \mathbf{x}_i\boldsymbol{\alpha}(U_i)\}^2$. In this case, $\ell(\boldsymbol{\beta})$ in (4.1.5) is

$\ell(\boldsymbol{\beta}) = -2^{-1}\sum_{i=1}^{n}(Y_i - \mathbf{z}_i^T\boldsymbol{\beta})^2$. Thus, $-\ell''(\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i^T = \mathbf{Z}^T\mathbf{Z}$, where $\mathbf{Z}$ is $n \times p_t$ matrix with i-th row being $\mathbf{z}_i^T$. Thus,

$$\rho^{(t)} = \lambda_{\max}(\text{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}(\mathbf{Z}^T\mathbf{Z})\text{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}),$$

which does not depend on the step of iteration t. If $\mathbf{z}_i$'s are marginally standardized so that its marginal sample mean and sample standard deviation equal 0 and 1, respectively, then $\text{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}(\mathbf{Z}^T\mathbf{Z})\text{diag}(\mathbf{Z}^T\mathbf{Z})^{-1/2}$ is the corresponding sample correlation matrix of $\mathbf{z}_i$'s. Thus, ρ is the largest eigenvalue of the sample correlation matrix.

3.3 Sure Screening Property

For a subset s of $\{1, \ldots, p\}$ with size $\tau(s)$, recall notation $\mathbf{x}_s = \{x_j, j \in s\}$ and associated coefficients $\boldsymbol{\alpha}_s(U) = \{\alpha_j(U), j \in s\}$ corresponding to $\boldsymbol{\beta}_s = \{\boldsymbol{\beta}_j, j \in s\}$ with $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_{nj}})$. We denote the true model by $s^* = \{j : E\alpha_j^2(U) > 0, 1 \leqslant j \leqslant p\}$ with $\tau(s^*) = q$. The objective of feature selection is to obtain a subset $\widehat{s}$ such that $s^* \subset \widehat{s}$ with very high probability.

We now provide some theoretical justifications for the screening procedure for the GVCM. The sure screening property (Fan and Lv, 2008)) is referred to as

$$Pr(s^* \subset \widehat{s}) \longrightarrow 1, \quad as \quad n \to \infty. \tag{3.3.10}$$

To establish this sure screening property for the proposed feature screening method, we introduce some additional notations as follows. For any model s, let $\ell'(\boldsymbol{\beta}_s) = \partial\ell(\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s$ and $\ell''(\boldsymbol{\beta}_s) = \partial^2\ell(\boldsymbol{\beta}_s)/\partial\boldsymbol{\beta}_s\partial\boldsymbol{\beta}_s^T$ be the score function and the Hessian matrix of $\ell(\cdot)$ as a function of $\boldsymbol{\beta}_s$, respectively. Assume that a screening procedure retains m out of p features such that $\tau(s^*) = q < m$. So, we define

$$S_+^m = \{s : s^* \subset s; \|s\|_0 \leqslant m\} \quad \text{and} \quad S_-^m = \{s : s^* \not\subset s; \|s\|_0 \leqslant m\} \tag{3.3.11}$$

as the collections of the over-fitted models and the under-fitted models. We investigate the asymptotic properties of $\widehat{\boldsymbol{\beta}}_m$ under the scenario where p, q, m and $\boldsymbol{\beta}^*$ are allowed to depend on the sample size n. We impose the following conditions, some

of which are purely technical and only serve to facilitate theoretical understanding of the proposed feature screening procedure.

(C1) The support of U is bounded and is assumed to be $[a, b]$.

(C2) The functions $\{\alpha_j(U)\}_{j=1}^p$ belong to a class of functions $\mathcal{F}$, whose rth derivative $\alpha_j^{(r)}$ exists and is Lipschitz of order η,

$$\mathcal{F} = \left\{\alpha_j(\cdot) : |\alpha_j^{(r)}(s) - \alpha_j^{(r)}(t)| \leqslant K|s-t|^\eta \text{ for } s, t \in [a, b]\right\},$$

for some positive constant K, where r is a nonnegative integer and $\eta \in (0, 1]$ such that $\upsilon = r + \eta > 0.5$.

(C3) There exists $w_1, w_2 > 0$ and for some non-negative constants τ_1, τ_2 such that $\tau_1 + \tau_2 < 2/5$ with

$$\min_{j \in s^*} \|\alpha_j(U)\|_2 \geqslant w_1 n^{-\tau_1} \quad \text{and} \quad q < m \leqslant w_2 n^{\tau_2}.$$

(C4) $\log p = O(n^\kappa / d_n)$ for some $1/5 \leqslant \kappa < 1 - 2(\tau_1 + \tau_2)$ and $d_n = O(n^{1/5})$.

(C5) $\mu'(\cdot)/V(\cdot)$ is bounded by some constant $M > 0$, where $\mu\{\cdot\}$ and $V(\cdot)$ are the mean and variance functions used for the quasi-likelihood function, respectively.

(C6) There exist constants $C_1, C_2 > 0$, $\delta > 0$, such that for sufficiently large n,

$$C_1 d_n^{-1} \leqslant \lambda_{\min}[-n^{-1}\ell''(\boldsymbol{\beta}_s)] \leqslant \lambda_{\max}[-n^{-1}\ell''(\boldsymbol{\beta}_s)] \leqslant C_2 d_n^{-1},$$

for $\boldsymbol{\beta}_s \in \{\boldsymbol{\beta} : \|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\|_2 \leqslant \delta\}$ and $s \in S_+^{2m}$, where $\lambda_{\min}[\cdot]$ and $\lambda_{\max}[\cdot]$ denote the smallest and largest eigenvalues of a matrix.

(C7) There exists positive constants a and b such that $P(|X_j Y| \geqslant x) \leqslant a \exp(-bx)$ for all $x > 0$.

Under Conditions (C1) and (C2), the following two properties of B-splines are valid.

(a) (de Boor, 1978) For $k = 1, \ldots, d_n$, $\psi_{jk}(U) \geqslant 0$ and $\sum_{k=1}^{d_n} \psi_{jk}(U)^2 = 1$, $U \in [a, b]$. In addition, there exist positive constants C_3 and C_4 such that $C_3 d_n^{-1} \leqslant E\psi_{jk}^2(U) \leqslant C_4 d_n^{-1}$.

(b) (Stone, 1982, 1985) If $\{\alpha_j, j = 1, 2, \cdots, p\}$ is a set of functions in $\mathcal{F}$ described in condition (C2), there exists a positive constant C_5 that does not depend on $\alpha_j(U)$ so that the uniform approximation error has the following bound. $\rho = \sup_{U \in [a,b]} \|\alpha_j(U) - \alpha_{nj}(U)\|_2 \leqslant C_5 d_n^{-v}, \forall j$, as $d_n \to \infty$.

Conditions (C1) and (C2) ensure properties (a) and (b), which are required for the B-spline approximation and establishing the sure screening properties.

Note that $\|\alpha_{nj}(U)\|_2^2 = \boldsymbol{\beta}_j^T E\{\boldsymbol{\psi}_j(U)\boldsymbol{\psi}_j(U)^T\}\boldsymbol{\beta}_j$, based on the properties (a), (b) and Condition (C3), we can derive that

$$\min_{j \in s^*} \|\boldsymbol{\beta}_j\|_2 \geqslant w_1 d_n n^{-\tau_1}. \tag{3.3.12}$$

Condition (C3) states a few requirements for establishing the sure screening property of the proposed procedure. The first one is the sparsity of $\boldsymbol{\beta}^*$ which makes the sure screening possible with $\tau(\widehat{s}) = m > q$. Condition (C3) requires that the signal of the active components ($\|\alpha_j(U)\|_2, j \in s^*$) does not vanish. This is referred to as minimal signal condition in the literature. Minimal signal condition is a commonly-imposed assumption in existing work on marginal feature screening for other model (e.g, Liu, et al., 2014). By (3.3.12), it is equivalent to requiring that the minimal component in $\boldsymbol{\beta}^*$ does not degenerate too fast, so that the signal is detectable in the asymptotic sequence. Condition (C4) has p diverge with n at up to an exponential rate. Meanwhile, together with (C6), it confines an appropriate order of m that guarantees the identifiability of s^* over s for $\tau(s) \leqslant m$. For varying coefficient model discussed in Section 2.1, Condition (C6) requires

$$C_1 d_n^{-1} \leqslant \lambda_{\min}[n^{-1}\mathbf{Z}_s^T\mathbf{Z}_s] \leqslant \lambda_{\max}[n^{-1}\mathbf{Z}_s^T\mathbf{Z}_s] \leqslant C_2 d_n^{-1},$$

where $\mathbf{Z}_s$ is the corresponding design matrix of model s. We establish the sure screening property of the quasi-likelihood estimation by the following theorem. Condition C7 indicates that there exists a positive constant t_0 and g such that for all $|t| \leqslant t_0$, $E\{\exp[t(X_jY - E(X_jY)]\} < \exp(gt^2/2)$ uniformly for $j = 1, \cdots, p$.

Theorem 3.3.1. *Suppose we have n independent observations with p candidate features from model (4.1.2) and conditions (C1)—(C7) are satisfied. Let $\widehat{s}$ be the features obtained by (3.2.5) of size m. Then, we have*

$$Pr(s^* \subset \widehat{s}) \to 1, \ \ as \ n \to \infty.$$

The proof is given in the Appendix. The sure screening property is an appealing property of a screening procedure since it ensures that the true active predictors are retained in the model selected by the screening procedure. We establish the sure screening property under weaker conditions imposed in Fan, et al. (2014) and Xia, et al. (2016).

One has to specify the value of m in practical implementation. As to the choice of m, there are two scenarios. The first one chooses m by a data-driven method that described in Section 2.3. The second one is an ad hoc method. In the literature of feature screening, it is typical to set $m = [n/\log(n)]$ for a parametric model, where $[a]$ indicates the integer part of a (Fan and Lv, 2008). Since we use a linear combination of d_n B-spline bases in our proposed screening procedure for the GVCM, we set $m = [(n/d_n)/\log(n/d_n)]$ throughout in Examples 3.1, 3.2 and 3.3. Although it is an ad hoc choice, it works reasonably well in our numerical examples. With this choice of m, one is ready to further apply existing methods such as the penalized quasi-likelihood method to further remove inactive predictors. To be distinguished from the SIS procedure, the proposed procedure is referred to as sure joint screening (SJS) procedure.

3.3.1 Choice of m

Feature screening may be used in various contexts. In some contexts, people may treated m as a pre-specified value. For example, due to budget constraint, a biologist may be able to examine up to m genes that potentially associate with a certain phenotype. In other contexts, people may treat m as a tuning parameter to control model complexity. In such cases, it is desirable to develop an automatic data-driven method to determine m. We propose to select m by minimizing the

high-dimensional BIC score:

$$HBIC(m) = -2\ell(\widehat{\boldsymbol{\beta}}_m) + d_n m \frac{C_n \log(d_n p)}{n},$$

where $\widehat{\boldsymbol{\beta}}_j = (\widehat{\beta}_{j1}, \ldots, \widehat{\beta}_{jd_n})$, $j = 1, \ldots, m$, and C_n is a sequence of numbers that diverges to ∞. Wang, et al. (2013) proposed the HBIC for selecting tuning parameter in the penalized least squares method for high dimensional linear models. Here we modified their proposal for high dimensional generalized varying-coefficient models. In our simulation, we take $C_n = \log\log n$, and compare its performance with AIC and BIC tuning parameter selectors defined in the same manner. It is worth to noting that the proposed tuning parameter HBIC selector requires to search over $m = 1, 2, \cdots, [n/d_n]$. This is distinguished from that the classical AIC and BIC used for subset selection requires to search over subsets. Thus, the tuning parameter selector does not require expensive computational cost.

Recall notation S_+^m and S_-^m defined in (3.3.11). Theorem 3.3.2 below shows that the HBIC selects the right model size almost surely.

Theorem 3.3.2. *Suppose we have n independent observations with p candidate features from model (4.1.2) and conditions (C3)—(C6) are satisfied. Let $\widehat{s}$ be the features obtained by (4.1.5) and (3.2.7) of size m. Then, we have*

$$Pr\left\{\min_{s \in S_-^m} HBIC(\tau(s)) \leqslant HBIC(q)\right\} \longrightarrow 0, \tag{3.3.13}$$

where $q = \tau(s^*)$, and

$$Pr\left\{\min_{s \in S_+^m, s \neq s^*} HBIC(\tau(s)) \leqslant HBIC(q)\right\} \longrightarrow 0. \tag{3.3.14}$$

In Example 3.4, we will examine the performance of the proposed HBIC tuning parameter selector.

3.4 Numerical Studies

In this section, we conduce numerical studies to examine the finite sample performance of the proposed feature screening procedures and compare it with the

existing ones. All simulation are conducted by using R code.

3.4.1 Simulation Studies

In our simulation, the covariate u and $\mathbf{x}$ are generated as follows: first draw $(U^*, \mathbf{x})^T$ from a $p+1$ dimensional normal distribution $N(0, \Sigma)$. Then set $U = \Phi(U^*)$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$. Thus, U follows a uniform distribution $U(0,1)$ and is correlated with x, and all the predictors $X_1, ..., X_p$ are correlated with each other. In our simulation, we consider two scenarios for $\Sigma = (\sigma_{ij})$

Σ_1: Compound symmetric correlation structure: $\sigma_{ij} = 1$ if $i = j$ and ρ otherwise.

Σ_2: AR(1) correlation structure: $\sigma_{ij} = \rho^{|i-j|}$.

In our numerical studies, we set the number of B-spline basis functions to be $d_n = 5$ for each coefficient function and set the threshold in (**??**) $m = [n/\log(n)]$. We use the following two criteria to assess the performance of the proposed procedure.

P_a: The proportion of submodels $\widehat{\mathcal{M}}$ with size d that contain all the true predictors among 1000 simulations.

P_j: The proportion of submodels $\widehat{\mathcal{M}}$ with size d that contain X_j among 1000 simulations.

Example 3.2.1.1. This example is designated to compare the proposed screening procedure with existing SIS procedures for VCM. Since the proposal of Fan, Ma and Dai (2014) shares the same spirit as that of Liu, Li and Wu (2014), and Song, Yi and Zou (2014) and Chu, Li and Reimherr (2016) were proposed for longitudinal data, we will concentrate on our comparison with CC-SIS proposed by Liu, Li and Wu (2014). Given $\{U, \mathbf{x}\}$, we generate a continuous response from

$$Y = \alpha_1(U)X_1 + \alpha_2(U)X_2 + \alpha_3(U)X_3 + \alpha_4(U)X_4 + \varepsilon, \tag{3.4.1}$$

where $\varepsilon \sim N(0,1)$. Model (3.4.1) implies that $\alpha_j(\cdot) = 0$ for $j > 4$ and $\mathcal{M}_* = \{1, 2, 3, 4\}$. We consider two sets of coefficient functions:

$\boldsymbol{\alpha}_1$: Let $\alpha_1(u) = \alpha_2(u) = \alpha_3(u) = 2 + 2\sin^2(2\pi u)$, and $\alpha_4(u) = -3\rho * \alpha_1(u)$.

$\boldsymbol{\alpha}_2$: $\alpha_1(u) = -(3 + 2\cos^2(\frac{\pi}{2}u))$, $\alpha_2(u) = -(3 + 3u)$, $\alpha_3(u) = (2 - u)^2 + 2$, $\alpha_4(u) = 3 + 2\sin^2(\frac{\pi}{2}u)$.

In this example, we consider $p = 1000$ and 2000, and the sample size $n = 200$ and 400. All simulation results are based on 1000 replications. Simulation results are summarized in Tables 3.4.1—3.4.2.

Table 3.4.1 shows the values of $\mathcal{P}_1, \cdots, \mathcal{P}_4$ and $\mathcal{P}_a$ for continuous response with $\Sigma = \Sigma_1$. Under the design of $\boldsymbol{\alpha}_1$, X_4 is jointly dependent but marginally independent of Y. In this setting, the marginal screening procedure always fails to identify X_4. As shown in Table 3.4.1, when there exists marginal independence, it is hard for CC-SIS to detect X_4 whose values of $\mathcal{P}_4$ and $\mathcal{P}_a$ are small as expected. However, our method can identify X_4 in this setting and the corresponding values of $\mathcal{P}_4$ and $\mathcal{P}_a$ are close to one. Therefore, our new procedure outperforms CC-SIS in the marginal independence setting. Under the design of $\boldsymbol{\alpha}_2$, there is no predictor that is jointly dependent but marginal independent of Y. The performances of both CC-SIS and the proposed procedure are good, as the detecting probabilities are close to one. However, CC-SIS performs better when the sample size increases and the dimensionality decreases. On the other hand, those factors have less influences on the new procedure than CC-SIS. Furthermore, the corresponding values of $\mathcal{P}_j$s and $\mathcal{P}_a$ of our new procedure are closer to one in every case in this setting. In a word, when $\Sigma = \Sigma_1$, regardless of whether marginal independence exists, our new procedure outperforms CC-SIS which suggests its sure screening property.

Table 3.4.2 shows the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ for continuous response with $\Sigma = \Sigma_2$. There is no predictor that is jointly dependent but marginal independent of Y. Hence both of the CC-SIS and the new procedure perform well, as most of the values of $\mathcal{P}_a$ are greater than 0.9. Table 3.4.2 also indicates that when the sample size increases and the dimensionality decreases, both CC-SIS and our new procedure perform better. Futhermore, this table also shows that those factors have less effect on our new procedure. For instance, when $n = 200$, some values of $\mathcal{P}_a$ obtained by CC-SIS are less than 0.8, but the corresponding values of $\mathcal{P}_a$ of the new procedure are close to one. Besides, Table 3.4.2 shows that the new procedure performs better than CC-SIS in every case, which is consistent with our theoretical

analysis since our new procedure has the sure screening property. Hence, our new procedure also beats CC-SIS in the setting of $\Sigma = \Sigma_2$.

In addition, comparing the two methods with different ρ's, Table 3.4.1 and Table 3.4.2 show that when ρ increases, the performance of CC-SIS and the new procedure become worse. This is because when the predictors are highly correlated, the unimportant predictors may be selected due to their strong correlations with the true predictors.

Table 3.4.1. The proportions of $\mathcal{P}_j$s and $\mathcal{P}_a$ for Continuous Response with $\Sigma = \Sigma_1$

				CC-SIS					New (SJS)				
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$
200	1000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	0	0	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	0.995	1	1	0.992	0.987	1	1	1	1	1
200	1000	1/2	$\boldsymbol{\alpha}_1$	1	1	1	0.015	0.015	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	0.994	0.999	0.996	0.979	0.970	1	1	1	1	1
200	1000	2/3	$\boldsymbol{\alpha}_1$	0.995	0.997	0.995	0.302	0.297	1	1	0.999	1	0.999
			$\boldsymbol{\alpha}_2$	0.976	0.995	0.984	0.942	0.909	1	1	1	1	1
200	2000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	0.001	0.001	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	0.992	0.999	0.998	0.989	0.979	1	1	1	1	1
200	2000	1/2	$\boldsymbol{\alpha}_1$	0.999	0.997	0.998	0.008	0.008	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	0.991	0.998	0.994	0.973	0.958	1	1	1	1	1
200	2000	2/3	$\boldsymbol{\alpha}_1$	0.989	0.987	0.985	0.284	0.274	1	1	0.993	1	0.993
			$\boldsymbol{\alpha}_2$	0.974	0.999	0.976	0.932	0.892	1	1	1	1	1
400	1000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	0	0	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
400	1000	1/2	$\boldsymbol{\alpha}_1$	1	1	1	0.023	0.023	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
400	1000	2/3	$\boldsymbol{\alpha}_1$	1	1	1	0.623	0.623	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
400	2000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	0	0	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
400	2000	1/2	$\boldsymbol{\alpha}_1$	1	1	1	0.011	0.011	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
400	2000	2/3	$\boldsymbol{\alpha}_1$	1	1	1	0.549	0.549	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1

Our method has advantages in terms of the computing efficiency as well. Table 3.4.3 shows the medians and MADs of computing time (seconds), and the number of iterations for continuous responses. When $p = 1000$, most of the medians of the computing times are below 5 seconds, and the MAD is pretty small; when $p = 2000$, the computing times become larger, but the medians are still mostly below 9 seconds and the MADs are also small. In general, the algorithm converges faster as the sample size increases. As shown in Table 3.4.3, the algorithm always converges

Table 3.4.2. The proportions of $\mathcal{P}_j$s and $\mathcal{P}_a$ for Continuous Response with $\Sigma = \Sigma_2$

				CC-SIS					New (SJS)				
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$
200	1000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	0.644	0.644	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
200	1000	1/2	$\boldsymbol{\alpha}_1$	1	1	1	0.887	0.887	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	0.996	0.999	0.995	1	1	1	1	1
200	1000	2/3	$\boldsymbol{\alpha}_1$	1	1	0.741	0.990	0.731	1	1	0.952	1	0.952
			$\boldsymbol{\alpha}_2$	1	0.745	0.999	1	0.744	1	1	0.998	1	0.998
200	2000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	0.551	0.551	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
200	2000	1/2	$\boldsymbol{\alpha}_1$	1	1	0.997	0.858	0.855	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	0.991	0.999	1	0.990	1	1	1	1	1
200	2000	2/3	$\boldsymbol{\alpha}_1$	1	1	0.678	0.991	0.669	1	1	0.903	1	0.903
			$\boldsymbol{\alpha}_2$	0.999	0.693	0.999	1	0.692	1	1	0.996	1	0.996
400	1000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	0.982	0.982	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
400	1000	1/2	$\boldsymbol{\alpha}_1$	1	1	1	1	1	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
400	1000	2/3	$\boldsymbol{\alpha}_1$	1	1	0.993	1	0.993	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	0.996	1	1	0.996	1	1	1	1	1
400	2000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	0.951	0.951	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
400	2000	1/2	$\boldsymbol{\alpha}_1$	1	1	1	0.999	0.999	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
400	2000	2/3	$\boldsymbol{\alpha}_1$	1	1	0.991	1	0.991	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	0.986	1	1	0.986	1	1	1	1	1

after 5 iterations when $n = 400$ and it usually converges after 10 iterations when $n = 200$. All of the facts above show that our new procedure is highly efficient.

Example 3.2.1.2. This example is designated to examine the performance of the proposed procedures for binary response. Given $\{U, \mathbf{x}\}$, we generate a binary response with the probability of $Y = 1$ being $p(U, \mathbf{x})$ defined below.

$$\text{logit}\{p(U, \mathbf{x})\} = \alpha_1(U)X_1 + \alpha_2(U)X_2 + \alpha_3(U)X_3 + \alpha_4(U)X_4, \tag{3.4.2}$$

where $\text{logit}(t) = \log\{t/(1-t)\}$, the logit link in the logistic regression. Model (3.4.2) implies that $\alpha_j(\cdot) = 0$ for $j > 4$ and $\mathcal{M}_* = \{1, 2, 3, 4\}$. In this example, the coefficients are set to be the same as those in Example 3.2.1.1.

In this example, we consider $p = 1000$ and 2000, and the sample size $n = 300$ and 500. All simulation results are based on 1000 replications, and are summarized in Tables 3.4.4—3.4.5.

Table 3.4.3. Computing times (Seconds) and the Number of Iterations for Continuous Response

	Σ_1				Σ_2			
	$\boldsymbol{\alpha}_1$		$\boldsymbol{\alpha}_2$		$\boldsymbol{\alpha}_1$		$\boldsymbol{\alpha}_2$	
ρ	Time	Iterations	Time	Iterations	Time	Iterations	Time	Iterations
	$(n, p) = (200, 1000)$							
1/3	3.97(0.17)	10(0)	4.10(0.36)	10(0)	4.13(0.45)	10(0)	3.90(0.20)	10(0)
1/2	4.22(0.24)	10(0)	5.03(0.87)	10(0)	3.98(0.83)	10(0)	4.25(0.37)	10(0)
2/3	3.93(0.11)	10(0)	4.08(0.83)	10(0)	4.25(0.36)	10(0)	4.21(0.32)	10(0)
	$(n, p) = (200, 2000)$							
1/3	7.87(0.47)	10(0)	7.37(0.63)	10(0)	8.04(0.70)	10(0)	7.24(0.20)	10(0)
1/2	7.91(0.59)	10(0)	8.40(0.53)	10(0)	7.98(0.53)	10(0)	7.25(0.21)	10(0)
2/3	7.75(0.61)	10(0)	7.03(0.64)	10(0)	8.05(0.35)	10(0)	7.15(0.39)	10(0)
	$(n, p) = (400, 1000)$							
1/3	2.73(0.37)	5(1)	2.03(0.3)	4(1)	2.98(0.41)	5(1)	2.89(0.46)	5(0)
1/2	2.20(0.21)	4(0)	1.44(0.10)	3(0)	2.91(0.40)	5(1)	2.86(0.46)	5(1)
2/3	1.98(0.30)	4(1)	1.50(0.22)	3(0)	2.42(0.39)	5(1)	2.58(0.33)	5(1)
	$(n, p) = (400, 2000)$							
1/3	4.87(0.67)	5(1)	3.73(0.47)	4(0)	4.87(0.57)	5(1)	6.01(0.98)	5(1)
1/2	3.69(0.29)	4(0)	3.34(0.55)	3(0)	5.97(1.05)	5(1)	6.03(0.93)	5(1)
2/3	3.18(0.43)	4(0)	2.34(0.68)	3(0)	4.67(0.68)	5(1)	6.54(1.72)	5(1)

Table 3.4.4 shows the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ for the binary responses. Under the design of Σ_1 and $\boldsymbol{\alpha}_1$, X_4 is jointly dependent but marginally independent of Y. As shown in Table 3.4.4, the values of $\mathcal{P}_4$ and $\mathcal{P}_a$ are very close to one, which means our method is able to identify the predictor that is jointly important but marginally independent of the response. In general, $\mathcal{P}_4$ is the largest and this is because the absolute value of $\alpha_4(U)$ is no less than those of the other three coefficient functions, which makes X_4 much easier to be identified. If there is no marginal independence, the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ are very close to one. From the table, we see that the values of $\mathcal{P}_a$ are mostly greater than 0.9. In addition, our procedure performs better as the sample size increases and the dimensionality decreases, which is also consistent to the sure screening property of the new method.

Furthermore, comparing the performance of the new procedure under different ρ's, Table 3.4.4 shows that the new procedure performs better as the value of ρ decreases. This is the same as that happened in linear regression model setting. The reason is also that the unimportant predictors may be detected because of their strong correlations with the true predictors.

The computing efficiency of the proposed procedure for binary response can be seen from Table 3.4.5, where the medians and MADs of computing time (seconds)

Table 3.4.4. The proportions of $\mathcal{P}_j$s and $\mathcal{P}_a$ for Binary Response

				$\Sigma = \Sigma_1$					$\Sigma = \Sigma_2$				
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$
300	1000	1/3	$\boldsymbol{\alpha}_1$	0.999	0.998	1	1	0.997	1	1	0.998	0.994	0.992
			$\boldsymbol{\alpha}_2$	0.999	1	1	1	0.999	1	1	1	1	1
300	1000	1/2	$\boldsymbol{\alpha}_1$	0.983	0.987	0.987	1	0.958	1	1	0.984	1	0.984
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	0.996	1	0.996
300	1000	2/3	$\boldsymbol{\alpha}_1$	0.925	0.928	0.946	1	0.813	1	1	0.896	0.996	0.894
			$\boldsymbol{\alpha}_2$	0.995	1	0.996	0.994	0.988	1	0.997	0.976	1	0.973
300	2000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	1	1	1	1	0.998	0.99	0.988
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
300	2000	1/2	$\boldsymbol{\alpha}_1$	0.974	0.98	0.984	1	0.941	0.998	1	0.955	0.999	0.952
			$\boldsymbol{\alpha}_2$	0.999	1	1	0.998	0.997	1	1	0.994	1	0.994
300	2000	2/3	$\boldsymbol{\alpha}_1$	0.898	0.903	0.923	1	0.75	0.998	0.999	0.821	0.994	0.816
			$\boldsymbol{\alpha}_2$	0.991	1	0.996	0.99	0.979	1	0.99	0.952	1	0.943
500	1000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	1	1	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
500	1000	1/2	$\boldsymbol{\alpha}_1$	1	1	1	1	1	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
500	1000	2/3	$\boldsymbol{\alpha}_1$	0.998	0.998	0.998	1	0.994	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
500	2000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	1	1	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
500	2000	1/2	$\boldsymbol{\alpha}_1$	1	1	1	1	1	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
500	2000	2/3	$\boldsymbol{\alpha}_1$	0.987	0.995	0.998	1	0.980	1	1	0.998	1	0.998
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1

and the number of iterations for binary response are recorded. In general, the computing times become larger as the sample size and the dimension of predictors increases. The algorithm converges in 5 iterations and it is not influenced by the sample sizes and the dimension of the predictors, which also shows the high efficiency of the proposed method.

Example 3.2.1.3. This example is designated to examine the performance of the proposed procedures for count response. Given $\{U, \mathbf{x}\}$, we generate a count response from a Poisson distribution with mean $\lambda(U, \mathbf{x})$ defined below.

$$\log\{\lambda(U, \mathbf{x})\} = \alpha_1(U)X_1 + \alpha_2(U)X_2 + \alpha_3(U)X_3 + \alpha_4(U)X_4. \tag{3.4.3}$$

Model (3.4.3) implies that $\alpha_j(\cdot) = 0$ for $j > 4$ and $\mathcal{M}_* = \{1, 2, 3, 4\}$. In this example, the $\alpha_j(\cdot)$s are set to be $\frac{1}{4}\alpha_j(\cdot)$ as those in Example 1.

Table 3.4.5. Computing times (Seconds) and the Number of Iterations for Binary Response

	Σ_1				Σ_2			
	$\boldsymbol{\alpha}_1$		$\boldsymbol{\alpha}_2$		$\boldsymbol{\alpha}_1$		$\boldsymbol{\alpha}_2$	
ρ	Time	Iterations	Time	Iterations	Time	Iterations	Time	Iterations
	$(n,p) = (300, 1000)$							
1/3	15.65(2.51)	5(1)	13.18(2.37)	4(1)	12.36(1.69)	4(1)	14.52(2.62)	4(0)
1/2	17.39(2.56)	4(0)	8.17(0.28)	3(0)	14.70(2.39)	4(1)	14.48(2.67)	4(0)
2/3	15.44(2.39)	4(0)	9.19(1.75)	3(0)	14.55(1.98)	4(1)	16.76(3.19)	4(1)
	$(n,p) = (300, 2000)$							
1/3	23.63(4.09)	5(1)	19.80(3.31)	4(1)	17.76(3.55)	4(1)	16.93(3.21)	4(1)
1/2	17.70(1.08)	4(0)	13.54(0.39)	3(0)	22.61(4.13)	5(1)	18.79(3.60)	4(1)
2/3	16.94(1.94)	4(0)	13.46(0.64)	3(0)	22.24(3.89)	5(1)	21.50(3.56)	4(1)
	$(n,p) = (500, 1000)$							
1/3	75.23(11.43)	5(0)	50.36(8.00)	4(0)	55.09(8.95)	5(1)	55.03(7.53)	5(1)
1/2	64.40(8.98)	4(0)	33.64(3.32)	3(0)	62.36(8.52)	5(1)	56.10(9.03)	5(1)
2/3	55.52(8.34)	4(0)	31.63(3.18)	3(0)	63.35(8.16)	5(1)	56.07(9.19)	5(1)
	$(n,p) = (500, 2000)$							
1/3	112.07(18.07)	5(0)	57.70(4.09)	4(0)	70.14(12.46)	5(1)	71.20(10.52)	5(1)
1/2	75.85(13.67)	4(0)	49.28(7.43)	3(0)	69.76(11.67)	5(1)	70.23(12.71)	5(1)
2/3	78.53(11.51)	4(0)	44.31(3.67)	3(0)	79.09(13.66)	5(1)	72.74(11.21)	5(1)

In this example, we consider $p = 1000$ and 2000, and the sample size $n = 300$, and 500. All the simulation results are based on 1000 replications, and are summarized in Tables 3.4.6—3.4.7.

Table 3.4.6 shows the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ for the count responses. In most cases, the values of $\mathcal{P}_j$s and $\mathcal{P}_a$ are very close to one, regardless of whether there exists the marginal independence. We find that if there exists a significant difference between the absolute values of the coefficient functions, our proposed method can easily detect the one with the larger absolute value, but sometimes fails to detect others, which makes some values of $\mathcal{P}_a$ small. In general, our new procedure performs better when the sample size increases and the dimensionality decreases, which is consistent to the sure screening property of the new procedure. In addition, the new procedure has a better performance with smaller ρ's, which is happened in both linear and logistic setting, this is also because our new method mistakenly selects some unimportant predictors due to their high correlations with the true ones.

The computing efficiency of the proposed procedure for count responses can be seen from Table 3.4.7. Compared to the binary response, the computing time is

relatively shorter. In general, the computing times also become larger as n and p increases. The algorithm converges in fewer steps than the binary case, which indicates the high efficiency of the proposed method when dealing with count responses.

Table 3.4.6. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ for Count Response

				$\Sigma = \Sigma_1$					$\Sigma = \Sigma_2$				
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$
300	1000	1/3	$\boldsymbol{\alpha}_1$	0.982	0.976	0.978	0.983	0.942	0.998	0.998	0.983	0.989	0.975
			$\boldsymbol{\alpha}_2$	0.998	0.999	1	0.997	0.996	1	0.998	0.998	0.998	0.995
300	1000	1/2	$\boldsymbol{\alpha}_1$	0.945	0.941	0.928	0.989	0.842	0.999	1	0.884	0.994	0.883
			$\boldsymbol{\alpha}_2$	0.982	0.988	0.994	0.98	0.95	1	0.981	0.979	0.999	0.968
300	1000	2/3	$\boldsymbol{\alpha}_1$	0.815	0.848	0.808	0.979	0.554	0.993	0.998	0.622	0.994	0.617
			$\boldsymbol{\alpha}_2$	0.866	0.917	0.894	0.852	0.626	1	0.825	0.793	0.997	0.703
300	2000	1/3	$\boldsymbol{\alpha}_1$	0.965	0.966	0.956	0.973	0.895	0.998	1	0.966	0.97	0.955
			$\boldsymbol{\alpha}_2$	0.987	0.994	0.997	0.989	0.976	1	0.99	0.99	0.999	0.987
300	2000	1/2	$\boldsymbol{\alpha}_1$	0.897	0.895	0.88	0.994	0.739	0.996	0.997	0.811	0.991	0.806
			$\boldsymbol{\alpha}_2$	0.962	0.982	0.985	0.964	0.909	0.999	0.95	0.938	0.997	0.913
300	2000	2/3	$\boldsymbol{\alpha}_1$	0.744	0.743	0.748	0.986	0.421	0.992	0.99	0.489	0.988	0.479
			$\boldsymbol{\alpha}_2$	0.811	0.879	0.858	0.806	0.534	1	0.694	0.676	0.995	0.54
500	1000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	1	1	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
500	1000	1/2	$\boldsymbol{\alpha}_1$	0.999	0.999	1	1	0.998	0.999	1	0.991	1	0.990
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
500	1000	2/3	$\boldsymbol{\alpha}_1$	0.989	0.983	0.991	1	0.965	0.999	1	0.958	1	0.958
			$\boldsymbol{\alpha}_2$	0.996	1	1	0.993	0.989	1	0.996	0.997	1	0.994
500	2000	1/3	$\boldsymbol{\alpha}_1$	1	1	1	1	1	1	1	1	1	1
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
500	2000	1/2	$\boldsymbol{\alpha}_1$	0.999	1	0.999	1	0.998	1	1	0.988	1	0.988
			$\boldsymbol{\alpha}_2$	1	1	1	1	1	1	1	1	1	1
500	2000	2/3	$\boldsymbol{\alpha}_1$	0.981	0.976	0.972	1	0.933	1	1	0.929	1	0.929
			$\boldsymbol{\alpha}_2$	0.988	0.995	0.996	0.994	0.974	1	0.987	0.979	1	0.973

Example 3.2.1.4 Effect of minimum model size. We also examine the effect of m for our new method. In this example, we consider the following scenario:
(1) $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2$;
(2) $n = 300, 500$, $p = 1000, 2000$ and $\rho = 0.5$;
(3) $m = 4, 5, 6, 7, 8, 9$.

We examine continuous, binomial and count responses in this example, and the simulation results are based on 100 replicates. The results are presented in Table 3.4.8, 3.4.9 and 3.4.10.

Table 3.4.7. Computing times (Seconds) and the Number of Iterations for Count Response

	Σ_1				Σ_2			
	$\boldsymbol{\alpha}_1$		$\boldsymbol{\alpha}_2$		$\boldsymbol{\alpha}_1$		$\boldsymbol{\alpha}_2$	
ρ	Time	Iterations	Time	Iterations	Time	Iterations	Time	Iterations
	$(n,p) = (300, 1000)$							
1/3	13.62(2.44)	4(1)	11.10(2.10)	4(1)	16.17(2.40)	5(1)	11.86(2.39)	4(1)
1/2	10.51(2.23)	4(1)	12.61(2.03)	3(1)	12.90(2.46)	5(1)	15.39(2.65)	5(1)
2/3	9.76(0.67)	3(0)	11.15(1.51)	3(0)	12.84(2.46)	5(1)	13.04(2.44)	5(1)
	$(n,p) = (300, 2000)$							
1/3	17.24(3.16)	4(1)	18.50(3.96)	4(1)	22.47(3.79)	5(1)	20.40(3.48)	5(1)
1/2	17.12(3.23)	4(1)	16.64(2.84)	4(1)	20.38(3.67)	5(1)	20.53(3.61)	5(1)
2/3	13.84(0.62)	3(0)	13.67(0.51)	3(0)	19.84(3.73)	5(1)	21.20(3.98)	5(1)
	$(n,p) = (500, 1000)$							
1/3	56.39(9.94)	4(1)	43.94(6.90)	4(1)	54.58(8.08)	5(1)	63.15(9.99)	5(1)
1/2	43.14(6.40)	4(0)	39.69(6.17)	4(1)	51.78(9.01)	5(1)	52.92(8.86)	5(1)
2/3	47.08(7.45)	4(1)	29.25(1.14)	3(0)	51.12(9.04)	5(1)	52.86(8.80)	5(1)
	$(n,p) = (500, 2000)$							
1/3	77.70(11.08)	4(1)	53.43(10.93)	4(1)	70.14(12.30)	5(1)	71.47(12.31)	5(1)
1/2	61.36(8.73)	4(0)	52.00(11.15)	4(1)	70.80(12.03)	5(1)	74.42(10.20)	5(1)
2/3	50.81(11.06)	4(1)	50.32(8.40)	3(0)	70.83(11.98)	5(1)	76.46(11.58)	6(1)

Table 3.4.8. The proportions of $\mathcal{P}_a$ for continuous response

				Σ_1					
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	m=4	m=5	m=6	m=7	m=8	m=9
300	1000	1/2	$\boldsymbol{\alpha}_2$	0.97	1	1	1	1	1
300	2000	1/2	$\boldsymbol{\alpha}_2$	0.96	1	1	1	1	1
500	1000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1
500	2000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1
				Σ_2					
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	m=4	m=5	m=6	m=7	m=8	m=9
300	1000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1
300	2000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1
500	1000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1
500	2000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1

Based on the results in the three tables, we can see that there is no significant relationship between m and the success probability. Sometimes the success probabilities under smaller m are larger than those under larger values of m. However, we need the value of m to be larger than the true model size. When $m = 4$, which is the true model size, the values of success probability for the count response are not satisfying. When m is larger than the true model size, we can always have great chance to select all the important variables.

Table 3.4.9. The proportions of $\mathcal{P}_a$ for binary response

				Σ_1					
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	m=4	m=5	m=6	m=7	m=8	m=9
300	1000	1/2	$\boldsymbol{\alpha}_2$	0.99	1	1	1	1	1
300	2000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1
500	1000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1
500	2000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1
				Σ_2					
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	m=4	m=5	m=6	m=7	m=8	m=9
300	1000	1/2	$\boldsymbol{\alpha}_2$	0.97	1	1	1	1	1
300	2000	1/2	$\boldsymbol{\alpha}_2$	1	0.99	0.99	1	1	1
500	1000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1
500	2000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1

Table 3.4.10. The proportions of $\mathcal{P}_a$ for count response

				Σ_1					
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	m=4	m=5	m=6	m=7	m=8	m=9
300	1000	1/2	$\boldsymbol{\alpha}_2$	0.75	0.93	0.98	0.99	0.99	0.96
300	2000	1/2	$\boldsymbol{\alpha}_2$	0.66	0.87	0.93	0.96	0.94	0.97
500	1000	1/2	$\boldsymbol{\alpha}_2$	0.98	1	1	1	1	1
500	2000	1/2	$\boldsymbol{\alpha}_2$	0.98	0.98	1	1	1	1
				Σ_2					
n	p	ρ	$\boldsymbol{\alpha}(\cdot)$	m=4	m=5	m=6	m=7	m=8	m=9
300	1000	1/2	$\boldsymbol{\alpha}_2$	0.92	0.99	0.98	0.93	0.98	0.99
300	2000	1/2	$\boldsymbol{\alpha}_2$	0.86	0.90	0.97	0.92	0.98	0.92
500	1000	1/2	$\boldsymbol{\alpha}_2$	0.97	0.99	1	1	1	1
500	2000	1/2	$\boldsymbol{\alpha}_2$	1	1	1	1	1	1

Example 3.2.1.5. This example is designed to examine the performance of HBIC tuning parameter selector. We set $n = 500$, $p = 1000, 2000$, $\Sigma = \Sigma_2$ with $\rho = 0.5$ and $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2$ is the coefficient functions. We set $C_n = \log(\log n)$ in HBIC, and compare the performance of HBIC with those of the AIC and BIC tuning parameter selectors. The following three criteria are used to evaluate the performances:

1. P: the probability that the true model is selected;
2. C: the number of correctly selected predictors from four active predictors;
3. I: the number of predictors incorrectly selected as active ones from all inactive predictors.

The simulation results based on 200 replications are summarized in Table 4.2.8.

Table 3.4.11. Comparing AIC, BIC and HBIC (mean and sd)

		Continuous response		Binary response		Count response	
		p=1000	p=2000	p=1000	p=2000	p=1000	p=2000
AIC	P	0.100	0.060	0.055	0.020	0.420	0.370
	C	4(0)	4(0)	4(0.100)	4(0)	4(0)	4(0.141)
	I	10.200(7.366)	9.850(7.262)	11.425(6.889)	13.63(6.030)	1.64(2.242)	2.030(2.901)
BIC	P	0.745	0.715	0.760	0.710	0.665	0.570
	C	4(0)	4(0)	4(0.571)	4(0)	4(0.262)	4(0.278)
	I	0.305(0.560)	0.325(0.549)	0.300(0.481)	0.220(0.503)	0.530(0.956)	0.720(1.161)
HBIC	P	0.970	0.975	0.915	0.710	0.700	0.620
	C	4(0)	4(0)	3.73(0.954)	4(0)	4(0)	4(0)
	I	0.030(0.171)	0.025(0.157)	0.005(0.171)	0.320(0.509)	0.600(1.143)	0.660(1.002)

Table 3.4.11 shows that the AIC, BIC and HBIC tuning parameter selectors can reduce model complexity significantly, while retain all active predictors. As seen from Table 3.4.11, the HBIC performs much better than the AIC and theBIC in terms of controlling the false positives in linear varying coefficient model. For the HBIC, the probability of obtaining the true model is close to one and the number of false positives is close to zero. For logistic model and Poisson model, the HBIC performs much better than the AIC and the BIC in terms of selecting the true model. The BIC also works well for logistic model and Poisson model, since the probabilities of obtaining the true model are very close to those of the HBIC.

3.4.2 An Application

We illustrate the proposed methodology by an empirical analysis of a subset of data collected the Framingham Heart Study (FHS, for short) See Dawber, Meadors, and Moore (1951) and Jaquish (2007) for details about FHS. The data subset consists of data for 977 subjects. Of interest is to investigate the impact of dynamic genetic effects on obesity. In our analysis, we focus on nonrare SNPs. Here, nonrate SNPs are referred to those SNP whose the minor allele frequency of a SNP is great than 0.05. In our analysis, we include 4395 nonrare SNPs with missing rates being less than 0.02. Define the response variable to be 1 if this subject's BMI is greater than 25 and 0 oterwise. The goal is to identify the SNPs strongly associated with the obesity. To examine the dynamic (age-dependent) effect of SNPs and gender

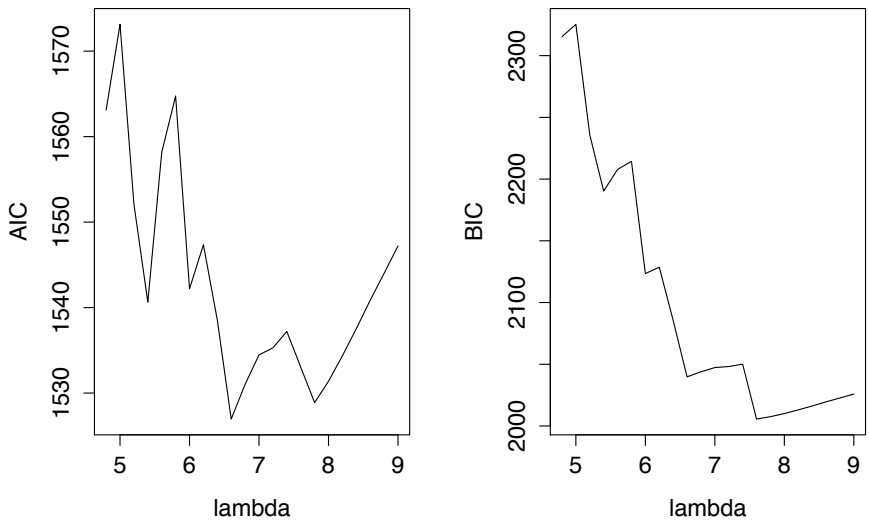

Figure 3.1. AIC and BIC versus λ

on obesity. We consider a logistic varying coefficient models with u being age, and 8791 covariates since for each SNP, both dominant effect and additive effect are considered, in addition to include gender as a covariate in our analysis. This leads to high-dimensional logistic varying coefficient model with the sample size $n = 977$.

We first apply the proposed screening procedure to the logistic varying coefficient model with the number of knots being $d_n = [\log(n)] = 6$ and $m = [(n/d_n)/\log(n/d_n)] = 28$ where $[a]$ indicates the integer part of a. Note that the gender variable is not subject to screening. Thus, there are total 29 variables after screening.

We further apply SCAD (Fan and Li, 2001) (**group SCAD**) to the model obtained from the screening procedure. Both Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used to select the tuning parameter in the SCAD. Figure 3.1 provides the plots of AIC and BIC scores versus λ. The optimal values for λ of SCAD-AIC and SCAD-BIC are 6.6 and 7.6, respectively. The SCAD-AIC selects a model with 13 SNPs, while the SCAD-BIC selects a model with 12 SNPs. All SNPs selected by the SCAD-BIC are selected by the SCAD-AIC, which also selects SS66164135_A. Figures 3.2 and 3.3 depict the plots

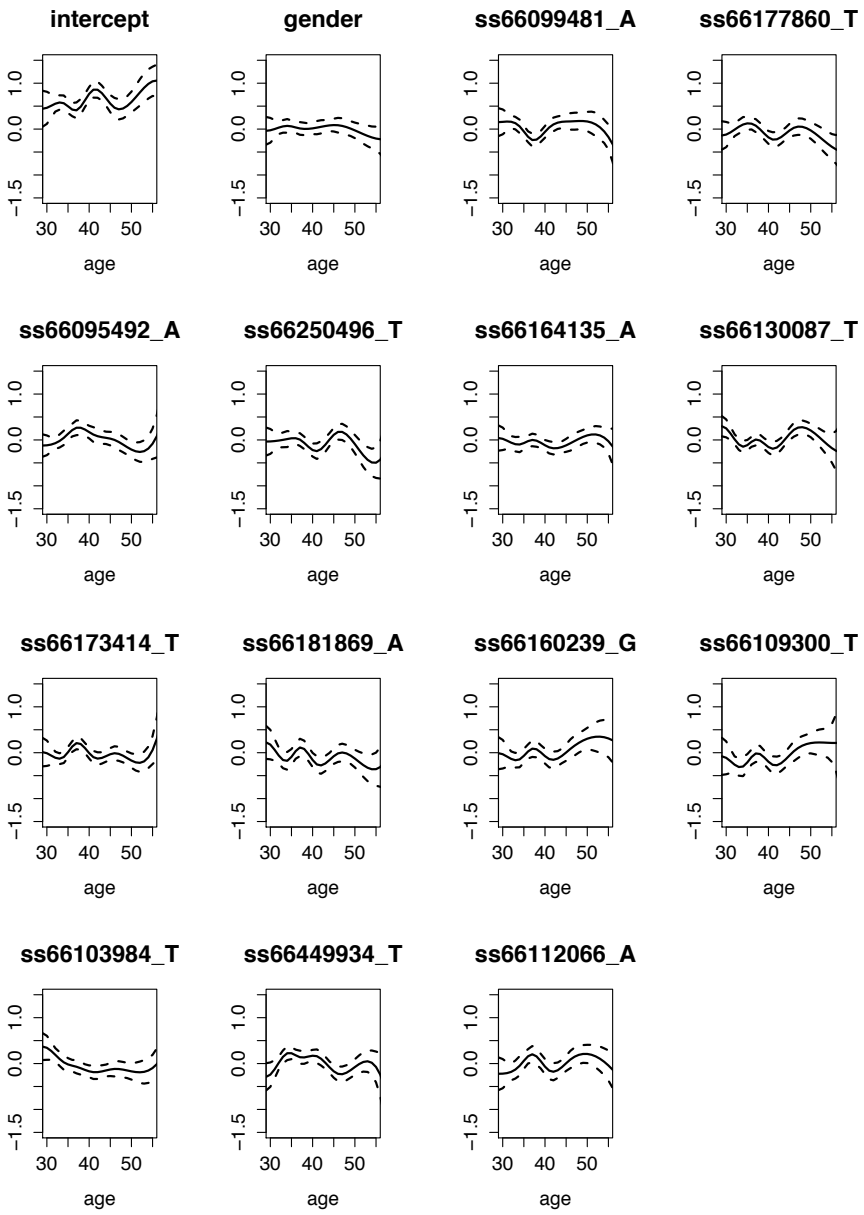

Figure 3.2. Estimated Coefficient Functions for $\lambda = 6.6$

of the estimated coefficient functions, and their pointwise confidence intervals.

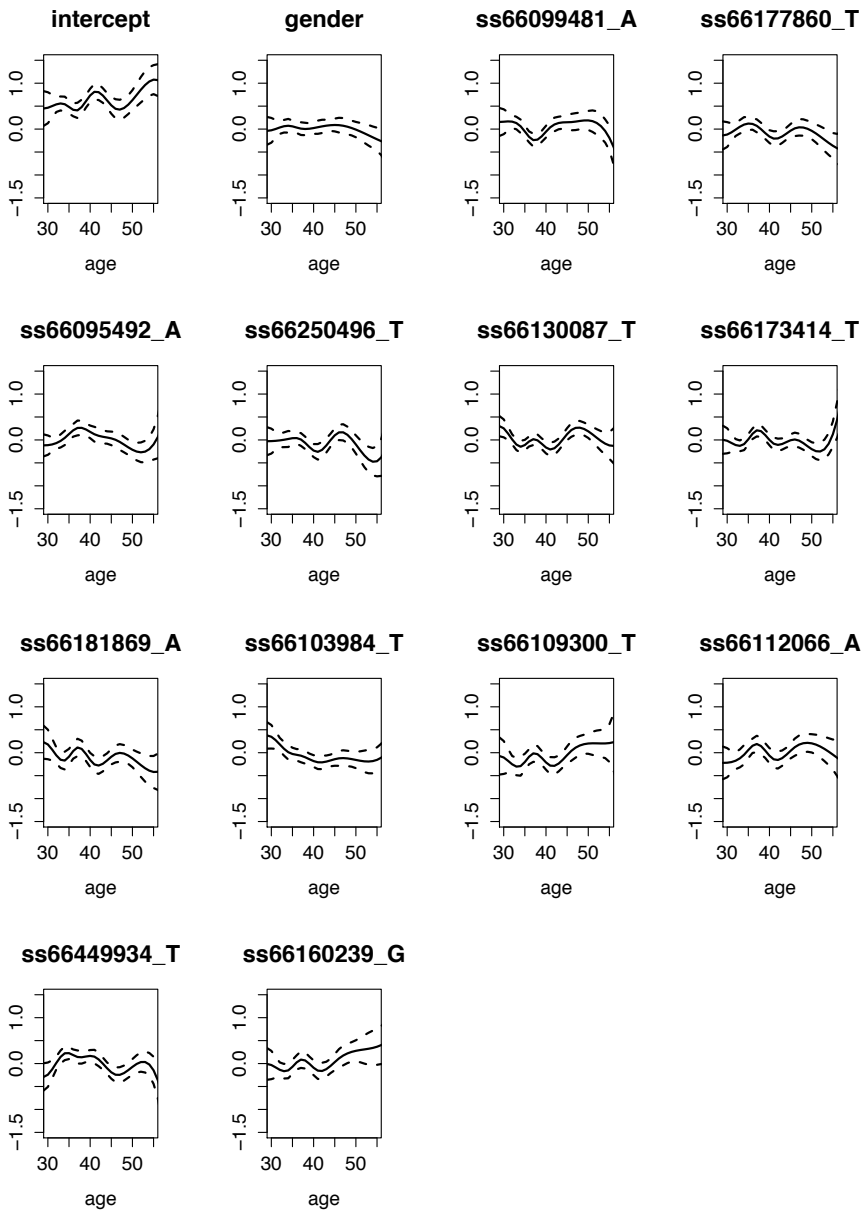

Figure 3.3. Estimated Coefficient Functions for $\lambda = 7.6$

3.5 Discussions

We have proposed a sure joint screening (SJS) procedure for feature screening in the generalized varying-coefficient models with ultrahigh dimensional covariates. The proposed SJS is distinguished from the existing SIS in that SJS is based on the joint likelihood of potential candidate features. We propose an effective algorithm

to carry out the feature screening procedure, and show that the proposed algorithm possesses an ascent property. We study the sample property of GVCM-SJS, and establish the sure screening property for GVCM-SJS.

Theorem 3.2.1 ensures the ascent property of the proposed algorithm under certain conditions, but it does not implies that the proposed algorithm converges to the global optimizer. If the proposed algorithm converges to a global maximizer of (3.2.5), then Theorem 3.3.1 shows that such a solution enjoys the sure screening property. We have simply set $m = [n/\log(n)]$ and $m = [(n/d_n)/\log(n/d_n)]$ in our numerical studies. It is of interest to derive a data-driven method to determine m and reduce false positive rate in the screening stage.

3.6 Technical Proof

Proof of Theorem 3.2.1. It follows by the Taylor expansion for the quasi-likelihood function $\ell(\boldsymbol{\gamma})$ at $\boldsymbol{\beta}$ lying within a neighbor of $\boldsymbol{\gamma}$ that

$$\ell(\boldsymbol{\gamma}) = \ell(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell'(\boldsymbol{\beta}) + \frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell''(\tilde{\boldsymbol{\beta}})(\boldsymbol{\gamma} - \boldsymbol{\beta}),$$

where $\tilde{\boldsymbol{\beta}}$ lies between $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. For $(\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell''(\tilde{\boldsymbol{\beta}})(\boldsymbol{\gamma} - \boldsymbol{\beta})$ term, we have

$$\begin{aligned} & (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \{-\ell''(\tilde{\boldsymbol{\beta}})\}(\boldsymbol{\gamma} - \boldsymbol{\beta}) \\ = \; & (\boldsymbol{\gamma} - \boldsymbol{\beta})^T W^{1/2}(\boldsymbol{\beta}) W^{-1/2}(\boldsymbol{\beta}) \{-\ell''(\tilde{\boldsymbol{\beta}})\} W^{-1/2}(\boldsymbol{\beta}) W^{1/2}(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) \\ \leqslant \; & \lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell''(\tilde{\boldsymbol{\beta}})\}W^{-1/2}(\boldsymbol{\beta})](\boldsymbol{\gamma} - \boldsymbol{\beta})^T W(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}), \end{aligned}$$

where $W(\boldsymbol{\beta})$ is a block diagonal matrix with $W_j(\boldsymbol{\beta})$ being a $d_{nj} \times d_{nj}$ matrix. Since $-\ell''(\boldsymbol{\beta})$ is non-negative definite, $\lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell''(\tilde{\boldsymbol{\beta}})\}W^{-1/2}(\boldsymbol{\beta})] \geqslant 0$ Thus, if

$$u > \lambda_{\max}[W^{-1/2}(\boldsymbol{\beta})\{-\ell''(\tilde{\boldsymbol{\beta}})\}W^{-1/2}(\boldsymbol{\beta})],$$

then

$$\ell(\boldsymbol{\gamma}) \geqslant \ell(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T \ell'(\boldsymbol{\beta}) - \frac{u}{2}(\boldsymbol{\gamma} - \boldsymbol{\beta})^T W(\boldsymbol{\beta})(\boldsymbol{\gamma} - \boldsymbol{\beta}) = h(\boldsymbol{\gamma}|\boldsymbol{\beta}).$$

Thus it follows that $\ell(\boldsymbol{\gamma}) \geqslant h(\boldsymbol{\gamma}|\boldsymbol{\beta})$ and $\ell(\boldsymbol{\beta}) = h(\boldsymbol{\beta}|\boldsymbol{\beta})$ by the definition of $h(\boldsymbol{\gamma}, \boldsymbol{\beta})$. The solution of $\partial h(\boldsymbol{\gamma}|\boldsymbol{\beta})/\partial\boldsymbol{\gamma} = 0$ is $\boldsymbol{\gamma} = \boldsymbol{\beta} + u^{-1}W(\boldsymbol{\beta})\ell'(\boldsymbol{\beta})$. Hence, under the conditions of Theorem 3.2.1, it follows that

$$\ell(\boldsymbol{\beta}^{*(t+1)}) \geqslant h(\boldsymbol{\beta}^{*(t+1)}|\boldsymbol{\beta}^{(t)}) \geqslant h(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^{(t)}) = \ell(\boldsymbol{\beta}^{(t)}).$$

The second inequality is due to the fact that $\tau(\{j : \|\boldsymbol{\beta}_j^{*(t+1)}\|_2 > 0\}) = \tau(\{j : \|\boldsymbol{\beta}_j^{(t)}\|_2 > 0\}) = m$, and $\boldsymbol{\beta}^{*(t+1)} = \arg\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}|\boldsymbol{\beta}^{(t)})$ subject to $\tau(\{j : \|\boldsymbol{\gamma}_j\|_2 > 0\}) \leqslant m$. By definition of $\boldsymbol{\beta}^{(t+1)}$, $\ell(\boldsymbol{\beta}^{(t+1)}) \geqslant \ell(\boldsymbol{\beta}^{*(t+1)})$ and $\tau(\{j : \|\boldsymbol{\beta}_j^{(t+1)}\|_2 > 0\}) = m$. This proves Theorem 1. □

Proof of Theorem 3.3.1. For a given model s, a subset of $\{1, \ldots, p\}$, let $\widehat{\boldsymbol{\alpha}}_s(\cdot)$ be the unrestricted maximum likelihood estimation of $\boldsymbol{\alpha}_s(\cdot)$ based on the spline approximation. It suffices to show that

$$Pr\left[\max_{s\in S_-^m} \ell\{\widehat{\boldsymbol{\alpha}}_s(U)\} \geqslant \min_{s\in S_+^m} \ell\{\widehat{\boldsymbol{\alpha}}_s(U)\}\right] \longrightarrow 0, \tag{A.1}$$

as $n \to \infty$.

We approximate $\alpha_j(U)$ by

$$\alpha_{nj}(U) = \sum_{k=1}^{d_n} \beta_{jk}\psi_{jk}(U) = \boldsymbol{\beta}_j^T\boldsymbol{\psi}_j(U), \quad j = 1, \cdots, p, \tag{A.2}$$

where $\psi_{jk}(U)$, $k = 1, \ldots, d_n$, are basis functions and d_n is the number of basis functions, which is allowed to increase with the sample size n.

Let $\mathcal{S}_j$ denote all functions that have the form $\sum_{k=1}^{d_n} \beta_{jk}\psi_{jk}(U)$ for a given set of basis $\{\psi_{jk}, k = 1, \ldots, d_n\}$. For $\alpha_{nj}(U)$, define the approximation error by

$$\rho_j(U) = \alpha_j(U) - \alpha_{nj}(U) = \alpha_j(U) - \sum_{k=1}^{d_n} \beta_{jk}\psi_{jk}(U), \quad j = 1, \ldots, p.$$

Let $\text{dist}(\alpha_j(\cdot), \mathcal{S}_j) = \inf_{\alpha_{nj}(U)\in\mathcal{S}_j} \sup_{U\in[a,b]} \|\rho_j(U)\|_2$, and take $\rho = \max_{1\leqslant j\leqslant p} \text{dist}(\alpha_j(\cdot), \mathcal{S}_j)$.

Let $\boldsymbol{\alpha}_n(U) = (\alpha_{n1}(U), \ldots, \alpha_{np}(U))^T$ and $\boldsymbol{\alpha}(U) = (\alpha_1(U), \ldots, \alpha_p(U))^T$. For any s,

$$\begin{aligned}\boldsymbol{\alpha}_s(U) &= \begin{pmatrix} \boldsymbol{\psi}_1(U) & & \\ & \ddots & \\ & & \boldsymbol{\psi}_s(U) \end{pmatrix}_{s\times sd_n} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_s \end{pmatrix}_{sd_n\times 1} + \begin{pmatrix} \rho_1(U) \\ \vdots \\ \rho_s(U) \end{pmatrix} \\ &\widehat{=} \Psi_s(U)\boldsymbol{\beta}_s + \rho_s(U),\end{aligned}$$

where $\Psi_s(U) = \mathrm{diag}(\boldsymbol{\psi}_1(U), \ldots, \boldsymbol{\psi}_s(U))$ with $\boldsymbol{\psi}_j(U) = (\psi_{j1}(U), \ldots, \psi_{jd_n}(U))$ and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_n})^T$, $j = 1, \ldots, s$.

For any $s \in S_-^m$, define $s' = s \cup s^* \in S_+^{2m}$. So, we have

$$\begin{aligned}&\ell\{\boldsymbol{\alpha}_{s'}(U)\} - \ell\{\boldsymbol{\alpha}^*_{s'}(U)\} \\ =\ &\ell\{\Psi_{s'}(U)\boldsymbol{\beta}_{s'} + \rho_{s'}(U)\} - \ell\{\Psi_{s'}(U)\boldsymbol{\beta}^*_{s'} + \rho^*_{s'}(U)\} \\ =\ &\ell\{\Psi_{s'}(U)\boldsymbol{\beta}_{s'}\} + \ell'\{\Psi_{s'}(U)\tilde{\boldsymbol{\beta}}_{s'}\}\rho_{s'}(U) - \ell\{\Psi_{s'}(U)\boldsymbol{\beta}^*_{s'}\} - \ell'\{\Psi_{s'}(U)\tilde{\boldsymbol{\beta}}^*_{s'}\}\rho^*_{s'}(U),\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{s'}$ and $\tilde{\boldsymbol{\beta}}^*_{s'}$ are two immediate values. Denote

$$\Delta_1 = \ell(\boldsymbol{\beta}_{s'}) - \ell(\boldsymbol{\beta}^*_{s'}), \quad \Delta_2 = \ell'(\tilde{\boldsymbol{\beta}}_{s'})\rho_{s'}(U), \quad \Delta_3 = \ell'(\tilde{\boldsymbol{\beta}}^*_{s'})\rho^*_{s'}(U).$$

Thus,

$$\ell\{\boldsymbol{\alpha}_{s'}(U)\} - \ell\{\boldsymbol{\alpha}^*_{s'}(U)\} = \Delta_1 + \Delta_2 - \Delta_3.$$

For Δ_2, by the Cauchy-Schwartz inequality, we have

$$E|\Delta_2| = E|\ell'(\tilde{\boldsymbol{\beta}}_{s'})\rho_{s'}(U)| \leqslant \sqrt{E\|\ell'(\tilde{\boldsymbol{\beta}}_{s'})\|^2}\sqrt{E\|\rho_{s'}(U)\|^2}.$$

According to the property of quasi-likelihood, we have

$$E\|\ell'(\tilde{\boldsymbol{\beta}}_{s'})\|^2 = \mathrm{tr}E\{\ell'(\tilde{\boldsymbol{\beta}}_{s'})\ell'(\tilde{\boldsymbol{\beta}}_{s'})^T\} = -\mathrm{tr}E\ell''(\tilde{\boldsymbol{\beta}}_{s'}).$$

By condition (C6) and Corollary 1 in Wei, Huang, and Li (2011), it follows $\Delta_2 = o_p(1)$. Similarly Δ_2, we have $\Delta_3 = o_p(1)$.

Next, we consider Δ_1. By Wedderburn (Part 5, 1974), the quasi-score function

of $\boldsymbol{\beta}_s$ is given by

$$S_n(\boldsymbol{\beta}_s) = \frac{\partial \ell(\boldsymbol{\beta}_s)}{\partial \boldsymbol{\beta}_s} = \sum_{i=1}^{n} \frac{\mu'(\mathbf{z}_{is}^T \boldsymbol{\beta}_s)}{V(\mathbf{z}_{is}^T \boldsymbol{\beta}_s)} [Y_i - E(Y_i|\mathbf{z}_i)] \mathbf{z}_{is},$$

where $\mu'(t)$ is the first-order derivative of $\mu(t)$. Let $H_n(\boldsymbol{\beta}_s) = -\partial^2 \ell(\boldsymbol{\beta}_s)/\partial \boldsymbol{\beta}_s \partial \boldsymbol{\beta}_s^T$ be the Hessian matrix of $\ell(\boldsymbol{\beta}_s)$ corresponding to $\boldsymbol{\beta}_s$.

Under (C3), we consider $\boldsymbol{\beta}_{s'}$ close to $\boldsymbol{\beta}_{s'}^*$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\| = w_1 d_n n^{-\tau_1}$ for some $w_1, \tau_1 > 0$. Clearly, when n is sufficiently large, $\boldsymbol{\beta}_{s'}$ falls into a neighborhood of $\boldsymbol{\beta}_{s'}^*$, so that condition (C6) becomes applicable. Thus, it follows by Condition (C6) and the Cauchy-Schwarz inequality that, we have

$$\begin{aligned}
\Delta_1 &= \ell(\boldsymbol{\beta}_{s'}) - \ell(\boldsymbol{\beta}_{s'}^*) \\
&= [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*]^T S_n(\boldsymbol{\beta}_{s'}^*) - (1/2)[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*]^T H_n(\tilde{\boldsymbol{\beta}}_{s'})[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*] \\
&\leqslant [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*]^T S_n(\boldsymbol{\beta}_{s'}^*) - (C_1/2) n d_n^{-1} \|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\|_2^2 \\
&\leqslant w_1 d_n n^{-\tau_1} \|S_n(\boldsymbol{\beta}_{s'}^*)\|_2 - (C_1/2) d_n^{-1} w_1^2 d_n^2 n^{1-2\tau_1}, \qquad \text{(A.3)}
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{s'}$ is an intermediate value between $\boldsymbol{\beta}_{s'}$ and $\boldsymbol{\beta}_{s'}^*$. Thus, we have

$$\begin{aligned}
Pr\{\ell(\boldsymbol{\beta}_{s'}) - \ell(\boldsymbol{\beta}_{s'}^*) \geqslant 0\} &\leqslant Pr\{\|S_n(\boldsymbol{\beta}_{s'}^*)\|_2 \geqslant (C_1 w_1/2) n^{1-\tau_1}\} \\
&\leqslant \sum_{j \in s'} Pr\{S_{nj}^2(\boldsymbol{\beta}_{s'}^*) \geqslant (2m)^{-1} (C_1 w_1/2)^2 n^{2-2\tau_1}\} \\
&\leqslant \sum_{j \in s'} \sum_{k=1}^{d_n} Pr\{S_{njk}^2(\boldsymbol{\beta}_{s'}^*) \geqslant (2m d_n)^{-1} (C_1 w_1/2)^2 n^{2-2\tau_1}\},
\end{aligned}$$

where

$$\Delta_4 = S_{njk}(\boldsymbol{\beta}_{s'}^*) = \sum_{i=1}^{n} \frac{\mu'(\mathbf{z}_{is'}^T \boldsymbol{\beta}_{s'})}{V(\mathbf{z}_{is'}^T \boldsymbol{\beta}_{s'})} [Y_i - E(Y_i|\mathbf{z}_i)] z_{ijk}.$$

We assume that $\mathbf{z}_i$ is marginally standardized. Since $\mu'(\mathbf{z}_{is'}^T \boldsymbol{\beta}_{s'})/V(\mathbf{z}_{is'}^T \boldsymbol{\beta}_{s'})$ is bounded by constant M under condition (C5), so according to the property of B-splines and condition (C7), there exists a positive constant $t_1 = t_0/M$ and g such that for all $|t| \leqslant t_1$, $E\{\exp(t \frac{\mu'(\mathbf{z}_{is'}^T \boldsymbol{\beta}_{s'})}{V(\mathbf{z}_{is'}^T \boldsymbol{\beta}_{s'})} [Y_i - E(Y_i|\mathbf{z}_i)] z_{ijk})\} < e^{gt^2/2}$.

By Petrov Exponential Inequalities (Lin and Bai, 2009, Page 68), we have

$$\begin{aligned} & Pr\{\Delta_4 \geqslant (C_1 w_1/2)(2d_n m)^{-1/2} n^{1-\tau_1}\} \\ \leqslant\ & Pr\{\Delta_4 \geqslant (C_1 w_1/2)(2w_2)^{-1/2} n^{-0.5\tau_2} n^{1-\tau_1} d_n^{-1/2}\} \\ \leqslant\ & \exp\left(-(c^2/2) n^{1-2\tau_1-\tau_2} d_n^{-1}\right), \end{aligned} \tag{A.4}$$

where $c = C_1 w_1/(2M\sqrt{2w_2})$. Also, by the same arguments, we have

$$Pr\{\Delta_4 \leqslant -(C_1 w_1/2)(2m)^{-1/2} n^{1-\tau_1}\} \leqslant \exp\left(-(c^2/2) n^{1-2\tau_1-\tau_2} d_n^{-1}\right), \tag{A.5}$$

The inequalities (A.4) and (A.5) imply that,

$$Pr\{\ell(\boldsymbol{\beta}_{s'}) \geqslant \ell(\boldsymbol{\beta}^*_{s'})\} \leqslant 4md_n \exp\left(-(c^2/2) n^{1-2\tau_1-\tau_2} d_n^{-1}\right).$$

So under condition (C4), we have

$$\begin{aligned} & Pr\left\{\max_{s\in S_-^m} \ell(\boldsymbol{\beta}_{s'}) \geqslant \ell(\boldsymbol{\beta}^*_{s'})\right\} \\ \leqslant\ & \sum_{s\in S_-^m} Pr\{\ell(\boldsymbol{\beta}_{s'}) \geqslant \ell(\boldsymbol{\beta}^*_{s'})\} \\ \leqslant\ & 4md_n p^m \exp\{-0.5c^2 n^{1-2\tau_1-\tau_2} d_n^{-1}\} \\ =\ & 4\exp\{\log m + m\log p - 0.5c^2 n^{1-2\tau_1-\tau_2} d_n^{-1} + 1/5\log n\} \\ \leqslant\ & 4\exp\{\log w_2 + (\tau_2+1/5)\log n + w_2 n^{\tau_2}\log p - 0.5c^2 n^{1-2\tau_1-\tau_2} d_n^{-1}\} \\ =\ & 4w_2 \exp\{\tau_2\log n + w_2 n^{\tau_2}\log p - 0.5c^2 n^{1-2\tau_1-\tau_2} d_n^{-1}\} \\ =\ & o(1) \quad \text{as} \quad n\to\infty. \end{aligned} \tag{A.6}$$

By Condition (C6), $\ell(\boldsymbol{\beta}_{s'})$ is concave in $\boldsymbol{\beta}_{s'}$, (A.6) holds for any $\boldsymbol{\beta}_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}^*_{s'}\| = w_1 d_n n^{-\tau_1}$.

For any $s \in S_-^m$, let $\breve{\boldsymbol{\beta}}_{s'}$ be $\widehat{\boldsymbol{\beta}}_s$ augmented with zeros corresponding to the elements in $s'\backslash s^*$ (i.e. $s' = \{s \cup (s^*\backslash s)\} \cup (s'\backslash s^*)$). By Condition (C1), it is seen that $\|\breve{\boldsymbol{\beta}}_{s'} - \boldsymbol{\beta}^*_{s'}\|_2 = \|\breve{\boldsymbol{\beta}}_{s^*\cup(s'\backslash s^*)} - \boldsymbol{\beta}^*_{s^*\cup(s'\backslash s^*)}\|_2 = \|\breve{\boldsymbol{\beta}}_{s^*\cup(s'\backslash s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \geqslant \|\boldsymbol{\beta}^*_{s^*\cup(s'\backslash s^*)} - \boldsymbol{\beta}^*_{s^*}\|_2 \geqslant \|\boldsymbol{\beta}^*_{s'\backslash s^*}\|_2 = w_1 d_n n^{-\tau_1}$. Consequently,

$$Pr\left\{\max_{s\in S_-^m} \ell(\widehat{\boldsymbol{\beta}}_s) \geqslant \min_{s\in S_+^m} \ell(\widehat{\boldsymbol{\beta}}_s)\right\} \leqslant Pr\left\{\max_{s\in S_-^m} \ell_p(\breve{\boldsymbol{\beta}}_{s'}) \geqslant \ell_p(\boldsymbol{\beta}^*_{s'})\right\} = o(1).$$

So, we have shown that

$$Pr\left[\max_{s\in S_-^m}\ell\{\widehat{\boldsymbol{\alpha}}_s(U)\}\geqslant\min_{s\in S_+^m}\ell\{\widehat{\boldsymbol{\alpha}}_s(U)\}\right]\longrightarrow 0,$$

as $n\to\infty$. The theorem is proved. □

Proof of Theorem 3.3.2. According to the definition of HBIC, for any model s, $HBIC(\tau(s))\leqslant HBIC(q)$ implies that

$$\begin{aligned}\ell(\widehat{\boldsymbol{\beta}}_s)-\ell(\widehat{\boldsymbol{\beta}}_{s*}) &\geqslant d_n\{\tau(s)-q\}\frac{C_n\log(d_np)}{2n}\\ &\geqslant -d_nq\frac{C_n\log(d_np)}{2n}.\end{aligned}\tag{A.7}$$

We show that the probability that (A.7) occurs at any $s\in S_-^m$ goes to 0. For any $s\in S_-^m$, let $\tilde{s}=s\cup s^*$. To consider those $\boldsymbol{\beta}_{\tilde{s}}$ near $\boldsymbol{\beta}^*_{\tilde{s}}$, we have

$$\ell(\boldsymbol{\beta}_{\tilde{s}})-\ell(\boldsymbol{\beta}^*_{\tilde{s}})=\{\boldsymbol{\beta}_{\tilde{s}}-\boldsymbol{\beta}^*_{\tilde{s}}\}^T\ell'(\boldsymbol{\beta}^*_{\tilde{s}})-\frac{1}{2}\{\boldsymbol{\beta}_{\tilde{s}}-\boldsymbol{\beta}^*_{\tilde{s}}\}^T[-\ell''(\tilde{\boldsymbol{\beta}}^*_{\tilde{s}})]\{\boldsymbol{\beta}_{\tilde{s}}-\boldsymbol{\beta}^*_{\tilde{s}}\},$$

for some $\tilde{\boldsymbol{\beta}}^*_{\tilde{s}}$ between $\boldsymbol{\beta}_{\tilde{s}}$ and $\boldsymbol{\beta}^*_{\tilde{s}}$. By Condition (C6),

$$\{\boldsymbol{\beta}_{\tilde{s}}-\boldsymbol{\beta}^*_{\tilde{s}}\}^T[-\ell''(\tilde{\boldsymbol{\beta}}^*_{\tilde{s}})]\{\boldsymbol{\beta}_{\tilde{s}}-\boldsymbol{\beta}^*_{\tilde{s}}\}\geqslant C_1d_n^{-1}n\|\boldsymbol{\beta}_{\tilde{s}}-\boldsymbol{\beta}^*_{\tilde{s}}\|^2.$$

Therefore,

$$\ell(\boldsymbol{\beta}_{\tilde{s}})-\ell(\boldsymbol{\beta}^*_{\tilde{s}})\leqslant\{\boldsymbol{\beta}_{\tilde{s}}-\boldsymbol{\beta}^*_{\tilde{s}}\}^T\ell'(\boldsymbol{\beta}^*_{\tilde{s}})-\frac{C_1}{2}d_n^{-1}n\|\boldsymbol{\beta}_{\tilde{s}}-\boldsymbol{\beta}^*_{\tilde{s}}\|^2.$$

Hence, for any $\boldsymbol{\beta}_{\tilde{s}}$ such that $\|\boldsymbol{\beta}_{\tilde{s}}-\boldsymbol{\beta}^*_{\tilde{s}}\|=w_1d_nn^{-\tau_1}$, we have

$$\ell(\boldsymbol{\beta}_{\tilde{s}})-\ell(\boldsymbol{\beta}^*_{\tilde{s}})\leqslant w_1d_nn^{-\tau_1}\|\ell'(\boldsymbol{\beta}^*_{\tilde{s}})\|-\frac{C_1}{2}d_n^{-1}n(w_1d_nn^{-\tau_1})^2.$$

By (A.4), (A.5) and (A.6), we can get

$$Pr\left\{\sup_{s\in S_-^m}\ell(\boldsymbol{\beta}_{\tilde{s}})\geqslant\ell(\boldsymbol{\beta}^*_{\tilde{s}})\right\}=o(1).$$

Now let $\breve{\boldsymbol{\beta}}_{\tilde{s}}$ be $\widehat{\boldsymbol{\beta}}_s$ augmented with zeros corresponding to the elements in $\tilde{s}\backslash s$.

It can be seen that

$$\|\breve{\boldsymbol{\beta}}_{\tilde{s}} - \boldsymbol{\beta}^*_{\tilde{s}}\| \geqslant \|\boldsymbol{\beta}^*_{s*\setminus s}\| = w_1 d_n n^{-\tau_1},$$

by (C3). Therefore, uniformly over $s \in S^m_-$ and with probability tending to 1,

$$Pr\left\{\sup_{s\in S^m_-} \ell(\widehat{\boldsymbol{\beta}}_{\tilde{s}}) \geqslant \ell(\boldsymbol{\beta}^*_{\tilde{s}})\right\} \leqslant Pr\left\{\sup_{s\in S^m_-} \ell(\breve{\boldsymbol{\beta}}_{\tilde{s}}) \geqslant \ell(\boldsymbol{\beta}^*_{\tilde{s}})\right\} = o(1).$$

Hence, the probability that (A.7) occurs at any $s \in S^m_-$ tends to 0 which is (3.3.13).

On the other hand, for $s \in S^m_+$, let $k = \tau(s) - q$. It suffices to consider a fixed k, since k takes only the values $1, \ldots, m-q$. By definition, $HBIC(\tau(s)) \leqslant HBIC(q)$ if and only if

$$\ell(\widehat{\boldsymbol{\beta}}_s) - \ell(\widehat{\boldsymbol{\beta}}_{s*}) \geqslant k d_n \frac{C_n \log(d_n p)}{2n}.$$

We show that, uniformly in $s \in S^m_+$ with $\tau(s) = k + q$, this inequality does not occur. For large n, by condition (C6),

$$\begin{aligned}
\ell(\widehat{\boldsymbol{\beta}}_s) - \ell(\widehat{\boldsymbol{\beta}}_{s*}) &\leqslant \ell(\widehat{\boldsymbol{\beta}}_s) - \ell(\boldsymbol{\beta}^*_s) \\
&\leqslant \{\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s\}^T \ell'(\boldsymbol{\beta}^*_s) - \frac{1}{2}\{\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s\}^T [-\ell''(\tilde{\boldsymbol{\beta}}^*_s)]\{\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s\} \\
&\leqslant \{\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s\}^T \ell'(\boldsymbol{\beta}^*_s) - \frac{1}{2} C_1 d_n^{-1} n \{\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s\}^T \{\widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s\}.
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}^*_s$ lies between $\widehat{\boldsymbol{\beta}}_s$ and $\widehat{\boldsymbol{\beta}}^*_s$. Denote $\Delta = \widehat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}^*_s$, and define

$$f(\Delta) = \Delta^T \ell'(\boldsymbol{\beta}^*_s) - \frac{1}{2} C_1 d_n^{-1} n \Delta^T \Delta.$$

So, we have

$$\frac{\partial f(\Delta)}{\partial \Delta} = \ell'(\boldsymbol{\beta}^*_s) - C_1 d_n^{-1} n \Delta = 0.$$

This implies that $f(\Delta)$ reaches its maximum at $\Delta = d_n \ell'(\widehat{\boldsymbol{\beta}}^*_s)/(C_1 n)$. Thus,

$$\ell(\widehat{\boldsymbol{\beta}}_s) - \ell(\widehat{\boldsymbol{\beta}}_{s*}) \leqslant \frac{1}{2}(C_1 n d_n^{-1})^{-1} \ell'(\boldsymbol{\beta}^*_s)^T \ell'(\boldsymbol{\beta}^*_s).$$

Hence, we show that, uniformly over $s \in S^m_+$ with $\tau(s) = k + q$,

$$\frac{1}{2}(C_1 n d_n^{-1})^{-1} \ell'(\boldsymbol{\beta}^*_s)^T \ell'(\boldsymbol{\beta}^*_s) \geqslant k d_n \frac{C_n \log(d_n p)}{2n},$$

occurs with diminishing probability. Thus, under conditions (C4) and (C6), by Markov inequality, for each $s \in S_+^m$, we have

$$\begin{aligned} & Pr\left[\frac{1}{2}(C_1 n d_n^{-1})^{-1}\ell'(\boldsymbol{\beta}_s^*)^T\ell'(\boldsymbol{\beta}_s^*) \geqslant k d_n \frac{C_n \log(d_n p)}{2n}\right] \\ = \; & Pr\left[\ell'(\boldsymbol{\beta}_s^*)^T\ell'(\boldsymbol{\beta}_s^*) \geqslant C_1 k C_n \log(d_n p)\right] \\ \leqslant \; & \frac{E[\ell'(\boldsymbol{\beta}_s^*)^T\ell'(\boldsymbol{\beta}_s^*)]}{C_1 k C_n \log(d_n p)} = \frac{E[\ell'(\boldsymbol{\beta}_s^*)^T\ell'(\boldsymbol{\beta}_s^*)]}{C_1 k C_n (\log(d_n) + n^\kappa)} \longrightarrow 0. \end{aligned}$$

the number of models in S_+^m is lower than p^κ, we have shown that

$$Pr\left[\frac{1}{2}(C_1 n d_n^{-1})^{-1}\ell'(\boldsymbol{\beta}_s^*)^T\ell'(\boldsymbol{\beta}_s^*) \geqslant k d_n \frac{C_n \log(d_n p)}{2n}, \forall s \in S_+^m\right] \longrightarrow 0,$$

This completes the proof. □

Chapter 4

New Test on High-Dimensional Mean Vectors With Consideration of Correlation Structure

4.1 New Test Method Considering the Linear Structure of Precision Matrix

Suppose that for $k = 1, 2$, $\mathbf{x}_{ki}$, $i = 1, \cdots, n_k$, is a random sample from a p-dimensional population with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}$, which is assumed to be finite and positive definite. We use $\bar{\mathbf{x}}_k = n_k^{-1} \sum_{i=1}^{n_k} \mathbf{x}_{ki}$ to be the sample mean of the k-th sample, and $\mathbf{S} = (n_1 + n_2 - 2)^{-1} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^T$ to be the pooled sample covariance matrix. The two-sample mean problem is to test

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{4.1.1}$$

In multivariate analysis, the Hotelling test statistic for the two sample problem is defined as

$$T_h = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

whose power function is an increasing function of $\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$, where $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. However, when $p > n_1 + n_2 - 2$, $\mathbf{S}$ is not invertible. Thus, Hotelling test is not well-defined for high dimensional data.

Bai and Sarandasa (1996) suggested the following test for the two sample problem

$$T_{bs} = \|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\|^2$$

and show that the power of T_{bs} is an increasing function of $\|\boldsymbol{\mu}_d\|^2/\sqrt{2\mathrm{tr}(\boldsymbol{\Sigma}^2)}$. It can be shown that

$$\|\boldsymbol{\mu}_d\|^2/\sqrt{2\mathrm{tr}(\boldsymbol{\Sigma}^2)} \leqslant \boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d \tag{4.1.2}$$

for any $\boldsymbol{\mu}_d$. This implies that the power of T_{bs} increases not as fast as Hotelling's type test because T_{bs} ignores the correlations among the variables. For a given constant symmetric positive definite $\mathbf{W}$ matrix, we consider the following test statistic

$$T = n(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{W}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{4.1.3}$$

to improve power over T_{bs}, where the weight matrix $\mathbf{W}$ is properly chosen and $n = n_1 + n_2$. Assume that for any $\mathbf{a}$ with $\|\mathbf{a}\| = 1$, $\sqrt{n}\mathbf{a}^T\boldsymbol{\Sigma}^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ asymptotically follows a normal distribution $N(\sqrt{n}\mathbf{a}^T\boldsymbol{\mu}_d, (n/n_1+n/n_2))$. As shown in the technical proof, the best choice of $\mathbf{W}$ should be proportional to $\boldsymbol{\Sigma}^{-1}$. This motivates us to model the precision matrix $\boldsymbol{\Sigma}^{-1}$ first.

4.1.1 Modelling precision matrix

If we decompose the covariance matrix $\boldsymbol{\Sigma} = \mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$, where $\mathbf{D}$ is the diagonal matrix with j-diagonal element being σ_{ii}, and $\mathbf{R}$ is the correlation matrix. The natural estimator of $\mathbf{D}$ is diag($\mathbf{S}$). Thus, it suffices to model $\mathbf{R}^{-1}$ in order to estimate $\boldsymbol{\Sigma}^{-1}$. Before we pursue further, let us examine some commonly-used correlation structures. Compound correlation matrix $\mathbf{R}_{cs} = (1-\rho)\mathbf{I}_p + \rho\mathbf{1}_p\mathbf{1}_p^T$, where $\mathbf{I}_p$ is the identity matrix and $\mathbf{1}_p$ is the p-dimensional column vector with all elements being 1. Then $\mathbf{R}_{cs}^{-1} = a_1\mathbf{A}_1 + a_2\mathbf{A}_2$ with $\mathbf{A}_1 = \mathbf{I}_p$, $\mathbf{A}_2 = \mathbf{1}\mathbf{1}_p^T$, $a_1 = (1-\rho)^{-1}$ and $a_2 = -\{(p-1)\rho+1\}\rho/(1-\rho)$. The correlation matrix $\mathbf{R}_{ar}$ from AR(1) model has its (i,j)-element $\rho^{|i-j|}$. Let $b_1 = (1+\rho^2)/(1-\rho^2)$, $b_2 = -\rho/(1-\rho^2)$ and

$b_3 = -\rho^2/(1-\rho^2)$. Then $\mathbf{R}_{ar}^{-1} = b_1\mathbf{A}_1 + b_2\mathbf{A}_2 + b_3\mathbf{A}_3$, where $\mathbf{A}_1 = \mathbf{I}_p$, $\mathbf{A}_2$ has 1 on the two main off-diagonals and 0 elsewhere, $\mathbf{A}_3$ has 1 on (p,p) and $(1,1)$, and 0 elsewhere. These examples motivate us to assume that $\mathbf{R}^{-1}$ can be represented as a linear combination of a set of matrix bases to achieve a parsimonious model for $\mathbf{R}$, that is

$$\mathbf{R}^{-1} = \theta_1\mathbf{A}_1 + \cdots + \theta_K\mathbf{A}_K.$$

Since the diagonal elements of $\mathbf{R}$ equal to 1, the linear representation implicitly imposes p constraints. To avoid solving constraint optimization problem, we propose a more flexible model for $\mathbf{R}^{-1}$. Let $\mathbf{A}_1, \cdots, \mathbf{A}_K$ be a set of known symmetric matrix bases and $\theta_1, \ldots, \theta_K$ be K unknown parameters. Assume that $\mathbf{R}^{-1}$ has the following structure

$$\mathbf{R}^{-1} = \theta_1\mathbf{A}_1 + \cdots + \theta_K\mathbf{A}_K. \tag{4.1.4}$$

It is worth pointing out that there always exists such a representation for the inverse of any correlation matrix. Since $\mathbf{R}^{-1}$ is a positive definite symmetric matrix, therefore $\mathbf{R}^{-1} = \sum_{j=1}^{p} r_{jj}^{(-1)}E_{jj} + \sum_{1\leqslant i<j\leqslant p} r_{ij}^{(-1)}E_{ij}$, where $\mathbf{R}^{-1} = (r_{ij}^{(-1)})$ and E_{ij} is a matrix with the (i,j)- and (j,i)-elements being 1 and all other elements being 0.

We first propose an estimation procedure for $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)^T$. Denote by $\widehat{\mathbf{R}}$ the sample correlation matrix, where $\widehat{\mathbf{R}} = [\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{S}[\mathrm{diag}(\mathbf{S})]^{-1/2}$ and $\mathrm{diag}(\mathbf{S})$ is the diagonal matrix from the diagonal elements of $\mathbf{S}$. We propose estimating $\boldsymbol{\theta}$ by minimizing the following quadratic loss

$$\min_{\boldsymbol{\theta}} \mathrm{tr}[\widehat{\mathbf{R}}(\theta_1\mathbf{A}_1 + \cdots \theta_K\mathbf{A}_K) - \mathbf{I}_p]^2. \tag{4.1.5}$$

Denote $\mathbf{B}$ to be a $K \times K$ matrix with (k,l)-element being $p^{-1}\mathrm{tr}(\widehat{\mathbf{R}}\mathbf{A}_k\widehat{\mathbf{R}}\mathbf{A}_l)$ and $\mathbf{b}$ to be a $K \times 1$ vector with k-the element being $p^{-1}\mathrm{tr}(\widehat{\mathbf{R}}\mathbf{A}_k)$. The minimizer of (4.1.5) has $\widehat{\boldsymbol{\theta}} = \mathbf{B}^{-1}\mathbf{b}$.

Theorem 4.1.1. *Suppose that* $\{\mathbf{x}_{ki}, i = 1, \cdots, n_k\}$, $k = 1, 2$, *is a random sample from a* p*-dimensional population* $\mathbf{x}^{(k)}$, *which can be represented as* $\mathbf{x}^{(k)} = \boldsymbol{\Sigma}^{1/2}\mathbf{w} + \boldsymbol{\mu}_k$, *where the components of* $\mathbf{w} = (w_1, \ldots, w_p)^T$ *are independent and identically distributed, and having the eighth moment with* $E(w_j) = 0$, $E(w_j)^2 = 1$, $E(w_i)^4 =$

κ. It follows that

$$\widehat{\theta}_k \to (1+y)^{-1}\theta_k, \quad k = 1, \ldots, K$$

where $y = \lim_{n\to\infty} p_n/(n-2)$ *and* $n = n_1 + n_2$.

Theorem 4.1.1 implies that if $y > 0$, $\widehat{\theta}_k$ is not consistent estimate for θ_k, but $[(n+p-2)/(n-2)]\widehat{\theta}_k$ is consistent. The joint distribution of θ_k's is derived by Theorem (4.1.2)

Theorem 4.1.2. *Under the conditions of Theorem 4.1.1, for any constants* $(\pi_1, \ldots, \pi_K)$, *we have*

$$\sigma^{-1}\left\{p\sum_{k=1}^{K}\pi_k[\widehat{\theta}_k - (1+y_{n-2})^{-1}\theta_k] - \nu\right\} \to N(0,1),$$

where $y = \lim_{n\to\infty} p_n/(n-2)$, $n = n_1 + n_2$,

$$\begin{aligned}
\nu &= \mathrm{tr}\mathbf{R}\mathbf{D}_1 - \frac{n}{(n-2)^2}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right] \\
&+\frac{n(n-1)}{4(n-2)^3}\mathrm{tr}\mathbf{D}_0\mathbf{D}_1 + \frac{3n(n-1)}{4(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^4\right] \\
&-\frac{1}{1+y_{n-2}}\left(\frac{n^2-3n+4}{(n-2)^2} + \frac{(n-4)p}{(n-2)^2} + \frac{(n_1-1)p}{n_1(n-2)^2} + \frac{(n_2-1)p}{n_2(n-2)^2} + \frac{\beta_w n}{(n-2)^2}\right)\mathrm{tr}\mathbf{R}\mathbf{D}_1 \\
&+(1+y_{n-2})^{-1}\frac{n[3(n-1)+p]}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right] \\
&+\frac{n(n-1)}{(1+y_{n-2})(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1\mathbf{R} + \beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^2\right] \\
&+(1+y_{n-2})^{-1}\frac{n}{(n-2)^3}\mathrm{tr}\mathbf{R}\mathbf{D}_1\left[2p + \beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right] \\
&-\frac{3n(n-1)p + 6n(n-1)(n-2)}{4(1+y_{n-2})(n-2)^4}\left[2\mathrm{tr}\mathbf{D}_1\mathbf{R} + \beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{D}_1\mathbf{R}\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right] \\
&-(1+y_{n-2})^{-1}\frac{3n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{4(n-2)^4}\left[2p + \beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right] \\
&-(1+y_{n-2})^{-1}\frac{2n(n-1)}{4(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{i=1}^{p}\mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{e}_i\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^4\right]
\end{aligned}$$

$$-(1+y_{n-2})^{-1}\frac{n(n-1)}{4(n-2)^3}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_j\Bigg[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 +\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\Bigg]$$

$$-(1+y_{n-2})^{-1}\frac{n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{4(n-2)^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\Bigg[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 +\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\Bigg]$$

$$-(1+y_{n-2})^{-1}\frac{n(n-1)(3n-6+p)}{4(n-2)^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{D}_1\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\Bigg[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 +\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\Bigg].$$

and

$$\begin{aligned}
&\sigma^2\\
=\;&\left(\frac{1+2y}{1-y}\right)^2\left(2n^{-1}\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2+\beta_w n^{-1}\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\right)\\
&+\left(\frac{1+2y}{1-y}\right)^2\frac{1}{n}\sum_{\ell_1=1}^{p}\sum_{\ell_2=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\mathbf{e}_{\ell_2}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_2}[2(\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{e}_{\ell_2})^2+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_1})^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_2})^2]\\
&-2\left(\frac{1+2y}{1-y}\right)^2 n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell[2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\cdot\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k]\\
&+(1+y)^{-2}(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)^2(2-2y-\beta_w y)+2(1+y)^{-2}y[n^{-1}\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2]\\
&+(1+y)^{-2}(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)^2(2n^{-1}\mathrm{tr}\mathbf{R}^2+\beta_w y)\\
&-\frac{2}{(1+y)^2}(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)\left(2n^{-1}\mathrm{tr}\mathbf{R}^2\mathbf{D}_1+\beta_w n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1\right)\\
&+\frac{2}{(1+y)^2}(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left[2\mathbf{e}_\ell^T\mathbf{R}^2\mathbf{e}_\ell+\beta_w\right].
\end{aligned}$$

$\mathbf{B}=\{p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_k\mathbf{R}\mathbf{A}_\ell+y_{n-2}(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_k)(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_\ell)\}_{k,\ell=1}^{K}$, $\mathbf{D}_1=\eta_1\mathbf{A}_1+\ldots+\eta_K\mathbf{A}_K$, $(\eta_1,\ldots,\eta_K)=(\pi_1,\ldots,\pi_K)\mathbf{B}^{-1}$, $\mathbf{D}_0$ *is the* $p\times p$ *dimensional matrix with the* (i,j)

element being $u_{ij} = 2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^3 + \beta_w\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_j)^2$, *and* $\mathbf{e}_k$ *is the* k*-column of the* $p \times p$ *identity matrix and* $\beta_w = \kappa - 3$.

In practical implementation, we may introduce a relative large number of $\mathbf{A}_k$s into the model (4.1.4) to reduce approximation error. For example, we may include bases for both the compound symmetric correlation structure and the AR correlation structure into model (4.1.4) if we are not sure which basis should be included. Thus, we introduce regularization method to reduce model complexity of model (4.1.4). Specifically, we consider the following penalized method

$$\mathrm{tr}\{\widehat{\mathbf{R}}(\theta_1\mathbf{A}_1 + \cdots\theta_K\mathbf{A}_K) - \mathbf{I}_p\}^2 + \sum_{k=1}^{K} p_\lambda(|\theta_k|), \tag{4.1.6}$$

where $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ. Minimizing (4.1.6) with respect to $\boldsymbol{\theta}$ results in a penalized estimator.

For a given estimate $\widehat{\boldsymbol{\theta}}$, define the estimate of $\mathbf{R}$ with linear structure (4.1.4) to be $\widehat{\mathbf{R}}_L = (\widehat{\theta}_1\mathbf{A}_1 + \cdots + \widehat{\theta}_K\mathbf{A}_K)^{-1}$. As a result, we estimate the population covariance matrix by

$$\widehat{\boldsymbol{\Sigma}} = [\mathrm{diag}(\mathbf{S})]^{1/2}\widehat{\mathbf{R}}_L[\mathrm{diag}(\mathbf{S})]^{1/2}.$$

The estimate of the population precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is

$$\widehat{\boldsymbol{\Omega}} = [\mathrm{diag}(\mathbf{S})]^{-1/2}\widehat{\mathbf{R}}_L^{-1}[\mathrm{diag}(\mathbf{S})]^{-1/2}, \tag{4.1.7}$$

where $\mathbf{S}$ is the pooled sample covariance matrix.

4.1.2 Limiting null distribution and power comparison

By replacing W in (4.1.3) by $\widehat{\boldsymbol{\Omega}}$, we obtain the test statistics

$$T_n = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\widehat{\boldsymbol{\Omega}}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2). \tag{4.1.8}$$

We next study the limiting distributions under both null hypothesis and alternative hypothesis.

Theorem 4.1.3. *Suppose that* $\{\mathbf{x}_{ki}, i = 1, \cdots, n_k\}$, $k = 1, 2$, *is a random sample from a p-dimensional population* $\mathbf{x}^{(k)}$, *which can be represented as* $\mathbf{x}^{(k)} = \boldsymbol{\Sigma}^{1/2}\mathbf{w} + \boldsymbol{\mu}_k$, *where the components of* $\mathbf{w} = (w_1, \ldots, w_p)^T$ *are independent and identically distributed and having the eighth moment with* $E(w_j) = 0$, $E(w_j)^2 = 1$, $E(w_i)^4 = \kappa$. *Denote* $\beta_w = \kappa - 3$, *then under* H_0 *and as* $y = \lim_{n\to\infty} p_n/(n-2)$, *it follows that*

$$\frac{nT_n - \widehat{c}\widehat{\mu}_0}{\widehat{c}\widehat{\sigma}_0} \xrightarrow{H_0} N(0,1),$$

where $\tilde{\mathbf{R}} = \{diag[\mathbf{R}_L]\}^{-1/2}\mathbf{R}_L\{diag[\mathbf{R}_L]\}^{-1/2}$,

$$\begin{aligned}\sigma_0^2 &= (2p + \kappa p n_1^{-1})\frac{n^2}{n_1^2} + (2p + \kappa p n_2^{-1})\frac{n^2}{n_2^2} + \frac{4n^2 p}{n_1 n_2} \\ &\quad + \left(\frac{n}{n_1^2} + \frac{n}{n_2^2}\right)\left[4\mathrm{tr}\mathbf{R}^2 + 2p\beta_w + 2\sum_{h=1}^{p}\mathbf{e}_h^T\mathbf{R}^2\mathbf{e}_h\mathbf{e}_h^T\mathbf{R}^{-1}\mathbf{e}_h + \beta_w\mathrm{tr}\mathbf{R}^{-1}\right],\end{aligned}$$

and

$$\begin{aligned}\mu_0 &= np(n_1^{-1} + n_2^{-1}) - \frac{n}{n-2}\left(n_1^{-1} + n_2^{-1}\right)\left[p + \beta_w\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_\ell\right] \\ &\quad + \frac{3n(n-1)}{4(n-2)^2}\left(n_1^{-1} + n_2^{-1}\right)\left[2p + \beta_w\sum_{h=1}^{p}\sum_{\ell=1}^{p}(\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right] \\ &\quad + \frac{n(n-1)}{4(n-2)^2}\left(n_1^{-1} + n_2^{-1}\right)\mathrm{tr}\mathbf{R}^{-1}\mathbf{A}_0.\end{aligned}$$

$\mathbf{e}_k$ *is the k-column of the identity matrix* $\mathbf{I}_p$, $\mathbf{A}_0$ *is a* $p \times p$ *matrix with* (h,ℓ)*-element being* $a_{h,\ell} = 2(\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell)^3 + \beta_w\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell\sum_{f=1}^{p}(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_h)^2(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2$, *and* $\widehat{\mu}_0$, $\widehat{\sigma}_0^2$ *and the estimate of* $\mathbf{A}_0$ *are obtained by replacing* $\mathbf{R}$ *in* μ_0, σ_0^2 *and* $\mathbf{A}_0$ *by* $\tilde{\mathbf{R}}$ *and* $\widehat{c} = \frac{1}{1+p/(n-2)}$.

This theorem proves that under the null hypothesis, the asymptotic distribution

of T_n is normal distribution. We next derive the limiting distribution of T_n under the alternative hypothesis.

Theorem 4.1.4. *Under conditions of Theorem 4.1.3 and* $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, *it follows that*

$$\frac{nT_n - c\mu_0 - cn\delta_n}{c\sqrt{\sigma_0^2 + 4n^2(n_1^{-1} + n_2^{-1})\delta_n}} \to N(0,1)$$

where $y = \lim_{n\to\infty} p_n/(n-2)$, $c = (1+y)^{-1}$, $\delta_n = \boldsymbol{\mu}_d^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d$ *with* $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. *Furthermore, the asymptotic power function of* T_n *is*

$$Q(\delta_n, \sigma_0|\alpha) = \Phi\left(\frac{-z_\alpha \sigma_0 + n\delta_n}{\sqrt{\sigma_0^2 + 4n^2(n_1^{-1} + n_2^{-1})\delta_n}}\right)$$

where $\Phi(\cdot)$ *is the cumulative distribution function of* $N(0,1)$ *and* $\Phi(-z_\alpha) = \alpha$.

Theorem 4.1.4 implies that T_n is an unbiased test (i.e., $Q(\delta_n, \sigma_0|\alpha) \geqslant \alpha$ for any δ_n and σ_0) since

$$\Phi\left(\frac{-z_\alpha \sigma_0 + n\delta_n}{\sqrt{\sigma_0^2 + 4n^2(n_1^{-1} + n_2^{-1})\delta_n}}\right) \geqslant \Phi(-z_\alpha).$$

In fact, Bai and Saranadasa (1996) and Chen and Qin (2010) also studied the two sample mean testing problem $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ *v.s.* $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. When the two population covariance matrices of $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are equal, Bai and Saranadasa (1996) and Chen and Qin (2010) have the same asymptotic powers as follows

$$Q_{bs}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}|\alpha) = \Phi\bigg(-z_\alpha + \frac{\|\boldsymbol{\mu}_d\|^2}{(n_1^{-1} + n_2^{-1})\sqrt{2\mathrm{tr}\boldsymbol{\Sigma}^2}}\bigg).$$

In fact, when $\delta_n = o(1)$ and n, p is large enough, we have

$$Q(\delta_n, \sigma_0|\alpha) \geqslant Q_{bs}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}|\alpha).$$

Especially, when $\boldsymbol{\Sigma}$ is the identity matrix, we approximately have $Q(\delta_n, \sigma_0|\alpha) = Q_{bs}(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}|\alpha)$.

4.2 Simulation

4.2.1 Performance of selecting basis matrices

In this part, we use our proposed method to estimate the precision matrix of $\mathbf{X}$ from $N(\mathbf{0}, \mathbf{\Sigma})$. In our motivation example, we consider two scenarios for $\mathbf{\Sigma} = (\sigma_{ij})$:

$\mathbf{\Sigma}_1$: Compound symmetric correlation structure: $\sigma_{ij} = 1$ if $i = j$ and ρ otherwise.

$\mathbf{\Sigma}_2$: Correlation structure from AR(1): $\sigma_{ij} = \rho^{|i-j|}$.

In the first scenario, $\mathbf{R}^{-1}$ can be written as $\theta_1\mathbf{A}_1 + \theta_2\mathbf{A}_2$, where $\mathbf{A}_1$ is the identity matrix, $\mathbf{A}_2 = \mathbf{1_p 1_p^T}$ is a matrix with 0 on the diagonal and 1 off the diagonal, $\theta_1 = -[(p-2)\rho+1]/\theta_0$ and $\theta_2 = \rho/\theta_0$, with $\theta_0 = (p-1)\rho^2 - (p-2)\rho - 1$. In the second scenario, $\mathbf{R}^{-1}$ can be written as a combination of three basis matrices: $\mathbf{R}^{-1} = \theta_1\mathbf{A}_1 + \theta_2\mathbf{A}_2 + \theta_3\mathbf{A}_3$, where$\mathbf{A}_1$ is the identity matrix, $\mathbf{A}_2$ is a matrix with 0 on the diagonal and 1 off the diagonal, $\mathbf{A}_3$ has the (p, p) and $(1, 1)$ elements being 1 and other elements being zeros, $\theta_1 = (1+\rho^2)/\theta_0$, $\theta_2 = -\rho/\theta_0$ and $\theta_3 = -\rho^2/\theta_0$ with $\theta_0 = 1 - \rho^2$.

First, we evaluate the performance of (4.1.6). Let the basis matrices be $\mathbf{A}_1, \ldots, \mathbf{A}_{[2n^{1/3}]}$, $\mathbf{A}_1, \ldots, \mathbf{A}_4$ are the basis matrix for compound symmetric and AR(1) covariance structure, and for $k \geqslant 5$, $\mathbf{A}_k$ has the $\{(i, j) : |i - j| = k - 3\}$ elements being 1 and other elements being zeros, where $[2n^{1/3}]$ is truncated integer of $2n^{1/3}$. (4.1.6) will be used to select the basis matrices $\mathbf{A}_j$. In the simulations, the following criteria will be reported.

- $||\widehat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}||_2$ where $\|\cdot\|_2$ is the quadratic norm;
- C: median of number of selected basis matrices;
- I: median of number of wrongly selected basis matrices.

In the simulation setup, ρ is taken as $\rho = 0.25, 0.5, 0.75$. The sample size is $n_1 = n_2 = 100, 200$ and the dimension is $p = 500, 1500$. The distribution of w_{ij} is Gaussian. We use the new estimate method under four different situations as follows:

1. Only the true basis matrices are included;

2. Redundant basis matrices are not removed;

3. Regularization method using MCP (Zhang, 2007) penalty;

4. Regularization method using SCAD (Fan and Li, 2001) penalty;

which are denoted by True, Full, MCP and SCAD. The simulation results based on 1000 replicates are given in Table 4.2.1-4.2.2. Table 4.2.1 show that when the full model is used to estimate $\boldsymbol{\theta}$, $\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}\|_2$ is much greater than TRUE, MCP and SCAD because the number of true basis matrices is smaller than the number of full models and full model leads to over-fit. Moreover, TRUE, MCP and SCAD have similar $\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}\|_2$ which shows that MCP and SCAD can almost select true basis matrices. Table 4.2.2 shows that the number of selected basis matrices is close to the number of true basis matrices, so the two selection methods perform very well.

Table 4.2.1. Precision Matrix Estimation

		Σ_1			Σ_2		
		n=100, p=500			n=100, p=500		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75
$\|\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\|\|_2$	True	0.0278(0.0208)	0.0828(0.0600)	0.2429(0.1854)	0.0922(0.0667)	0.0920(0.0620)	0.1257(0.0646)
	Full	0.1168(0.0548)	0.1801(0.0707)	0.4219(0.2052)	0.1392(0.0679)	0.1359(0.0556)	0.2668(0.1064)
	MCP	0.0275(0.0199)	0.0813(0.0589)	0.2422(0.1800)	0.0668(0.0051)	0.1214(0.0743)	0.1442(0.0870)
	SCAD	0.0274(0.0198)	0.0810(0.0583)	0.2405(0.1782)	0.0666(0.0051)	0.1451(0.0843)	0.1678(0.0916)
		n=200, p=500			n=200, p=500		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75
$\|\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\|\|_2$	True	0.0186(0.0143)	0.0577(0.0433)	0.1663(0.1267)	0.0458(0.0328)	0.0485(0.0301)	0.0695(0.0361)
	Full	0.0723(0.0243)	0.1275(0.0484)	0.2974(0.1308)	0.0764(0.0312)	0.0908(0.0292)	0.1966(0.0727)
	MCP	0.0193(0.0145)	0.0577(0.0432)	0.1656(0.1279)	0.0689(0.0271)	0.0584(0.0319)	0.0760(0.0443)
	SCAD	0.0193(0.0144)	0.0579(0.0427)	0.1669(0.1255)	0.0596(0.0195)	0.0585(0.0409)	0.0897(0.0542)
		n=100, p=1500			n=100, p=1500		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75
$\|\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\|\|_2$	True	0.0275(0.0195)	0.0849(0.0617)	0.2379(0.1817)	0.1526(0.1119)	0.1565(0.1111)	0.1771(0.1070)
	Full	0.1665(0.0993)	0.2125(0.1051)	0.4130(0.1915)	0.2157(0.1099)	0.1878(0.0988)	0.2667(0.1015)
	MCP	0.0280(0.0197)	0.0804(0.0599)	0.2435(0.2141)	0.0657(0.0031)	0.2080(0.1044)	0.2112(0.1023)
	SCAD	0.0279(0.0196)	0.0799(0.0590)	0.2432(0.2122)	0.0654(0.0030)	0.2197(0.1023)	0.2444(0.0973)
		n=200, p=1500			n=200, p=1500		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75
$\|\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}^{-1}\|\|_2$	True	0.0194(0.0141)	0.0565(0.0428)	0.1674(0.1267)	0.0797(0.0576)	0.0749(0.0549)	0.0864(0.0537)
	Full	0.0918(0.0490)	0.1320(0.0544)	0.2824(0.1343)	0.1099(0.0551)	0.1016(0.0445)	0.1720(0.0581)
	MCP	0.0192(0.0140)	0.0581(0.0436)	0.1703(0.1272)	0.0663(0.0022)	0.1013(0.0501)	0.1105(0.0537)
	SCAD	0.0192(0.0140)	0.0585(0.0434)	0.1718(0.1283)	0.0661(0.0022)	0.0994(0.0544)	0.1380(0.0543)

Table 4.2.2. Precision Matrix Estimation and Basis Selection

		Σ_1			Σ_2		
		n=100, p=500			n=100, p=500		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75
MCP	C	2	2	2	2	5	3
	I	0	0	0	0	2	0
SCAD	C	2	2	2	2	4	3
	I	0	0	0	0	1	0
		n=200, p=500			n=200, p=500		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75
MCP	C	2	2	2	5	5	3
	I	0	0	0	2	2	0
SCAD	C	2	2	2	5	5	3
	I	0	0	0	3	2	0
		n=100, p=1500			n=100, p=1500		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75
MCP	C	2	2	2	2	6	5
	I	0	0	0	0	3	2
SCAD	C	2	2	2	2	5	5
	I	0	0	0	0	3	2
		n=200, p=1500			n=200, p=1500		
	ρ	0.25	0.5	0.75	0.25	0.5	0.75
MCP	C	2	2	2	2	9	6
	I	0	0	0	0	6	3
SCAD	C	2	2	2	2	8	6
	I	0	0	0	0	5	3

4.2.2 Performance of the testing statistic T_n

Three scenarios for $\Sigma = (\sigma_{ij})$ are considered:

$\mathbf{\Sigma}_1$: Compound symmetric correlation structure with the diagonal elements being 1 and other elements being ρ;

$\mathbf{\Sigma}_2$: Correlation structure from AR(1): $\sigma_{ij} = \rho^{|i-j|}$;

$\mathbf{\Sigma}_3$: $0.5\mathbf{\Sigma}_1 + 0.5\mathbf{\Sigma}_2$.

The parameter ρ is taken as $\rho = 0.25, 0.50, 0.75$. Without loss of generality, $n_1 = n_2$ is assumed. The dimension is $p = 500, 1500$ and the sample size is $n_k = 100, 200, k = 1, 2$. w_{ij} is from $N(0, 1)$ or $Gamma(4, 2) - 2$. In the simulations, $\mu_1 = \mathbf{0_p}$ and $\mu_2 = c(\mathbf{1_{10}^T}, \mathbf{0_{p-10}^T})^\mathbf{T}$ and we consider $c = 0, 0.5, 1$. We compare the newly proposed method with other seven methods proposed before. In the simulation, we examine the performance of the proposed method under different settings as follows

1. Only the true basis matrices are included;

2. Redundant basis matrices are not removed;

3. Regularization method using MCP (Zhang, 2007) penalty;

4. Regularization method using SCAD (Fan and Li, 2001) penalty;

We compare those four methods with the other existing methods which are proposed by Bai and Saranadasa (1996), Chen and Qin (2010), Srivastava and Du (2008) with or without modification, Lopes, Jacob and Wainwright (2011, 2012) and Srivastava, Li and Ruppert (2014). The 11 different methods are denoted by New(true), New(full), New(MCP), New(SCAD), BS, CQ, SD1, SD2, LWJ1, LWJ2 and RAPTT. All the simulation results are based on $10,000$ replications and are summarized in Table 4.2.3-4.2.8.

Table 4.2.3 and Table 4.2.6 report the simulation results for the compound symmetric covariance structure for both the normal distribution and the gamma distribution. The new test methods return the type I error rate very well and the powers of the four new methods are extremely high when $c > 0$ and increase as ρ, c and n/p increase. All the new test methods outperform other existing methods. LWJ1, LWJ2 and RAPTT present the similar pattern to the proposed methods, however, LWJ2 fails to control the type I error rate. BS, CQ, SD1 and SD2 are affected by the value of ρ and their powers decrease significantly as ρ increases, since BS, CQ, SD1 and SD2 ignore the correlation among variables. In particular, when $c = 0.5$, the powers of these four test methods are always less than 0.3.

Table 4.2.4 and Table 4.2.7 present the simulation results for auto regressive correlation structure for the normal distribution and gamma distribution. Under this setting, the newly proposed method return the type I error rate very well and the powers of these new tests increase as c and n/p increase but decrease as ρ increases. LWJ1, LWJ2 and RAPTT tests also have the similar patterns as the new tests, but the new test methods outperform these methods. For the auto regressive covariance structure, the correlation between variables are weak if ρ is not large enough. BS, CQ, SD1 and SD2 beat the new methods in this setting. However, the performances of these methods are supposed to be good when the correlation between variables are weak since they neglect the correlation structure

Table 4.2.3. Power Comparison for $N_p(\boldsymbol{\mu}_1, \Sigma_1)$ and $N_p(\boldsymbol{\mu}_2, \Sigma_1)$

	$c = 0$			$c = 0.5$			$c = 1$		
ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
	$n = 100, p = 500$								
New(true)	0.0610	0.0463	0.0710	0.9979	1	1	1	1	1
New(full)	0.0613	0.0466	0.0720	0.9978	1	1	1	1	1
New(MCP)	0.0606	0.0456	0.0705	0.9979	1	1	1	1	1
New(SCAD)	0.0599	0.0427	0.0663	0.9978	1	1	1	1	1
BS	0.0676	0.0669	0.0680	0.1333	0.0954	0.0843	0.9570	0.2628	0.1665
CQ	0.0676	0.0669	0.0680	0.1333	0.0952	0.0843	0.9571	0.2630	0.1665
SD1	0.0311	0.0103	0.0026	0.0589	0.0145	0.0033	0.4731	0.0315	0.0061
SD2	0.0688	0.0681	0.0684	0.1367	0.0964	0.0845	0.9599	0.2653	0.1666
LWJ1	0.0528	0.0523	0.0525	0.4032	0.6189	0.9578	0.9958	1	1
LWJ2	0.0002	0.0002	0.0002	0.5188	0.9468	1	1	1	1
RAPTT	0.0521	0.0517	0.0515	0.4006	0.6269	0.9582	0.9943	1	1
	$n = 200, p = 500$								
New(true)	0.0517	0.0386	0.0560	1	1	1	1	1	1
New(full)	0.0509	0.0387	0.0564	1	1	1	1	1	1
New(MCP)	0.0513	0.0379	0.0550	1	1	1	1	1	1
New(SCAD)	0.0498	0.0344	0.0510	1	1	1	1	1	1
BS	0.0663	0.0662	0.0668	0.2496	0.1291	0.1014	1	0.9937	0.4166
CQ	0.0664	0.0663	0.0668	0.2494	0.1290	0.1014	1	0.9937	0.4169
SD1	0.0308	0.0088	0.0022	0.1075	0.0151	0.0030	1	0.0973	0.0104
SD2	0.0674	0.0666	0.0668	0.2522	0.1302	0.1017	1	0.9936	0.4179
LWJ1	0.0517	0.0518	0.0518	0.9697	0.9990	1	1	1	1
LWJ2	0.0019	0.0019	0.0019	1	1	1	1	1	1
RAPTT	0.0508	0.0508	0.0506	0.9735	0.9992	1	1	1	1
	$n = 100, p = 1500$								
New(true)	0.0677	0.0547	0.0704	0.9099	0.9925	1	1	1	1
New(full)	0.0681	0.0544	0.718	0.9090	0.9921	1	1	1	1
New(MCP)	0.0669	0.0524	0.0652	0.9091	0.9922	1	1	1	1
New(SCAD)	0.0652	0.0464	0.0575	0.9046	0.9913	1	1	1	1
BS	0.0721	0.0727	0.0732	0.0889	0.0807	0.0781	0.1661	0.1073	0.0966
CQ	0.0720	0.0727	0.0731	0.0891	0.0806	0.0781	0.1661	0.1071	0.0966
SD1	0.0225	0.0039	0.0004	0.0302	0.0044	0.0004	0.0512	0.0065	0.0006
SD2	0.0729	0.0733	0.0735	0.0913	0.0813	0.0782	0.1697	0.1086	0.0969
LWJ1	0.0476	0.0473	0.0474	0.1241	0.1809	0.4019	0.5606	0.8116	0.9943
LWJ2	0	0	0	0.0003	0.0112	0.5523	0.9518	1	1
RAPTT	0.0540	0.0535	0.0536	0.1348	0.1912	0.4151	0.5652	0.8086	0.9933
	$n = 200, p = 1500$								
New(true)	0.0552	0.0394	0.0643	1	1	1	1	1	1
New(full)	0.0554	0.0396	0.0644	1	1	1	1	1	1
New(MCP)	0.0543	0.0381	0.0623	1	1	1	1	1	1
New(SCAD)	0.0520	0.0323	0.0527	1	1	1	1	1	1
BS	0.0717	0.0719	0.0715	0.1059	0.0873	0.0817	0.4382	0.1612	0.1207
CQ	0.0717	0.0720	0.0715	0.1059	0.0873	0.0817	0.4384	0.1613	0.1207
SD1	0.0224	0.0050	0.0005	0.0347	0.0055	0.0006	0.1122	0.0105	0.0011
SD2	0.0732	0.0722	0.0716	0.1070	0.0877	0.0818	0.4441	0.1624	0.1211
LWJ1	0.0510	0.0510	0.0511	0.4069	0.6439	0.9745	0.9981	1.0000	1.0000
LWJ2	0.0001	0.0001	0.0001	0.4381	0.9669	1.0000	1.0000	1.0000	1.0000
RAPTT	0.0471	0.0472	0.0473	0.4123	0.6458	0.9714	0.9979	1.0000	1.0000

Table 4.2.4. Power Comparison for $N_p(\boldsymbol{\mu}_1, \Sigma_2)$ and $N_p(\boldsymbol{\mu}_2, \Sigma_2)$

	$c = 0$			$c = 0.5$			$c = 1$		
ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
	$n = 100, p = 500$								
New(true)	0.0641	0.0618	0.0543	0.8106	0.5503	0.4228	1	1	0.9996
New(full)	0.0639	0.0624	0.0547	0.8117	0.5502	0.4234	1	1	0.9995
New(MCP)	0.0641	0.0619	0.0538	0.8154	0.5511	0.4227	1	1	0.9996
New(SCAD)	0.0641	0.0611	0.0523	0.8152	0.5585	0.4198	1	1	0.9995
BS	0.0524	0.0580	0.0600	0.9491	0.8564	0.6158	1	1	1
CQ	0.0523	0.0580	0.0602	0.9492	0.8565	0.6158	1	1	1
SD1	0.0437	0.0471	0.0456	0.9387	0.8307	0.5606	1	1	1
SD2	0.0553	0.0611	0.0639	0.9488	0.8619	0.6225	1	1	1
LWJ1	0.0496	0.0500	0.0515	0.2661	0.2407	0.2119	0.9497	0.9159	0.8780
LWJ2	0.0001	0.0009	0.0048	0.1863	0.1575	0.1562	1	0.9998	0.9934
RAPTT	0.0499	0.0510	0.0489	0.2715	0.2449	0.2201	0.9511	0.9268	0.8871
	$n = 200, p = 500$								
New(true)	0.0588	0.0564	0.0529	0.9973	0.9363	0.8465	1	1	1
New(full)	0.0589	0.0568	0.0535	0.9973	0.9363	0.8470	1	1	1
New(MCP)	0.0589	0.0565	0.0527	0.9975	0.9360	0.8458	1	1	1
New(SCAD)	0.0589	0.0554	0.0514	0.9975	0.9361	0.8434	1	1	1
BS	0.0525	0.0539	0.0584	0.9999	0.9982	0.9608	1	1	1
CQ	0.0524	0.0540	0.0584	0.9999	0.9982	0.9608	1	1	1
SD1	0.0479	0.0463	0.0444	0.9999	0.9975	0.9476	1	1	1
SD2	0.0545	0.0554	0.0591	0.9999	0.9983	0.9615	1	1	1
LWJ1	0.0510	0.0501	0.0503	0.8035	0.6705	0.5403	1	1	0.9999
LWJ2	0.0025	0.0054	0.0106	0.9612	0.8414	0.6521	1	1	1
RAPTT	0.0498	0.0494	0.0502	0.8063	0.6843	0.5595	1	1	0.9999
	$n = 100, p = 1500$								
New(true)	0.0705	0.0664	0.0571	0.4952	0.3100	0.2294	1	0.9858	0.9309
New(full)	0.0710	0.0670	0.0574	0.4959	0.3103	0.2302	1	0.9847	0.9309
New(MCP)	0.0705	0.0666	0.0567	0.5002	0.3111	0.2285	1	0.9853	0.9306
New(SCAD)	0.0701	0.0656	0.0551	0.4997	0.3138	0.2241	1	0.9864	0.9228
BS	0.0531	0.0550	0.0572	0.6694	0.5332	0.3390	1	0.9999	0.9934
CQ	0.0531	0.0550	0.0571	0.6696	0.5332	0.3391	1	0.9999	0.9934
SD1	0.0388	0.0402	0.0419	0.6204	0.4811	0.2818	1	0.9999	0.9879
SD2	0.0552	0.0586	0.0592	0.6784	0.5422	0.3471	1	0.9999	0.9927
LWJ1	0.0499	0.0488	0.0502	0.0993	0.1009	0.1022	0.4026	0.3931	0.3719
LWJ2	0	0	0.0002	0.0001	0.0002	0.0044	0.5294	0.4724	0.4385
RAPTT	0.0527	0.0492	0.0503	0.1043	0.1066	0.1037	0.4129	0.3995	0.3851
	$n = 200, p = 1500$								
New(true)	0.0637	0.0619	0.0566	0.9004	0.6393	0.4985	1	1	1
New(full)	0.0643	0.0623	0.0563	0.9005	0.6396	0.4995	1	1	1
New(MCP)	0.0636	0.0623	0.0562	0.9052	0.6389	0.4974	1	1	1
New(SCAD)	0.0634	0.0613	0.0543	0.9051	0.6381	0.4922	1	1	1
BS	0.0551	0.0555	0.0565	0.9879	0.9420	0.7359	1.0000	1.0000	1.0000
CQ	0.0551	0.0555	0.0565	0.9879	0.9420	0.7359	1.0000	1.0000	1.0000
SD1	0.0459	0.0459	0.0436	0.9854	0.9316	0.6971	1.0000	1.0000	1.0000
SD2	0.0558	0.0560	0.0581	0.9873	0.9433	0.7404	1.0000	1.0000	1.0000
LWJ1	0.0487	0.0462	0.0487	0.2638	0.2397	0.2232	0.9641	0.9484	0.9265
LWJ2	0.0000	0.0000	0.0012	0.1065	0.1065	0.1298	1.0000	1.0000	0.9994
RAPTT	0.0487	0.0489	0.0491	0.2657	0.2487	0.2357	0.9653	0.9506	0.9355

Table 4.2.5. Power Comparison for $N_p(\boldsymbol{\mu}_1, \Sigma_3)$ and $N_p(\boldsymbol{\mu}_2, \Sigma_3)$

	$c = 0$			$c = 0.5$			$c = 1$		
ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
	$n = 100, p = 500$								
New(true)	0.0643	0.0527	0.0294	0.9408	0.8411	0.5719	1	1	1
New(full)	0.0648	0.0563	0.0383	0.9326	0.8040	0.5843	1	1	1
New(MCP)	0.0643	0.0527	0.0397	0.9423	0.8558	0.5884	1	1	1
New(SCAD)	0.0644	0.0521	0.0416	0.9420	0.8570	0.6047	1	1	1
BS	0.0654	0.0677	0.0680	0.2663	0.1319	0.1063	1	0.9272	0.4432
CQ	0.0653	0.0676	0.0681	0.2665	0.1320	0.1062	1	0.9273	0.4432
SD1	0.0481	0.0304	0.0169	0.1900	0.0590	0.0255	1	0.4625	0.0896
SD2	0.0673	0.0686	0.0694	0.2744	0.1346	0.1079	1	0.9292	0.4566
LWJ1	0.0459	0.0463	0.0467	0.3173	0.3401	0.3528	0.9761	0.9863	0.9894
LWJ2	0.0002	0.0006	0.0018	0.2837	0.3524	0.4032	1	1	1
RAPTT	0.0468	0.0508	0.0487	0.3123	0.3406	0.3665	0.9745	0.9841	0.9904
	$n = 200, p = 500$								
New(true)	0.0554	0.0448	0.0222	1	0.9992	0.9687	1	1	1
New(full)	0.0554	0.0505	0.0363	0.9999	0.9980	0.9733	1	1	1
New(MCP)	0.0550	0.0482	0.0370	1	0.9990	0.9736	1	1	1
New(SCAD)	0.0547	0.0445	0.0401	1	0.9993	0.9749	1	1	1
BS	0.0647	0.0661	0.0654	0.8724	0.2521	0.1588	1	1	1
CQ	0.0647	0.0661	0.0654	0.8724	0.2522	0.1587	1	1	0.9999
SD1	0.0491	0.0293	0.0135	0.7281	0.1091	0.0363	1	1	0.5085
SD2	0.0658	0.0666	0.0662	0.8774	0.2557	0.1594	1	1	1
LWJ1	0.0497	0.0502	0.0499	0.8965	0.8808	0.8327	1	1	1
LWJ2	0.0023	0.0037	0.0050	0.9945	0.9893	0.9628	1	1	1
RAPTT	0.0469	0.0471	0.469	0.8969	0.8878	0.8493	1	1	1
	$n = 100, p = 1500$								
New(true)	0.0687	0.0586	0.0367	0.6716	0.5170	0.2954	1	1	0.9966
New(full)	0.0692	0.0592	0.0459	0.6556	0.4739	0.3026	1	1	0.9950
New(MCP)	0.0570	0.0447	0.0339	0.7842	0.5248	0.3047	1	1	0.9965
New(SCAD)	0.0577	0.0424	0.0283	0.7912	0.5275	0.3286	1	1	0.9991
BS	0.0689	0.0689	0.0696	0.1054	0.0856	0.0794	0.4611	0.1605	0.1208
CQ	0.0690	0.0690	0.0696	0.1054	0.0856	0.0794	0.4611	0.1602	0.1208
SD1	0.0438	0.0221	0.0105	0.0671	0.0276	0.0114	0.2542	0.0495	0.0161
SD2	0.0705	0.0710	0.0708	0.1088	0.0883	0.0806	0.4767	0.1651	0.1232
LWJ1	0.0515	0.0523	0.0467	0.1131	0.1201	0.1358	0.4626	0.5313	0.6227
LWJ2	0.0000	0.0000	0.0001	0.0003	0.0010	0.0092	0.7636	0.8929	0.9465
RAPTT	0.0507	0.0518	0.0530	0.1149	0.1276	0.1429	0.4609	0.5326	0.6293
	$n = 200, p = 1500$								
New(true)	0.0583	0.0489	0.0257	0.9837	0.9273	0.6698	1	1	1
New(full)	0.0591	0.0533	0.0375	0.9801	0.8952	0.6916	1	1	1
New(MCP)	0.0582	0.0483	0.0389	0.9846	0.9394	0.6946	1	1	1
New(SCAD)	0.0590	0.0483	0.0409	0.9846	0.9367	0.7354	1	1	1
BS	0.0686	0.0636	0.0642	0.1616	0.1016	0.0868	1.0000	0.4438	0.2214
CQ	0.0686	0.0636	0.0642	0.1616	0.1016	0.0868	1.0000	0.4438	0.2212
SD1	0.0446	0.0218	0.0084	0.1040	0.0288	0.0128	0.9934	0.1038	0.0240
SD2	0.0700	0.0638	0.0648	0.1644	0.1038	0.0876	1.0000	0.4516	0.2228
LWJ1	0.0526	0.0508	0.0528	0.3140	0.3480	0.3852	0.9896	0.9932	0.9978
LWJ2	0.0000	0.0000	0.0006	0.2038	0.3150	0.4406	1.0000	1.0000	1.0000
RAPTT	0.0512	0.0544	0.0522	0.3283	0.3665	0.4075	0.9858	0.9926	0.9976

Table 4.2.6. Power Comparison for Gamma (4,2) with Σ_1

	$c=0$			$c=0.5$			$c=1$		
ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
	$n=100, p=500$								
New(true)	0.0580	0.0395	0.0634	0.9975	1	1	1	1	1
New(full)	0.0589	0.0396	0.0664	0.9976	1	1	1	1	1
New(MCP)	0.0573	0.0384	0.0625	0.9975	1	1	1	1	1
New(SCAD)	0.0561	0.053	0.0586	0.9973	1	1	1	1	1
BS	0.0699	0.0701	0.0696	0.1350	0.0973	0.0875	0.9405	0.2644	0.1693
CQ	0.0702	0.0704	0.0699	0.1357	0.0975	0.0877	0.9421	0.2662	0.1701
SD1	0.0306	0.0096	0.0028	0.0613	0.0135	0.0030	0.4948	0.0314	0.0057
SD2	0.0711	0.0706	0.0698	0.1392	0.0980	0.0875	0.9426	0.2677	0.1697
LWJ1	0.0543	0.0547	0.0547	0.3937	0.6207	0.9545	0.9956	0.9999	1
LWJ2	0	0.0001	0.0001	0.5232	0.9478	1	1	1	1
RAPTT	0.0497	0.0502	0.0500	0.4098	0.6317	0.9596	0.9936	1	1
	$n=200, p=500$								
New(true)	0.0505	0.0358	0.0574	1	1	1	1	1	1
New(full)	0.0504	0.0359	0.0588	1	1	1	1	1	1
New(MCP)	0.0501	0.0348	0.0560	1	1	1	1	1	1
New(SCAD)	0.0487	0.0325	0.0511	1	1	1	1	1	1
BS	0.0664	0.0682	0.0686	0.2590	0.1331	0.1053	1	0.9870	0.4300
CQ	0.0664	0.0682	0.0686	0.2597	0.1335	0.1056	1	0.9877	0.4316
SD1	0.0299	0.0074	0.0023	0.1149	0.0153	0.0032	0.9999	0.1030	0.0105
SD2	0.0672	0.0686	0.0686	0.2621	0.1336	0.1057	1	0.9859	0.4304
LWJ1	0.0529	0.0529	0.0526	0.9730	0.9994	1	1	1	1
LWJ2	0.003	0.003	0.003	1	1	1	1	1	1
RAPTT	0.0521	0.0520	0.0522	0.9728	1	1	1	1	1
	$n=100, p=1500$								
New(true)	0.0579	0.0459	0.0547	0.9041	0.9940	0.9791	1	1	1
New(full)	0.0579	0.0456	0.0674	0.9034	0.9936	0.9804	1	1	1
New(MCP)	0.0571	0.0441	0.0520	0.9031	0.9934	0.9797	1	1	1
New(SCAD)	0.0551	0.0394	0.0464	0.9004	0.9919	0.9785	1	1	1
BS	0.0721	0.0727	0.0723	0.0871	0.0804	0.0942	0.1641	0.1075	0.0942
CQ	0.0724	0.0731	0.0732	0.0874	0.0805	0.0777	0.1649	0.1079	0.0947
SD1	0.0223	0.0049	0.0012	0.0511	0.0055	0.0013	0.0729	0.0070	0.0016
SD2	0.0729	0.0736	0.0724	0.0886	0.0808	0.0774	0.1677	0.1090	0.0944
LWJ1	0.0498	0.0494	0.0498	0.1292	0.1919	0.4161	0.5648	0.8074	0.9944
LWJ2	0	0	0	0.0011	0.0109	0.5475	0.9515	1	1
RAPTT	0.0500	0.0501	0.0502	0.1337	0.1923	0.4153	0.5672	0.8078	0.9931
	$n=200, p=1500$								
New(true)	0.0513	0.0344	0.0594	0.9999	1	1	1	1	1
New(full)	0.0514	0.0346	0.0630	0.9999	1	1	1	1	1
New(MCP)	0.0506	0.0332	0.0568	0.9999	1	1	1	1	1
New(SCAD)	0.0481	0.0279	0.0470	0.9999	1	1	1	1	1
BS	0.0664	0.0664	0.0664	0.1007	0.0814	0.0758	0.4475	0.1564	0.1169
CQ	0.0664	0.0664	0.0665	0.1008	0.0814	0.0763	0.4491	0.1573	0.1176
SD1	0.0216	0.0031	0.0006	0.0314	0.0034	0.0007	0.1080	0.0074	0.0009
SD2	0.0670	0.0665	0.0664	0.1018	0.0817	0.0764	0.4577	0.1584	0.1173
LWJ1	0.0484	0.0484	0.0483	0.4081	0.6394	0.9729	0.9978	0.9999	1
LWJ2	0	0	0	0.4363	0.9595	1	1	1	1
RAPTT	0.0501	0.0498	0.0498	0.4046	0.6431	0.9736	0.9978	0.9998	1

Table 4.2.7. Power Comparison for Gamma (4,2) with Σ_2

	$c=0$			$c=0.5$			$c=1$		
ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
	$n=100, p=500$								
New(true)	0.0653	0.0617	0.0547	0.8110	0.5516	0.4168	1	1	0.9993
New(full)	0.0660	0.0617	0.0553	0.8112	0.5532	0.4190	1	1	0.9993
New(MCP)	0.0648	0.0612	0.0539	0.8172	0.5536	0.4175	1	1	0.9993
New(SCAD)	0.0647	0.0612	0.0525	0.8170	0.5599	0.4152	1	1	0.9993
BS	0.0527	0.0529	0.0568	0.9472	0.8545	0.6135	1	1	1
CQ	0.0535	0.0534	0.0569	0.9481	0.8554	0.6137	1	1	1
SD1	0.0441	0.0434	0.0405	0.9390	0.8327	0.5575	1	1	1
SD2	0.0572	0.0573	0.0612	0.9504	0.8618	0.6227	1	1	1
LWJ1	0.0499	0.0486	0.0470	0.2660	0.2388	0.2092	0.9471	0.9181	0.8792
LWJ2	0.0002	0.0008	0.0033	0.1837	0.1559	0.1544	1	0.9997	0.9937
RAPTT	0.0516	0.0520	0.0536	0.2751	0.2470	0.2182	0.9462	0.9228	0.8838
	$n=200, p=500$								
New(true)	0.0644	0.0624	0.0596	0.9978	0.9410	0.8463	1	1	1
New(full)	0.0649	0.0625	0.0594	0.9979	0.9421	0.8465	1	1	1
New(MCP)	0.0646	0.0623	0.0591	0.9978	0.9413	0.8457	1	1	1
New(SCAD)	0.0645	0.0617	0.0572	0.9978	0.9412	0.8427	1	1	1
BS	0.0568	0.0596	0.0609	0.9999	0.9988	0.9625	1	1	1
CQ	0.0570	0.0599	0.0612	0.9999	0.9988	0.9625	1	1	1
SD1	0.0526	0.0525	0.0465	0.9999	0.9981	0.9485	1	1	1
SD2	0.0599	0.0612	0.0631	0.9999	0.9987	0.9619	1	1	1
LWJ1	0.0513	0.0494	0.0509	0.7978	0.6711	0.5388	1	1	1
LWJ2	0.0030	0.0061	0.0111	0.9646	0.8420	0.6547	1	1	1
RAPTT	0.0500	0.0480	0.0510	0.8060	0.6852	0.5590	1	1	1
	$n=100, p=1500$								
New(true)	0.0695	0.0679	0.0599	0.4939	0.3075	0.2311	0.9999	0.9864	0.9377
New(full)	0.0702	0.0685	0.0603	0.4921	0.3099	0.2314	0.9999	0.9860	0.9371
New(MCP)	0.0694	0.0680	0.0596	0.4988	0.3096	0.2305	0.9999	0.9865	0.9373
New(SCAD)	0.0690	0.0676	0.0578	0.4978	0.3108	0.2256	0.9999	0.9872	0.9366
BS	0.0509	0.0502	0.0549	0.6690	0.5424	0.3442	1	1	0.9927
CQ	0.0517	0.0506	0.0550	0.6706	0.5438	0.3446	1	1	0.9927
SD1	0.0377	0.0383	0.0374	0.6248	0.4892	0.2848	1	1	0.9872
SD2	0.0529	0.0530	0.0578	0.6851	0.5541	0.3522	1	1	0.9920
LWJ1	0.0505	0.0492	0.0525	0.1022	0.1001	0.0936	0.4069	0.3891	0.3799
LWJ2	0	0	0.0001	0	0.0005	0.0035	0.5262	0.4733	0.4396
RAPTT	0.0531	0.0502	0.0465	0.1075	0.1040	0.1008	0.4045	0.3923	0.3823
	$n=200, p=1500$								
New(true)	0.0647	0.0629	0.0583	0.9025	0.6508	0.5075	1	1	1
New(full)	0.0643	0.0629	0.0589	0.9025	0.6511	0.5078	1	1	1
New(MCP)	0.0646	0.0628	0.0580	0.9067	0.6503	0.5063	1	1	1
New(SCAD)	0.0645	0.0618	0.0557	0.9065	0.6505	0.5018	1	1	1
BS	0.0548	0.0545	0.0549	0.9883	0.9439	0.7379	1	1	1
CQ	0.0549	0.0546	0.0549	0.9883	0.9439	0.7381	1	1	1
SD1	0.0449	0.0438	0.0432	0.9861	0.9330	0.6976	1	1	1
SD2	0.0547	0.0563	0.0558	0.9885	0.9440	0.7396	1	1	1
LWJ1	0.0510	0.0485	0.0484	0.2716	0.2491	0.2232	0.9656	0.9498	0.9308
LWJ2	0	0	0.0013	0.1151	0.1065	0.1279	1	1	0.9993
RAPTT	0.0534	0.0524	0.0500	0.2732	0.2526	0.2348	0.9670	0.9522	0.9346

Table 4.2.8. Power Comparison for Gamma (4,2) with Σ_3

	$c=0$			$c=0.5$			$c=1$		
ρ	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75
	$n=100, p=500$								
New(true)	0.0648	0.0530	0.0319	0.9417	0.8412	0.5675	1	1	1
New(full)	0.0644	0.0577	0.0419	0.9353	0.8003	0.5785	1	1	1
New(MCP)	0.0647	0.0530	0.0423	0.9433	0.8554	0.5831	1	1	1
New(SCAD)	0.0646	0.0526	0.0433	0.9428	0.8571	0.6009	1	1	1
BS	0.0731	0.0741	0.0744	0.2717	0.1396	0.1116	1	0.9162	0.4524
CQ	0.0735	0.0741	0.0748	0.2731	0.1401	0.1118	1	0.9186	0.4549
SD1	0.0549	0.0342	0.0181	0.1931	0.0661	0.0277	0.9998	0.4772	0.1004
SD2	0.0745	0.0748	0.0761	0.2805	0.1426	0.1141	1	0.9188	0.4644
LWJ1	0.0485	0.0493	0.0522	0.3100	0.3344	0.3566	0.9746	0.9835	0.9893
LWJ2	0.0001	0	0.0012	0.2845	0.3491	0.3988	1	1	1
RAPTT	0.0504	0.0496	0.0497	0.3162	0.3477	0.3702	0.9777	0.9871	0.9925
	$n=200, p=500$								
New(true)	0.0564	0.0469	0.0253	1	0.9990	0.9673	1	1	1
New(full)	0.0582	0.0507	0.0384	1	0.9986	0.9713	1	1	1
New(MCP)	0.0564	0.0487	0.0163	1	0.9986	0.8559	1	1	1
New(SCAD)	0.0560	0.0469	0.0140	1	0.9992	0.9028	1	1	1
BS	0.0670	0.0685	0.0692	0.8707	0.2667	0.1678	1	1	0.9998
CQ	0.0671	0.0687	0.0692	0.8719	0.2676	0.1682	1	1	0.9998
SD1	0.0522	0.0310	0.0162	0.7306	0.1147	0.0395	1	1	0.5238
SD2	0.0678	0.0688	0.0695	0.8734	0.2690	0.1704	1	1	0.9997
LWJ1	0.0491	0.0484	0.0493	0.8969	0.8850	0.8354	1	1	1
LWJ2	0.0024	0.0032	0.0070	0.9966	0.9898	0.9624	1	1	1
RAPTT	0.0525	0.0510	0.0511	0.8929	0.8841	0.8485	1	1	1
	$n=100, p=1500$								
New(true)	0.0683	0.0617	0.0412	0.6796	0.5276	0.2981	1	1	0.9965
New(full)	0.0679	0.0642	0.0503	0.6636	0.4822	0.3057	1	1	0.9961
New(MCP)	0.0566	0.0454	0.0371	0.7803	0.5311	0.3019	1	1	0.9979
New(SCAD)	0.0570	0.0441	0.0334	0.7901	0.5332	0.3287	1	1	0.9997
BS	0.0736	0.0730	0.0727	0.1089	0.0881	0.0842	0.4717	0.1737	0.1299
CQ	0.0740	0.0734	0.0730	0.1089	0.0889	0.0847	0.4740	0.1746	0.1306
SD1	0.0474	0.0247	0.0097	0.0734	0.0298	0.0109	0.2734	0.0550	0.0170
SD2	0.0751	0.0741	0.0735	0.1112	0.0898	0.0858	0.4862	0.1790	0.1313
LWJ1	0.0535	0.0529	0.0536	0.1121	0.1264	0.1379	0.4675	0.5267	0.6150
LWJ2	0	0	0	0.0002	0.0012	0.0098	0.7534	0.8866	0.9417
RAPTT	0.0568	0.0534	0.0503	0.1217	0.1295	0.1396	0.4778	0.5501	0.6251
	$n=200, p=1500$								
New(true)	0.0573	0.0467	0.0262	0.9864	0.9240	0.6733	1	1	1
New(full)	0.0573	0.0510	0.0389	0.9825	0.8926	0.6967	1	1	1
New(MCP)	0.0570	0.0462	0.0401	0.9868	0.9371	0.7009	1	1	1
New(SCAD)	0.0575	0.0466	0.0419	0.9870	0.9348	0.7392	1	1	1
BS	0.0674	0.0678	0.0679	0.1631	0.1028	0.0901	0.9998	0.4508	0.2226
CQ	0.0674	0.0678	0.0680	0.1635	0.1030	0.0902	0.9998	0.4518	0.2231
SD1	0.0430	0.0189	0.0083	0.1019	0.0298	0.0095	0.9848	0.1113	0.0255
SD2	0.0679	0.0682	0.0683	0.1651	0.1038	0.0908	0.9998	0.4593	0.2251
LWJ1	0.0506	0.0512	0.0487	0.3142	0.3567	0.4003	0.9867	0.9940	0.9969
LWJ2	0	0	0.0001	0.2031	0.3147	0.4329	1	1	1
RAPTT	0.0485	0.0476	0.0488	0.3179	0.3601	0.4051	0.9873	0.9945	0.9970

between variables. We can see that as ρ increases, the powers of these four methods also decrease significantly.

To make a fair comparison, we consider the mixture structure of compound symmetric and auto regressive structure, and the simulation results are reported in Table 4.2.5 and Table 4.2.8. The basis matrices we used for Σ_3 is the union of the basis matrices for Σ_1 and Σ_2, which are actually not the real basis matrices for Σ_3. However, the newly proposed methods outperform all other existing methods. The type I error rates are controlled very well and the power increases as c and n/p increase. For instance, when $(n, p, c) = (100, 1500, 0.5)$, all other methods almost have no powers.

In general, the power of the new test is better for the compound symmetric structure than auto regressive structure when other situations are same. For example when $(n, p, c) = (200, 1500, 0.5)$, the power for the compound symmetric structure and autoregressive structure are 0.905 and 0.52, respectively. Therefore, the performances of the new proposed methods are affected by the covariance structure of the variables.

The performances of the four different sorts of new proposed methods are very close in the simulation for Σ_1 amd Σ_2, since the true basis matrices are always considered in the estimation procedure. For Σ_3, the New (full) method controls the type I error rate better, but the when we compare the power of the results, other three methods beat the New(full). Hence, the regularization methods help increasing the power of the test when the true basis matrices are unknown. The fact shows the necessity of the regularization procedure in the new test technique.

4.3 Real data example

To examine the effectiveness of our newly proposed method, we apply our method to a high resolution microcomputed tomography dataset. The data were collected by the Center for Quantitative X-ray Imaging at The Pennsylvania State University, which contains the bone volume measured at different density levels in a genetic study. Our target is to detect the difference of bone density patterns for the mice with different genotypes. The data are normalized by dividing the bone

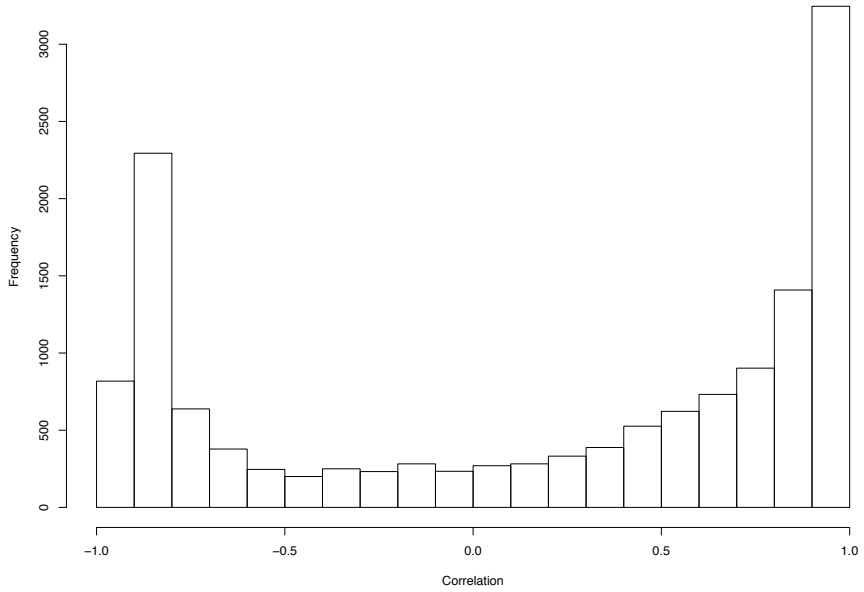

Figure 4.1. Histogram of the correlations

volume at each density level by the total bone volume, so that the bone size effect is removed. Let μ_1 and μ_2 be the population bone volumes for genotype T0A0 and T1A1, respectively. The hypothesis test is

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2.$$

In the selected data set, there are $p = 120$ measurements of bone volume for each observation, and there are $n_1 = 16$ mice with genotype T0A0 and $n_2 = 13$ with T1A1 in the data set. It is a two sample mean testing problem for high-dimensional data, since $p > n_1 + n_2 - 2$. Figure 1 is the histogram of the off-diagonal elements of the correlation matrix, which shows the high correlation between variables. It implies the irrationality of using diagonal matrix to approximate the correlation matrix. We consider using $[2Newn^{1/3}]$ basis matrices selected by following the same rule stated in the simulation studies. We apply the new test methods without and with the regularization process to the high resolution microcomputed tomography dataset, and compare the results with those of BS, CQ, SD1, SD2, LWJ1, LWJ2 and RAPTT. The results are summarized in Table 4.3.1.

Table 4.3.1. P-values of the tests

Method	P-value	Method	P-value
New(full)	0	New(MCP)	0
New(SCAD)	0	BS	0.2003
CQ	0.2579	SD1	0.4313
SD2	0.3310	LWJ1	0.0458
LWJ2	0.5712	RAPTT	0.0495

Table 4.3.1 shows that the new method with or without regularization process can reject H_0 and the p-values are 0, which is a strong evidence showing that the difference of bone volume between two different genotypes has been detected. BS, CQ, SD1 and SD2 cannot reject H_0 as we expected, because they neglect the correlation structure between variables. For the three projection methods, LWJ2 fails to reject H_0 and other two methods can reject H_0. However, the corresponding p-values are very close to 0.05. The results imply the effectiveness of our proposed method when the high correlations between variables and the proposed method is also more powerful than the existing methods.

4.4 Technical Proofs

Proof of (4.1.2). Let $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^T$, the eigen-decomposition of $\mathbf{\Sigma}$, where $\mathbf{\Gamma}$ is a $p \times p$ orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix with k-th element λ_k. We represent $\boldsymbol{\mu}_d$ in the coordinate system of $\mathbf{\Gamma}$ as $\boldsymbol{\mu}_d = \mathbf{\Gamma}\mathbf{a}$. Then

$$\boldsymbol{\mu}_d^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_d = \sum_{k=1}^{p} a_k^2/\lambda_k,$$

where $\mathbf{a} = (a_1, \cdots, a_p)^T$ and

$$\|\boldsymbol{\mu}_d\|^2/\sqrt{2\mathrm{tr}(\mathbf{\Sigma}^2)} = \sum_{k=1}^{p} a_k^2/(2\sum_{k=1}^{p} \lambda_k^2)^{1/2}.$$

Note that $(2\sum_{k=1}^{p}\lambda_k^2)^{1/2} \geqslant \lambda_k$ for every k. Thus,

$$\|\boldsymbol{\mu}_d\|^2/\sqrt{2\mathrm{tr}(\boldsymbol{\Sigma}^2)} \leqslant \boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$$

for any $\boldsymbol{\mu}_d$. This is (4.1.2).

The best choice of W. By the property of noncentral χ^2-distribution, it can be shown that the asymptotic mean and variance of T are

$$E_a(T) = (n/n_1 + n/n_2)\mathrm{tr}(\boldsymbol{\Sigma}^{1/2}\mathbf{W}\boldsymbol{\Sigma}^{1/2}) + \boldsymbol{\epsilon}_n^T\mathbf{W}\boldsymbol{\epsilon}_n,$$

and

$$\mathrm{Var}_a(T) = 2(n/n_1 + n/n_2)^2\mathrm{tr}(\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}) + 4(n/n_1 + n/n_2)\boldsymbol{\epsilon}_n^T\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\epsilon}_n$$

where $\boldsymbol{\epsilon}_n = \sqrt{n}\boldsymbol{\mu}_d$. Then the asymptotic power function of T can be expressed as

$$\begin{aligned}
& \Phi\left\{\frac{-z_\alpha\sqrt{2(\frac{n}{n_1}+\frac{n}{n_2})^2\mathrm{tr}(\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}\mathbf{W})} + \boldsymbol{\epsilon}_n^T\mathbf{W}\boldsymbol{\epsilon}_n}{\sqrt{2(\frac{n}{n_1}+\frac{n}{n_2})^2\mathrm{tr}(\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}) + 4(\frac{n}{n_1}+\frac{n}{n_2})\boldsymbol{\epsilon}_n^T\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\boldsymbol{\epsilon}_n}}\right\} \\
= \quad & \Phi\left\{-\frac{z_\alpha}{\sqrt{1+2(\frac{n}{n_1}+\frac{n}{n_2})^{-1}\frac{\boldsymbol{\nu}^T\mathbf{A}^2\boldsymbol{\nu}}{\mathrm{tr}(\mathbf{A}^2)}}} + \frac{\boldsymbol{\nu}^T\mathbf{A}\boldsymbol{\nu}}{\sqrt{2(\frac{n}{n_1}+\frac{n}{n_2})^2\mathrm{tr}(\mathbf{A}^2) + 4(\frac{n}{n_1}+\frac{n}{n_2})\boldsymbol{\nu}^T\mathbf{A}^2\boldsymbol{\nu}}}\right\} \\
\widehat{=} \quad & \beta(\boldsymbol{\nu}, \mathbf{A}) \qquad (4.4.1)
\end{aligned}$$

where $\boldsymbol{\nu} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}_n$ and $\mathbf{A} = \boldsymbol{\Sigma}^{1/2}\mathbf{W}\boldsymbol{\Sigma}^{1/2}$. Denote by $\eta_0 = \boldsymbol{\nu}^T\boldsymbol{\nu} = \boldsymbol{\epsilon}_n^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}_n$, and by $\lambda_{\min}(\mathbf{D})$ and $\lambda_{\max}(\mathbf{D})$ the minimal and maximal eigenvalues of $\mathbf{D}$, respectively. Note that

$$\frac{\lambda_{\min}^2(\mathbf{A})}{\lambda_{\max}^2(\mathbf{A})}\frac{\eta_0}{p} \leqslant \frac{\boldsymbol{\nu}^T\mathbf{A}^2\boldsymbol{\nu}}{\mathrm{tr}(\mathbf{A}^2)} \leqslant \frac{\lambda_{\max}^2(\mathbf{A})}{\lambda_{\min}^2(\mathbf{A})}\frac{\eta_0}{p}$$

and

$$\frac{\lambda_{\min}(\mathbf{A})}{\lambda_{\max}(\mathbf{A})}\frac{\eta_0}{\sqrt{2(p(\frac{n}{n_1}+\frac{n}{n_2})+2\eta_0)}} \leqslant \frac{\boldsymbol{\nu}^T\mathbf{A}\boldsymbol{\nu}}{\sqrt{2(\frac{n}{n_1}+\frac{n}{n_2})\mathrm{tr}(\mathbf{A}^2) + 4\boldsymbol{\nu}^T\mathbf{A}^2\boldsymbol{\nu}}}$$

$$\frac{\boldsymbol{\nu}^T\mathbf{A}\boldsymbol{\nu}}{\sqrt{2(\frac{n}{n_1}+\frac{n}{n_2})\mathrm{tr}(\mathbf{A}^2)+4\boldsymbol{\nu}^T\mathbf{A}^2\boldsymbol{\nu}}} \leqslant \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}\frac{\eta_0}{\sqrt{2(p(\frac{n}{n_1}+\frac{n}{n_2})+2\eta_0)}}.$$

Let $c(\mathbf{A})=\frac{\lambda_{\min}(\mathbf{A})}{\lambda_{\max}(\mathbf{A})}$. Then

$$\beta(\boldsymbol{\nu},\mathbf{A}) \geqslant \Phi\left\{-\frac{z_\alpha}{\sqrt{1+2\left(\frac{n^2}{n_1n_2}\right)^{-1}c^2(\mathbf{A})\frac{\eta_0}{p}}}+\frac{c(\mathbf{A})\eta_0}{\sqrt{2[\frac{pn^4}{n_1^2n_2^2}+\frac{n^22\eta_0}{n_1n_2}]}}\right\} \tag{4.4.2}$$

and

$$\beta(\boldsymbol{\nu},\mathbf{A}) \leqslant \Phi\left\{-\frac{z_\alpha}{\sqrt{1+2\left(\frac{n^2}{n_1n_2}\right)^{-1}c^{-2}(\mathbf{A})\frac{\eta_0}{p}}}+\frac{c^{-1}(\mathbf{A})\eta_0}{\sqrt{2[\frac{pn^4}{n_1^2n_2^2}+\frac{n^22\eta_0}{n_1n_2}]}}\right\} \tag{4.4.3}$$

where $n/n_1+n/n_2=n^2/(n_1n_2)$. Since $0\leqslant c(\mathbf{A})\leqslant 1$, then when $c(\mathbf{A})=1$, the right side of (4.4.2) achieve the maximum value and

$$\beta(\boldsymbol{\nu},\mathbf{A}) = \Phi\left\{-\frac{z_\alpha}{\sqrt{1+2\left(\frac{n^2}{n_1n_2}\right)^{-1}\frac{\eta_0}{p}}}+\frac{\eta_0}{\sqrt{2[\frac{pn^4}{n_1^2n_2^2}+\frac{n^22\eta_0}{n_1n_2}]}}\right\}.$$

In order to maximize the asymptotic power of T in the worst scenario, $c(\mathbf{A})$ should be taken to be 1 since $0\leqslant c(\mathbf{A})\leqslant 1$. $c(\mathbf{A})=1$ implies that $\mathbf{W}=\lambda\boldsymbol{\Sigma}^{-1}$ for some positive constant λ. This leads to the asymptotic power function $\Phi\{[-z_\alpha(n/n_1+n/n_2)\sqrt{2p}+\eta_0]/\sqrt{2p(n/n_1+n/n_2)^2+4(n/n_1+n/n_2)\eta_0}\}$.

Proof of Theorem 4.1.1. We have $\mathbf{R}^{-1}=\theta_1\mathbf{A}_1+\ldots+\theta_K\mathbf{A}_K$. Now, we compute the following optimal problem as follows $\min\limits_{\theta_1,\ldots,\theta_k}\mathrm{tr}[\widehat{\mathbf{R}}(\theta_1\mathbf{A}_1+\ldots+\theta_K\mathbf{A}_K)-\mathbf{I}_p]^2$. Then we have

$$\begin{cases} \widehat{\theta}_1\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_1\widehat{\mathbf{R}}\mathbf{A}_1+\widehat{\theta}_2\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_1\widehat{\mathbf{R}}\mathbf{A}_2+\ldots+\widehat{\theta}_K\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_1\widehat{\mathbf{R}}\mathbf{A}_K=\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_1 \\ \vdots \\ \widehat{\theta}_1\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_K\widehat{\mathbf{R}}\mathbf{A}_1+\widehat{\theta}_2\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_K\widehat{\mathbf{R}}\mathbf{A}_2+\ldots+\widehat{\theta}_K\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_K\widehat{\mathbf{R}}\mathbf{A}_K=\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_K. \end{cases} \tag{4.4.4}$$

That is, the estimate $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_K)^T$ satisfies

$$\widehat{\boldsymbol{\theta}} = \widehat{\mathbf{B}}^{-1}\widehat{\boldsymbol{\alpha}} \tag{4.4.5}$$

where $\widehat{\boldsymbol{\alpha}} = (p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_1, \ldots, p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_K)^T$ and $\widehat{\mathbf{B}}$ is a $K \times K$ dimensional matrix with the (k_1, k_2) element being $(p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_{k_1}\widehat{\mathbf{R}}\mathbf{A}_{k_2})$. To obtain the limit of $\widehat{\boldsymbol{\theta}}$, we will obtain the limits of $p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_k$ and $p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_{k_1}\widehat{\mathbf{R}}\mathbf{A}_{k_2}$ for $k, k_1, k_2 = 1, \ldots, K$.

Step 1. Proving $p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_k - p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_k = o_p(1)$ for $k = 1, \ldots, K$. Because $\widehat{\mathbf{R}}$ is not related to $\mathrm{diag}(\boldsymbol{\Sigma})$, then without loss of generality, we assume that $\mathrm{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_p$. Then we have

$$\begin{aligned}
p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_k &= p^{-1}\mathrm{tr}[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{S}[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{A}_k \\
&= p^{-1}\mathrm{tr}\mathbf{S}\mathbf{A}_k + p^{-1}\mathrm{tr}\{[\mathrm{diag}(\mathbf{S})]^{-1/2} - \mathbf{I}_p\}\mathbf{S}[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{A}_k \\
&\quad + p^{-1}\mathrm{tr}\mathbf{S}\{[\mathrm{diag}(\mathbf{S})]^{-1/2} - \mathbf{I}_p\}\mathbf{A}_k \\
&= \frac{n}{n-2}p^{-1}\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{A}_k\mathbf{R}^{1/2}\mathbf{r}_i - \frac{n}{n-2}p^{-1}n_1\bar{\mathbf{r}}_1^T\mathbf{R}^{1/2}\mathbf{A}_k\mathbf{R}^{1/2}\bar{\mathbf{r}}_1 \\
&\quad - \frac{n}{n-2}p^{-1}n_2\bar{\mathbf{r}}_2^T\mathbf{R}^{1/2}\mathbf{A}_k\mathbf{R}^{1/2}\bar{\mathbf{r}}_2 + o_p(1) \\
&= p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_k + o_p(1)
\end{aligned}$$

where $n = n_1 + n_2$, $\bar{\mathbf{r}}_1 = n_1^{-1}\sum_{i=1}^{n_1}\mathbf{r}_i$, $\bar{\mathbf{r}}_2 = n_2^{-1}\sum_{i=n_1+1}^{n_1+n_2}\mathbf{r}_i$, $\mathbf{r}_i = n^{-1/2}\mathbf{w}_{1i}$ for $i = 1, \ldots, n_1$ and $\mathbf{r}_{i+n_1} = n^{-1/2}\mathbf{w}_{2i}$ for $i = 1, \ldots, n_2$. Then $p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_k - p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_k = o_p(1)$ for $k = 1, \ldots, K$.

Step 2. Proving $p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_{k_1}\widehat{\mathbf{R}}\mathbf{A}_{k_2} - p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_1}\mathbf{R}\mathbf{A}_{k_2} - y_{n-2}(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_1})(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_2}) = o_p(1)$ for $k_1, k_2 = 1, \ldots, K$. We have

$$\begin{aligned}
& p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_{k_1}\widehat{\mathbf{R}}\mathbf{A}_{k_2} \\
= \; & p^{-1}\mathrm{tr}\left([\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{S}[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{A}_{k_1}[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{S}[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{A}_{k_2}\right) \\
= \; & p^{-1}\mathrm{tr}\mathbf{S}\mathbf{A}_{k_1}\mathbf{S}\mathbf{A}_{k_2} + o_p(1) \\
= \; & \frac{n^2}{(n-2)^2}p^{-1}\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{A}_{k_1}\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{A}_{k_2}\mathbf{R}^{1/2}\mathbf{r}_i
\end{aligned}$$

$$
\begin{aligned}
&+\frac{n^2}{(n-2)^2}p^{-1}\sum_{i\neq j}\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{A}_{k_1}\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{A}_{k_2}\mathbf{R}^{1/2}\mathbf{r}_i\\
&-2\frac{n^2}{(n-2)^2}p^{-1}n_1\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{A}_{k_1}\mathbf{R}^{1/2}\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T\mathbf{R}^{1/2}\mathbf{A}_{k_2}\mathbf{R}^{1/2}\mathbf{r}_i\\
&-2\frac{n^2}{(n-2)^2}p^{-1}n_2\sum_{i=1}^{n}\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{A}_{k_1}\mathbf{R}^{1/2}\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T\mathbf{R}^{1/2}\mathbf{A}_{k_2}\mathbf{R}^{1/2}\mathbf{r}_i\\
&+\frac{n^2}{(n-2)^2}p^{-1}n_1^2\bar{\mathbf{r}}_1^T\mathbf{R}^{1/2}\mathbf{A}_{k_1}\mathbf{R}^{1/2}\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T\mathbf{R}^{1/2}\mathbf{A}_{k_2}\mathbf{R}^{1/2}\bar{\mathbf{r}}_1\\
&+\frac{n^2}{(n-2)^2}p^{-1}n_2^2\bar{\mathbf{r}}_2^T\mathbf{R}^{1/2}\mathbf{A}_{k_1}\mathbf{R}^{1/2}\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T\mathbf{R}^{1/2}\mathbf{A}_{k_2}\mathbf{R}^{1/2}\bar{\mathbf{r}}_2\\
&+2\frac{n^2}{(n-2)^2}p^{-1}n_1n_2\bar{\mathbf{r}}_1^T\mathbf{R}^{1/2}\mathbf{A}_{k_1}\mathbf{R}^{1/2}\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T\mathbf{R}^{1/2}\mathbf{A}_{k_2}\mathbf{R}^{1/2}\bar{\mathbf{r}}_1+o_p(1)\\
=\;&\frac{n(n-1)}{(n-2)^2}p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_1}\mathbf{R}\mathbf{A}_{k_2}+\frac{pn}{(n-2)^2}(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_1})(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_2})+o_p(1)\\
=\;&p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_1}\mathbf{R}\mathbf{A}_{k_2}+y_{n-2}(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_1})(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_2})+o_p(1).
\end{aligned}
$$

That is, $p^{-1}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{A}_{k_1}\widehat{\mathbf{R}}\mathbf{A}_{k_2}-p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_1}\mathbf{R}\mathbf{A}_{k_2}-y_{n-2}(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_1})(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_2})=o_p(1)$ for $k_1,k_2=1,\ldots,K$. Let $\boldsymbol{\alpha}=(p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_1,\ldots,p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_K)^T$ and $\mathbf{A}$ be a $K\times K$ dimensional matrix with the (k_1,k_2) element being $p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_{k_1}\mathbf{R}\mathbf{A}_{k_2}$ for $k_1,k_2=1,\ldots,K$. Then we have

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}}&=\widehat{\mathbf{B}}^{-1}\widehat{\boldsymbol{\alpha}}\\
&=(\mathbf{A}+y\boldsymbol{\alpha}\boldsymbol{\alpha}^T)^{-1}\boldsymbol{\alpha}+o_p(1)\\
&=(\mathbf{A}^{-1}-\mathbf{A}^{-1}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\mathbf{A}^{-1}(y^{-1}+\boldsymbol{\alpha}^T\mathbf{A}^{-1}\boldsymbol{\alpha})^{-1})\boldsymbol{\alpha}+o_p(1)\\
&=\mathbf{A}^{-1}\boldsymbol{\alpha}(1+y\boldsymbol{\alpha}^T\mathbf{A}^{-1}\boldsymbol{\alpha})^{-1}+o_p(1)\\
&=\boldsymbol{\theta}(1+y\boldsymbol{\alpha}^T\mathbf{A}^{-1}\boldsymbol{\alpha})^{-1}+o_p(1)\\
&=(1+y)^{-1}\boldsymbol{\theta}+o_p(1)
\end{aligned}
$$

where $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_K)^T$, $(1+y\boldsymbol{\alpha}^T\mathbf{A}^{-1}\boldsymbol{\alpha})^{-1}=(1+y)^{-1}$ and $y_{n-2}=p/(n-2)\to y$ as $p,n\to\infty$.

Then the proof of Theorem 4.1.1 is completed.

Proof of Theorem 4.1.2. We will prove the CLT of $p\sum_{k=1}^K\pi_k[\widehat{\theta}_k-(1+y_{n-2})^{-1}\theta_k]$ for any constant vector $(\pi_1,\ldots,\pi_K)^T$. Step 1 will prove $p\sum_{k=1}^K\pi_k[\widehat{\theta}_k-$

$(1+y_{n-2})^{-1}\theta_k] = \mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1 - (1+y_{n-2})^{-1}\mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1\hat{\mathbf{R}}\mathbf{R}^{-1}$ where $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_K)^T = \mathbf{B}^{-1}(\pi_1,\ldots,\pi_K)^T$, $\mathbf{B} = \mathbf{A} + y_{n-2}\boldsymbol{\alpha}\boldsymbol{\alpha}^T$ and $\mathbf{D}_1 = \eta_1\mathbf{A}_1 + \ldots + \eta_K\mathbf{A}_K$. Step 2 will prove the CLT of $(\mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1 - \mathrm{E}\hat{\mathbf{R}}\mathbf{D}_1, \mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1\hat{\mathbf{R}}\mathbf{R}^{-1} - \mathrm{Etr}\hat{\mathbf{R}}\mathbf{D}_1\hat{\mathbf{R}}\mathbf{R}^{-1})$. Step 3 will obtain the CLT of $p\sum_{k=1}^{K}\pi_k[\hat{\theta}_k - (1+y_{n-2})^{-1}\theta_k]$.

Step 1. We have $\hat{\mathbf{B}}^{-1} = \mathbf{B}^{-1} - \mathbf{B}^{-1}(\hat{\mathbf{B}}-\mathbf{B})\hat{\mathbf{B}}^{-1}$. Thus

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \hat{\mathbf{B}}^{-1}\hat{\boldsymbol{\alpha}} \\
&= \mathbf{B}^{-1}\hat{\boldsymbol{\alpha}} - \mathbf{B}^{-1}(\hat{\mathbf{B}}-\mathbf{B})\hat{\mathbf{B}}^{-1}\hat{\boldsymbol{\alpha}} \\
&= \mathbf{B}^{-1}\hat{\boldsymbol{\alpha}} - \mathbf{B}^{-1}(\hat{\mathbf{B}}-\mathbf{B})\hat{\boldsymbol{\theta}} \\
&= \mathbf{B}^{-1}\hat{\boldsymbol{\alpha}} - (1+y_{n-2})^{-1}\mathbf{B}^{-1}(\hat{\mathbf{B}}-\mathbf{B})\boldsymbol{\theta} + o_p(p^{-1})
\end{aligned}$$

where $(\hat{\mathbf{B}}-\mathbf{B})\boldsymbol{\theta}$ is the K-dimensional vector with the kth element being $p^{-1}\mathrm{tr}\hat{\mathbf{R}}\mathbf{A}_k\hat{\mathbf{R}}\mathbf{R}^{-1} - (1+y_{n-2})p^{-1}\mathrm{tr}\mathbf{R}\mathbf{A}_k$ for $k=1,\ldots,K$. That is,

$$\begin{aligned}
& p(\hat{\boldsymbol{\theta}} - (1+y_{n-2})^{-1}\boldsymbol{\theta}) \\
= \;& p\mathbf{B}^{-1}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}) - (1+y_{n-2})^{-1}p\mathbf{B}^{-1}(\hat{\mathbf{B}}-\mathbf{B})\boldsymbol{\theta} \\
= \;& \mathbf{B}^{-1}[\mathrm{tr}\hat{\mathbf{R}}\mathbf{A}_1 - (1+y_{n-2})^{-1}\mathrm{tr}\hat{\mathbf{R}}\mathbf{A}_1\hat{\mathbf{R}}\mathbf{R}^{-1}, \ldots, \mathrm{tr}\hat{\mathbf{R}}\mathbf{A}_K - (1+y_{n-2})^{-1}\mathrm{tr}\hat{\mathbf{R}}\mathbf{A}_K\hat{\mathbf{R}}\mathbf{R}^{-1}]^T.
\end{aligned}$$

Then we have $p\boldsymbol{\pi}^T[\hat{\boldsymbol{\theta}} - (1+y_{n-2})^{-1}\boldsymbol{\theta}] = \mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1 - (1+y_{n-2})^{-1}\mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1\hat{\mathbf{R}}\mathbf{R}^{-1}$.

Step 2. We will prove the CLT of $(\mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1 - \mathrm{E}\hat{\mathbf{R}}\mathbf{D}_1, \mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1\hat{\mathbf{R}}\mathbf{R}^{-1} - \mathrm{Etr}\hat{\mathbf{R}}\mathbf{D}_1\hat{\mathbf{R}}\mathbf{R}^{-1})$. It is easy to verify that $\{(\mathrm{E}_i - \mathrm{E}_{i-1})\mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1, i=1,\ldots,n\}$ and $\{(\mathrm{E}_i - \mathrm{E}_{i-1})\mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1\hat{\mathbf{R}}\mathbf{R}^{-1}, i=1,\ldots,n\}$ are two martingale difference sequences and satisfying Lindeberg conditions and E_i is the conditional expectation based on $\mathbf{r}_1,\ldots,\mathbf{r}_{i-1}$. We will first derive $\nu = p\mathrm{E}\boldsymbol{\pi}^T[\hat{\boldsymbol{\theta}} - (1+y_{n-2})^{-1}\boldsymbol{\theta}] = \nu_1 - (1+y_{n-2})^{-1}\nu_2$ where $\nu_1 = \mathrm{E}\hat{\mathbf{R}}\mathbf{D}_1$ and $\nu_2 = \mathrm{Etr}\hat{\mathbf{R}}\mathbf{D}_1\hat{\mathbf{R}}\mathbf{R}^{-1}$. We have

$$\begin{aligned}
& \mathrm{tr}\hat{\mathbf{R}}\mathbf{D}_1 \\
= \;& \mathrm{tr}\mathbf{S}\mathbf{D}_1 + \mathrm{tr}(\mathrm{diag}^{-1/2}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathrm{diag}^{-1/2}(\mathbf{S})\mathbf{D}_1 + \mathrm{tr}\mathbf{S}(\mathrm{diag}^{-1/2}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1 \\
= \;& \mathrm{tr}\mathbf{S}\mathbf{D}_1 - \frac{1}{2}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1 \\
& + \frac{1}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1 + \frac{3}{8}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)^2\mathbf{S}\mathbf{D}_1
\end{aligned}$$

$$
\begin{aligned}
& -\frac{1}{2}\mathrm{tr}\mathbf{D}_1\mathbf{S}(\mathrm{diag}(\mathbf{S}-\mathbf{I}_p)) + \frac{3}{8}\mathrm{tr}\mathbf{D}_1\mathbf{S}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)^2 + o_p(1) \\
= \;& \mathrm{tr}\mathbf{S}\mathbf{D}_1 - \mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}\mathbf{D}_1 \\
& +\frac{1}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{D}_1 + \frac{3}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)^2\mathbf{S}\mathbf{D}_1 + o_p(1)
\end{aligned}
$$

where

$$
\begin{aligned}
& \mathrm{tr}(\mathrm{diag}^{-1/2}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}\mathrm{diag}^{-1/2}(\mathbf{S})\mathbf{D}_1 \\
= \;& -\frac{1}{2}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}\mathrm{diag}^{-1/2}(\mathbf{S})\mathbf{D}_1 + \frac{3}{8}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)^2\mathbf{S}\mathrm{diag}^{-1/2}(\mathbf{S})\mathbf{D}_1 \\
= \;& -\frac{1}{2}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}\mathbf{D}_1 + \frac{1}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{D}_1 \\
& +\frac{3}{8}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)^2\mathbf{S}\mathbf{D}_1 + o_p(1)
\end{aligned}
$$

and

$$
\mathrm{tr}\mathbf{D}_1\mathbf{S}(\mathrm{diag}^{-1/2}(\mathbf{S})-\mathbf{I}_p) = -\frac{1}{2}\mathrm{tr}\mathbf{D}_1\mathbf{S}(\mathrm{diag}(\mathbf{S}-\mathbf{I}_p)) + \frac{3}{8}\mathrm{tr}\mathbf{D}_1\mathbf{S}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)^2 + o_p(1).
$$

Thus we have

$$
\begin{aligned}
\nu_1 &= \mathrm{E}\mathrm{tr}\widehat{\mathbf{R}}\mathbf{D}_1 \\
&= \mathrm{E}\mathrm{tr}\mathbf{S}\mathbf{D}_1 - \mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}\mathbf{D}_1 \\
&\quad +\frac{1}{4}\mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{D}_1 + \frac{3}{4}\mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)^2\mathbf{S}\mathbf{D}_1 + o(1) \\
&= \mathrm{tr}\mathbf{R}\mathbf{D}_1 - \frac{n}{(n-2)^2}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right] \\
&\quad +\frac{n(n-1)}{4(n-2)^3}\mathrm{tr}\mathbf{D}_0\mathbf{D}_1 + \frac{3n(n-1)}{4(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^4\right] + o(1)
\end{aligned}
$$

because $\mathrm{E}\mathbf{S}\mathbf{D}_1 = \mathrm{tr}\mathbf{R}\mathbf{D}_1$,

$$
\begin{aligned}
& \mathrm{E}\mathrm{tr}\mathbf{S}\mathbf{D}_1(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p) \\
= \;& \frac{n^2}{(n-2)^2}\mathrm{E}\mathrm{tr}\mathbf{R}^{1/2}\mathbf{F}_n\mathbf{R}^{1/2}\mathbf{D}_1(\mathrm{diag}(\mathbf{R}^{1/2}\mathbf{F}_n\mathbf{R}^{1/2})-\mathbf{I}_p) + o(1)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n^2}{(n-2)^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{p}\mathrm{E}\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1}) + o(1)\\
&= \frac{n^2}{(n-2)^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{p}\mathrm{E}\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i(\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_j - n^{-1}) + o(1)\\
&= \frac{n}{(n-2)^2}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell\cdot\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell\right] + o(1)\\
&= \frac{n}{(n-2)^2}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right] + o(1)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\mathrm{tr}\mathbf{D}_1\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)^2\\
&= \frac{n^3}{(n-2)^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{h=1}^{n}\sum_{k=1}^{p}\mathrm{E}\mathbf{e}_k^T\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1}) + o(1)\\
&\quad(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_h\mathbf{r}_h^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1})\\
&= \frac{n^3}{(n-2)^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{h=1}^{n}\sum_{k=1}^{p}\mathrm{E}\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{r}_i(\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_j - n^{-1}) + o(1)\\
&\quad(\mathbf{r}_h^T\mathbf{R}^{1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_h - n^{-1})\\
&= \frac{n(n-1)}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\right] + o(1)\\
&= \frac{n(n-1)}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^4\right] + o(1)
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1\\
&= \frac{n^3}{(n-2)^3}\mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{R}^{1/2}\mathbf{F}_n\mathbf{R}^{1/2}) - \mathbf{I}_p)\mathbf{R}^{1/2}\mathbf{F}_n\mathbf{R}^{1/2}(\mathrm{diag}(\mathbf{R}^{1/2}\mathbf{F}_n\mathbf{R}^{1/2}) - \mathbf{I}_p)\mathbf{D}_1 + o(1)\\
&= \frac{n^3}{(n-2)^3}\sum_{k=1}^{p}\sum_{\ell=1}^{p}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{m=1}^{n}\mathrm{E}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1})\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_\ell\\
&\quad(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1})\mathbf{e}_\ell^T\mathbf{D}_1\mathbf{e}_k + o(1)\\
&= \frac{n(n-1)}{(n-2)^3}\mathrm{tr}\mathbf{D}_0\mathbf{D}_1 + o(1)
\end{aligned}
$$

where $\mathbf{F}_n = \sum_{i=1}^n \mathbf{r}_i\mathbf{r}_i^T$, $\mathbf{D}_0$ is the $p \times p$ dimensional matrix with the (i,j) element being $u_{ij} = 2(\mathbf{e}_i\mathbf{R}\mathbf{e}_j)^3 + \beta_w r_{ij} \sum_{\ell=1}^p (\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_j)^2$.

Moreover we have

$$
\begin{aligned}
& \mathrm{tr}(\widehat{\mathbf{R}}\mathbf{R}^{-1}\widehat{\mathbf{R}}\mathbf{D}_1) \\
= \; & \mathrm{tr}[\mathbf{S}\mathbf{R}^{-1} + (\mathrm{diag}^{-1/2}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathrm{diag}^{-1/2}(\mathbf{S})\mathbf{R}^{-1} + \mathbf{S}(\mathrm{diag}^{-1/2}(\mathbf{S}) - \mathbf{I}_p)\mathbf{R}^{-1}] \\
& \quad \cdot[\mathbf{S}\mathbf{D}_1 + (\mathrm{diag}^{-1/2}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathrm{diag}^{-1/2}(\mathbf{S})\mathbf{D}_1 + \mathbf{S}(\mathrm{diag}^{-1/2}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1] \\
= \; & \mathrm{tr}(\mathbf{S}\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1) - \mathrm{tr}[(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1] - \mathrm{tr}[(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1\mathbf{S}\mathbf{R}^{-1}] \\
& + \frac{1}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1\mathbf{S}\mathbf{R}^{-1} \\
& + \frac{1}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1 \\
& + \frac{3}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)^2\mathbf{D}_1\mathbf{S}\mathbf{R}^{-1}\mathbf{S} \\
& + \frac{3}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)^2\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1\mathbf{S} \\
& + \frac{1}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{R}^{-1}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1 \\
& + \frac{1}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{R}^{-1}\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1 \\
& + \frac{1}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{R}^{-1}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1\mathbf{S} \\
& + \frac{1}{4}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{R}^{-1}\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1\mathbf{S}. \qquad (4.4.6)
\end{aligned}
$$

Then in the following, we will derive $\mathrm{Etr}(\mathbf{S}\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1)$, $\mathrm{Etr}[(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1]$, $\mathrm{Etr}[(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1\mathbf{S}\mathbf{R}^{-1}]$, $\mathrm{Etr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1\mathbf{S}\mathbf{R}^{-1}$, $\mathrm{Etr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1$, $\mathrm{Etr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)^2\mathbf{D}_1\mathbf{S}\mathbf{R}^{-1}\mathbf{S}$, $\mathrm{Etr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)^2\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1\mathbf{S}$, $\mathrm{Etr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{R}^{-1}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1$, $\mathrm{Etr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{R}^{-1}\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1$, $\mathrm{Etr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{R}^{-1}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1\mathbf{S}$ and $\mathrm{Etr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{R}^{-1}\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1\mathbf{S}$. Because

$$
\begin{aligned}
\mathrm{Etr}(\mathbf{R}^{1/2}\mathbf{F}_n\mathbf{F}_n\mathbf{R}^{1/2}\mathbf{D}_1) &= \sum_{i=1}^n\sum_{j=1}^n \mathrm{Etr}\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{D}_1 \\
&= \sum_{i=1}^n\sum_{j=1}^n \mathrm{Etr}\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{D}_1
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n-1}{n}\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \frac{p}{n}\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \frac{1}{n}\bigg(2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w \mathrm{tr}\mathbf{R}\mathbf{D}_1\bigg) \\
&= \frac{n+1+p+\beta_w}{n}\mathrm{tr}\mathbf{R}\mathbf{D}_1,
\end{aligned}
$$

then we have

$$
\begin{aligned}
& \mathrm{E}\mathrm{tr}\mathbf{S}\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1 \\
= \; & \frac{n^2}{(n-2)^2}\mathrm{E}\mathrm{tr}\mathbf{R}^{1/2}(\mathbf{F}_n - n_1\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T - n_2\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T)(\mathbf{F}_n - n_1\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T - n_2\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T)\mathbf{R}^{1/2}\mathbf{D}_1 \\
= \; & \frac{n^2}{(n-2)^2}\mathrm{E}\mathrm{tr}\mathbf{R}^{1/2}\mathbf{F}_n\mathbf{F}_n\mathbf{R}^{1/2}\mathbf{D}_1 \\
& -\frac{n^2 n_1}{(n-2)^2}\mathrm{E}\mathrm{tr}\mathbf{R}^{1/2}\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T\mathbf{F}_n\mathbf{R}^{1/2}\mathbf{D}_1 - \frac{n^2 n_2}{(n-2)^2}\mathrm{E}\mathrm{tr}\mathbf{R}^{1/2}\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T\mathbf{F}_n\mathbf{R}^{1/2}\mathbf{D}_1 \\
& -\frac{n^2 n_1}{(n-2)^2}\mathrm{E}\mathrm{tr}\mathbf{R}^{1/2}\mathbf{F}_n\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T\mathbf{R}^{1/2}\mathbf{D}_1 - \frac{n^2 n_2}{(n-2)^2}\mathrm{E}\mathrm{tr}\mathbf{R}^{1/2}\mathbf{F}_n\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T\mathbf{R}^{1/2}\mathbf{D}_1 \\
& +\frac{n^2 n_1^2}{(n-2)^2}\mathrm{tr}\mathbf{R}^{1/2}\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T\mathbf{R}^{1/2}\mathbf{D}_1 + \frac{n^2 n_2^2}{(n-2)^2}\mathrm{tr}\mathbf{R}^{1/2}\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T\mathbf{R}^{1/2}\mathbf{D}_1 \\
& +\frac{n^2 n_1 n_2}{(n-2)^2}\mathrm{tr}\mathbf{R}^{1/2}\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T\mathbf{R}^{1/2}\mathbf{D}_1 + \frac{n^2 n_1 n_2}{(n-2)^2}\mathrm{tr}\mathbf{R}^{1/2}\bar{\mathbf{r}}_2\bar{\mathbf{r}}_2^T\bar{\mathbf{r}}_1\bar{\mathbf{r}}_1^T\mathbf{R}^{1/2}\mathbf{D}_1 \\
= \; & \frac{n(n-1)}{(n-2)^2}\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \frac{np}{(n-2)^2}\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \frac{n}{(n-2)^2}\bigg(2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\mathrm{tr}\mathbf{R}\mathbf{D}_1\bigg) \\
& -\frac{4(n-1)}{(n-2)^2}\mathrm{tr}\mathbf{R}\mathbf{D}_1 - \frac{4}{(n-2)^2}\mathrm{tr}\mathbf{R}\mathbf{D}_1 + p\bigg(\frac{n_1-1}{n_1(n-2)^2} + \frac{n_2-1}{n_2(n-2)^2}\bigg)\mathrm{tr}\mathbf{R}\mathbf{D}_1 \\
= \; & \frac{n^2-3n+4}{(n-2)^2}\mathrm{tr}\mathbf{R}\mathbf{D}_1 + p\left(\frac{n-4}{(n-2)^2} + \frac{n_1-1}{n_1(n-2)^2} + \frac{n_2-1}{n_2(n-2)^2}\right)\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \frac{\beta_w n}{(n-2)^2}\mathrm{tr}\mathbf{R}\mathbf{D}_1 \\
= \; & \left(\frac{n^2-3n+4}{(n-2)^2} + \frac{(n-4)p}{(n-2)^2} + \frac{(n_1-1)p}{n_1(n-2)^2} + \frac{(n_2-1)p}{n_2(n-2)^2} + \frac{\beta_w n}{(n-2)^2}\right)\mathrm{tr}\mathbf{R}\mathbf{D}_1. \qquad (4.4.7)
\end{aligned}
$$

$$
\begin{aligned}
& \mathrm{E}\mathrm{tr}[(\mathbf{S}\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1)(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)] \\
= \; & \frac{n^3}{(n-2)^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{h=1}^{n}\sum_{k=1}^{p}\mathrm{E}\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_h\mathbf{r}_h^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1}) + o(1) \\
= \; & \frac{n(n-1)}{(n-2)^3}\bigg[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\bigg]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{n(n-1)}{(n-2)^3}\sum_{k=1}^{p}\left[2\mathbf{e}_k^T\mathbf{R}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{e}_k+\beta_w\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^2\right]\\
&+\frac{np}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right]+o(1)\\
=\;&\frac{n[2(n-1)+p]}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right]+o(1). \qquad (4.4.8)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\mathrm{tr}[\mathbf{S}\mathbf{D}_1\mathbf{S}\mathbf{R}^{-1}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)]\\
=\;&\frac{n^3}{(n-2)^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{h=1}^{n}\sum_{k=1}^{p}\mathrm{E}\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{-1/2}\mathbf{e}_k\\
&\cdot(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_h\mathbf{r}_h^T\mathbf{R}^{1/2}\mathbf{e}_k-n^{-1})+o(1)\\
=\;&\frac{n(n-1)}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right]\\
&+\frac{n(n-1)}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1\mathbf{R}+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^2\right]\\
&+\frac{n}{(n-2)^3}\mathrm{tr}\mathbf{R}\mathbf{D}_1\left[2p+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right]+o(1) \qquad (4.4.9)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1 \qquad (4.4.10)\\
=\;&\mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{S}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)\mathbf{D}_1\mathbf{S}\mathbf{R}^{-1}\\
=\;&\frac{n(n-1)}{(n-2)^3}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\mathbf{e}_i^T\mathbf{D}_1\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right]+o(1)
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S})-\mathbf{I}_p)^2\mathbf{D}_1\mathbf{S}\mathbf{R}^{-1}\mathbf{S}\\
=\;&\frac{n(n-1)}{(n-2)^3}\left[2\mathrm{tr}\mathbf{D}_1\mathbf{R}+\beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{D}_1\mathbf{R}\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right]\\
&+\frac{n(n-1)p}{(n-2)^4}\left[2\mathrm{tr}\mathbf{D}_1\mathbf{R}+\beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{D}_1\mathbf{R}\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right]+o(1)
\end{aligned}
$$

$$= \frac{n(n-1)(n-2+p)}{(n-2)^4}\left[2\mathrm{tr}\mathbf{D}_1\mathbf{R} + \beta_w \sum_{k=1}^{p} \mathbf{e}_k^T\mathbf{D}_1\mathbf{R}\mathbf{e}_k \sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right] + o(1) \quad (4.4.11)$$

$$\begin{aligned} & \mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)^2\mathbf{R}^{-1}\mathbf{S}\mathbf{D}_1\mathbf{S} \\ = & \frac{n(n-1)}{(n-2)^3}\left[2\mathrm{tr}\mathbf{D}_1\mathbf{R} + \beta_w \sum_{k=1}^{p} \mathbf{e}_k^T\mathbf{D}_1\mathbf{R}\mathbf{e}_k \sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right] \\ & + \frac{n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{(n-2)^4}\left[2p + \beta_w \sum_{k=1}^{p}\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right] \end{aligned} \quad (4.4.12)$$

$$\begin{aligned} & \mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{R}^{-1}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1 \\ = & \mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{R}^{-1}\mathbf{S}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1\mathbf{S} \\ = & \frac{n(n-2)}{(n-1)^3}\left[2 + \beta_w \sum_{i=1}^{p} \mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{e}_i \sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^4\right] + o(1) \end{aligned} \quad (4.4.13)$$

$$\begin{aligned} & \mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{R}^{-1}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1\mathbf{S} \quad (4.4.14) \\ = & \frac{n(n-1)}{(n-2)^3}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right] \\ & + \frac{n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{(n-2)^4}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right] \end{aligned}$$

and

$$\begin{aligned} & \mathrm{E}\mathrm{tr}(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{D}_1(\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p)\mathbf{S}\mathbf{D}_1\mathbf{S} \quad (4.4.15) \\ = & \frac{n(n-1)}{(n-2)^3}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{e}_i^T\mathbf{D}_1\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right] \\ & + \frac{n(n-1)p}{(n-2)^4}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{e}_i^T\mathbf{D}_1\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right] \\ = & \frac{n(n-1)(n-2+p)}{(n-2)^4}\sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{e}_i^T\mathbf{D}_1\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right]. \end{aligned}$$

By (4.4.6)-(4.4.15), we have

$$
\begin{aligned}
& \nu_2 = \mathrm{tr}\mathrm{E}\widehat{\mathbf{R}}\mathbf{R}^{-1}\widehat{\mathbf{R}}\mathbf{D}_1 \\
= \; & \left(\frac{n^2-3n+4}{(n-2)^2}+\frac{(n-4)p}{(n-2)^2}+\frac{(n_1-1)p}{n_1(n-2)^2}+\frac{(n_2-1)p}{n_2(n-2)^2}+\frac{\beta_w n}{(n-2)^2}\right)\mathrm{tr}\mathbf{R}\mathbf{D}_1 \\
& -\frac{n[2(n-1)+p]}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right] \\
& -\frac{n(n-1)}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right] \\
& -\frac{n(n-1)}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1\mathbf{R}+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^2\right] \\
& -\frac{n}{(n-2)^3}\mathrm{tr}\mathbf{R}\mathbf{D}_1\left[2p+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right] \\
& +\frac{n(n-1)}{2(n-2)^3}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\mathbf{e}_i^T\mathbf{D}_1\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right] \\
& +\frac{3n(n-1)(n-2+p)}{4(n-2)^4}\left[2\mathrm{tr}\mathbf{D}_1\mathbf{R}+\beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{D}_1\mathbf{R}\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right] \\
& +\frac{3n(n-1)}{4(n-2)^3}\left[2\mathrm{tr}\mathbf{D}_1\mathbf{R}+\beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{D}_1\mathbf{R}\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right] \\
& +\frac{3n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{4(n-2)^4}\left[2p+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right] \\
& +\frac{2n(n-1)}{4(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1+\beta_w\sum_{i=1}^{p}\mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{e}_i\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^4\right] \\
& +\frac{n(n-1)}{4(n-2)^3}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right] \\
& +\frac{n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{4(n-2)^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right] \\
& +\frac{n(n-1)(n-2+p)}{4(n-2)^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{D}_1\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\right].
\end{aligned}
$$

That is,

$$
\begin{aligned}
& \nu_2 \\
= \; & \left(\frac{n^2-3n+4}{(n-2)^2} + \frac{(n-4)p}{(n-2)^2} + \frac{(n_1-1)p}{n_1(n-2)^2} + \frac{(n_2-1)p}{n_2(n-2)^2} + \frac{\beta_w n}{(n-2)^2} \right) \mathrm{tr}\mathbf{R}\mathbf{D}_1 \\
& - \frac{n[3(n-1)+p]}{(n-2)^3} \left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w \sum_{k=1}^{p}\sum_{\ell=1}^{p} \mathbf{e}_\ell^T \mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3 \right] \\
& - \frac{n(n-1)}{(n-2)^3} \left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1\mathbf{R} + \beta_w \sum_{k=1}^{p}\sum_{\ell=1}^{p} \mathbf{e}_\ell^T \mathbf{R}^{-1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell (\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^2 \right] \\
& - \frac{n}{(n-2)^3} \mathrm{tr}\mathbf{R}\mathbf{D}_1 \left[2p + \beta_w \sum_{k=1}^{p}\sum_{\ell=1}^{p} \mathbf{e}_\ell^T \mathbf{R}^{-1/2}\mathbf{e}_k (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3 \right] \\
& + \frac{3n(n-1)p + 6n(n-1)(n-2)}{4(n-2)^4} \left[2\mathrm{tr}\mathbf{D}_1\mathbf{R} + \beta_w \sum_{k=1}^{p} \mathbf{e}_k^T\mathbf{D}_1\mathbf{R}\mathbf{e}_k \sum_{\ell=1}^{p} (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4 \right] \\
& + \frac{3n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{4(n-2)^4} \left[2p + \beta_w \sum_{k=1}^{p}\sum_{\ell=1}^{p} (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4 \right] \\
& + \frac{2n(n-1)}{4(n-2)^3} \left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w \sum_{i=1}^{p} \mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{e}_i \sum_{k=1}^{p} (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^4 \right] \\
& + \frac{n(n-1)}{4(n-2)^3} \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_j \left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 + \beta_w \sum_{k=1}^{p} (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2 (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2 \right] \\
& + \frac{n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{4(n-2)^4} \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j \left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 + \beta_w \sum_{k=1}^{p} (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2 (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2 \right] \\
& + \frac{n(n-1)(3n-6+p)}{4(n-2)^4} \sum_{i=1}^{p}\sum_{j=1}^{p} \mathbf{e}_i^T\mathbf{D}_1\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j \left[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2 + \beta_w \sum_{k=1}^{p} (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2 (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2 \right].
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
& \nu = \nu_1 - (1+y_{n-2})^{-1}\nu_2 \\
= \; & \mathrm{tr}\mathbf{R}\mathbf{D}_1 - \frac{n}{(n-2)^2} \left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w \sum_{k=1}^{p}\sum_{\ell=1}^{p} \mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k (\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3 \right] \\
& + \frac{n(n-1)}{4(n-2)^3} \mathrm{tr}\mathbf{D}_0\mathbf{D}_1 + \frac{3n(n-1)}{4(n-2)^3} \left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1 + \beta_w \sum_{k=1}^{p} \mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{e}_k \sum_{\ell=1}^{p} (\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^4 \right]
\end{aligned}
$$

$$
\begin{aligned}
&-\frac{1}{1+y_{n-2}}\left(\frac{n^2-3n+4}{(n-2)^2}+\frac{(n-4)p}{(n-2)^2}+\frac{(n_1-1)p}{n_1(n-2)^2}+\frac{(n_2-1)p}{n_2(n-2)^2}+\frac{\beta_w n}{(n-2)^2}\right)\mathrm{tr}\mathbf{R}\mathbf{D}_1\\
&+(1+y_{n-2})^{-1}\frac{n[3(n-1)+p]}{(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right]\\
&+\frac{n(n-1)}{(1+y_{n-2})(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1\mathbf{R}+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k\mathbf{e}_k^T\mathbf{R}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^2\right]\\
&+(1+y_{n-2})^{-1}\frac{n}{(n-2)^3}\mathrm{tr}\mathbf{R}\mathbf{D}_1\left[2p+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^3\right]\\
&-\frac{3n(n-1)p+6n(n-1)(n-2)}{4(1+y_{n-2})(n-2)^4}\left[2\mathrm{tr}\mathbf{D}_1\mathbf{R}+\beta_w\sum_{k=1}^{p}\mathbf{e}_k^T\mathbf{D}_1\mathbf{R}\mathbf{e}_k\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right]\\
&-(1+y_{n-2})^{-1}\frac{3n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{4(n-2)^4}\left[2p+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4\right]\\
&-(1+y_{n-2})^{-1}\frac{2n(n-1)}{4(n-2)^3}\left[2\mathrm{tr}\mathbf{R}\mathbf{D}_1+\beta_w\sum_{i=1}^{p}\mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{e}_i\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^4\right]\\
&-(1+y_{n-2})^{-1}\frac{n(n-1)}{4(n-2)^3}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_j\Bigg[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2\\
&\qquad\qquad+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\Bigg]\\
&-(1+y_{n-2})^{-1}\frac{n(n-1)\mathrm{tr}\mathbf{R}\mathbf{D}_1}{4(n-2)^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{R}^{-1}\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\Bigg[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2\\
&\qquad\qquad+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\Bigg]\\
&-(1+y_{n-2})^{-1}\frac{n(n-1)(3n-6+p)}{4(n-2)^4}\sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{e}_i^T\mathbf{D}_1\mathbf{e}_j\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j\Bigg[2(\mathbf{e}_i^T\mathbf{R}\mathbf{e}_j)^2\\
&\qquad\qquad+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_i)^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_j)^2\Bigg].
\end{aligned}
$$

Moreover, we have

$$
\sigma_{110} = \sum_{i=1}^{n}\mathrm{E}_{i-1}(\mathrm{tr}\mathrm{E}_i\widehat{\mathbf{R}}\mathbf{D}_1-\mathrm{E}_{i-1}\widehat{\mathbf{R}}\mathbf{D}_1)^2
$$

$$
\begin{aligned}
&= 2n^{-1}\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2 + \beta_w n^{-1}\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell)^2 \\
&\quad +n^{-1}\sum_{\ell_1=1}^{p}\sum_{\ell_2=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\mathbf{e}_{\ell_2}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_2}\left(2(\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{e}_{\ell_2})^2 + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_1})^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_2})^2\right) \\
&\quad -2n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left(2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\cdot\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right)
\end{aligned}
$$

$$
\begin{aligned}
&\sigma_{220} \\
&= \sum_{i=1}^{n}\mathrm{E}_{i-1}(\mathrm{tr}\mathrm{E}_i\widehat{\mathbf{R}}\mathbf{R}^{-1}\widehat{\mathbf{R}}\mathbf{D}_1 - \mathrm{E}_{i-1}\widehat{\mathbf{R}}\mathbf{R}^{-1}\widehat{\mathbf{R}}\mathbf{D}_1)^2
\end{aligned}
$$

$$
\begin{aligned}
&= (y_n+2)^2n^{-1}\left[2\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2 + \beta_w\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\right] \\
&\quad +(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)^2(2-2y_n-\beta_w y_n) \\
&\quad +2y_n[n^{-1}\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2] \\
&\quad +(2+y_n)^2n^{-1}\sum_{\ell_1=1}^{p}\sum_{\ell_2=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\mathbf{e}_{\ell_2}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_2}\left[2(\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{e}_{\ell_2})^2 + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_1})^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_2})^2\right] \\
&\quad +\frac{(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)^2}{n}\left[2\mathrm{tr}\mathbf{R}^2 + \beta_w p\right] \\
&\quad -2(2+y_n)^2n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left[2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right] \\
&\quad -2(2+y_n)(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)n^{-1}\sum_{\ell=1}^{p}\left[2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right] \\
&\quad +(4+2y_n)(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)n^{-1}\sum_{\ell_1=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\left[2\mathbf{e}_{\ell_1}^T\mathbf{R}^2\mathbf{e}_{\ell_1} + \beta_w\right]
\end{aligned}
$$

$$
\begin{aligned}
\sigma_{120} &= \sum_{i=1}^{n}\mathrm{E}_{i-1}[(\mathrm{tr}\mathrm{E}_i\widehat{\mathbf{R}}\mathbf{D}_1 - \mathrm{E}_{i-1}\widehat{\mathbf{R}}\mathbf{D}_1)(\mathrm{tr}\mathrm{E}_i\widehat{\mathbf{R}}\mathbf{R}^{-1}\widehat{\mathbf{R}}\mathbf{D}_1 - \mathrm{E}_{i-1}\widehat{\mathbf{R}}\mathbf{R}^{-1}\widehat{\mathbf{R}}\mathbf{D}_1)] \\
&= (2+y_n)n^{-1}(2\mathrm{tr}\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell) \\
&\quad +(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)(2+\beta_w)
\end{aligned}
$$

$$
\begin{aligned}
& -(1+y_n)n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left(2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right)\\
& -n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left(2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\cdot\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right)\\
& -(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)n^{-1}\sum_{\ell=1}^{p}\left(2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right)\\
& -(2+y_n)n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left[2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right]\\
& -(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)(2+\beta_w)\\
& +(1+y_n)n^{-1}\sum_{\ell_1=1}^{p}\sum_{\ell_2=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\mathbf{e}_{\ell_2}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_2}\\
& \left[2(\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{e}_{\ell_2})^2+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_1})^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_2})^2\right]\\
& +n^{-1}\sum_{\ell_1=1}^{p}\sum_{\ell_2=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\mathbf{e}_{\ell_2}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_2}\left[2(\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{e}_{\ell_2})^2+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_1})^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_2})^2\right]\\
& +\frac{n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1}{n}\sum_{\ell_1=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\left[2\mathbf{e}_{\ell_1}^T\mathbf{R}^2\mathbf{e}_{\ell_1}+\beta_w\right]\\
= \;& (2+y_n)n^{-1}(2\mathrm{tr}\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{D}_1+\beta_w\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell)\\
& -(4+2y_n)n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left(2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right)\\
& -(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)n^{-1}\sum_{\ell=1}^{p}\left(2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right)\\
& +(2+y_n)n^{-1}\sum_{\ell_1=1}^{p}\sum_{\ell_2=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\mathbf{e}_{\ell_2}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_2}\\
& \left[2(\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{e}_{\ell_2})^2+\beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_1})^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_2})^2\right]\\
& +\frac{n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1}{n}\sum_{\ell_1=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\left[2\mathbf{e}_{\ell_1}^T\mathbf{R}^2\mathbf{e}_{\ell_1}+\beta_w\right]
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2 &= \sigma_{110} + (1+y)^{-2}\sigma_{220} - 2(1+y)^{-1}\sigma_{120} \\
&= 2n^{-1}\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2 + \beta_w n^{-1}\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell)^2 \\
&\quad + n^{-1}\sum_{\ell_1=1}^{p}\sum_{\ell_2=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\mathbf{e}_{\ell_2}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_2}\left(2(\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{e}_{\ell_2})^2 + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_1})^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_2})^2\right) \\
&\quad - 2n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left(2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\cdot\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right) \\
&\quad + (1+y)^{-2}(y+2)^2n^{-1}\left[2\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2 + \beta_w\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\right] \\
&\quad + (1+y)^{-2}(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)^2(2-2y-\beta_w y_n) + 2y(1+y)^{-2}[n^{-1}\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2] \\
&\quad + (1+y)^{-2}(2+y)^2n^{-1}\sum_{\ell_1=1}^{p}\sum_{\ell_2=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\mathbf{e}_{\ell_2}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_2}\Big[2(\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{e}_{\ell_2})^2 \\
&\quad + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_1})^2(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_{\ell_2})^2\Big] + (1+y)^{-2}\frac{(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)^2}{n}\Big[2\mathrm{tr}\mathbf{R}^2 + \beta_w p\Big] \\
&\quad - 2(1+y)^{-2}(2+y)^2n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left[2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right] \\
&\quad - 2(1+y)^{-2}(2+y)(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)n^{-1}\sum_{\ell=1}^{p}\Big[2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k \\
&\quad + (1+y)^{-2}(4+2y)(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)n^{-1}\sum_{\ell_1=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\Big[2\mathbf{e}_{\ell_1}^T\mathbf{R}^2\mathbf{e}_{\ell_1} + \beta_w\Big] \\
&\quad - 2(1+y)^{-1}(2+y)n^{-1}(2\mathrm{tr}\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{D}_1 + \beta_w\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_\ell) \\
&\quad + 2(1+y)^{-1}(4+2y)n^{-1}\sum_{\ell=1}^{p}\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{e}_\ell\left(2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right) \\
&\quad + 2(1+y)^{-1}(n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)n^{-1}\sum_{\ell=1}^{p}\left(2\mathbf{e}_\ell^T\mathbf{R}\mathbf{D}_1\mathbf{R}\mathbf{e}_\ell + \beta_w\sum_{k=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{D}_1\mathbf{R}^{1/2}\mathbf{e}_k\right) \\
&\quad - 2(1+y)^{-1}(2+y)n^{-1}\sum_{\ell_1=1}^{p}\sum_{\ell_2=1}^{p}\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_1}\mathbf{e}_{\ell_2}^T\mathbf{R}\mathbf{D}_1\mathbf{e}_{\ell_2}\Big[2(\mathbf{e}_{\ell_1}^T\mathbf{R}\mathbf{e}_{\ell_2})^2
\end{aligned}
$$

$$+\beta_w \sum_{k=1}^{p} (\mathbf{e}_k^T \mathbf{R}^{1/2} \mathbf{e}_{\ell_1})^2 (\mathbf{e}_k^T \mathbf{R}^{1/2} \mathbf{e}_{\ell_2})^2 \Bigg]$$
$$-2(1+y)^{-1} \frac{n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1}{n} \sum_{\ell_1=1}^{p} \mathbf{e}_{\ell_1}^T \mathbf{R}\mathbf{D}_1 \mathbf{e}_{\ell_1} \left[2\mathbf{e}_{\ell_1}^T \mathbf{R}^2 \mathbf{e}_{\ell_1} + \beta_w\right]$$

That is,

$$\begin{aligned}
\sigma^2 \;=\; & \left(\frac{1+2y}{1-y}\right)^2 \left(2n^{-1}\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2 + \beta_w n^{-1} \sum_{\ell=1}^{p} (\mathbf{e}_\ell^T \mathbf{R}^{1/2} \mathbf{D}_1 \mathbf{R}^{1/2} \mathbf{e}_\ell)^2\right) \\
& + \left(\frac{1+2y}{1-y}\right)^2 n^{-1} \sum_{\ell_1=1}^{p} \sum_{\ell_2=1}^{p} \mathbf{e}_{\ell_1}^T \mathbf{R}\mathbf{D}_1 \mathbf{e}_{\ell_1} \mathbf{e}_{\ell_2}^T \mathbf{R}\mathbf{D}_1 \mathbf{e}_{\ell_2} \left(2(\mathbf{e}_{\ell_1}^T \mathbf{R}\mathbf{e}_{\ell_2})^2 + \beta_w \sum_{k=1}^{p} (\mathbf{e}_k^T \mathbf{R}^{1/2} \mathbf{e}_{\ell_1})^2 (\mathbf{e}_k^T \mathbf{R}^{1/2} \mathbf{e}\right. \\
& -2\left(\frac{1+2y}{1-y}\right)^2 n^{-1} \sum_{\ell=1}^{p} \mathbf{e}_\ell^T \mathbf{R}\mathbf{D}_1 \mathbf{e}_\ell \\
& \left(2\mathbf{e}_\ell^T \mathbf{R}\mathbf{D}_1 \mathbf{R}\mathbf{e}_\ell + \beta_w \sum_{k=1}^{p} (\mathbf{e}_k^T \mathbf{R}^{1/2} \mathbf{e}_\ell)^2 \cdot \mathbf{e}_k^T \mathbf{R}^{1/2} \mathbf{D}_1 \mathbf{R}^{1/2} \mathbf{e}_k\right) \\
& + (1+y)^{-2} (n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)^2 (2 - 2y - \beta_w y_n) \\
& + 2(1+y)^{-2} y [n^{-1}\mathrm{tr}(\mathbf{R}\mathbf{D}_1)^2] \\
& + (1+y)^{-2} (n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1)^2 (2n^{-1}\mathrm{tr}\mathbf{R}^2 + \beta_w y_n) \\
& - \frac{2}{(1+y)^2} (n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1) \left(2n^{-1}\mathrm{tr}\mathbf{R}^2\mathbf{D}_1 + \beta_w n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1\right) \\
& + \frac{2}{(1+y)^2} (n^{-1}\mathrm{tr}\mathbf{R}\mathbf{D}_1) n^{-1} \sum_{\ell_1=1}^{p} \mathbf{e}_{\ell_1}^T \mathbf{R}\mathbf{D}_1 \mathbf{e}_{\ell_1} \left[2\mathbf{e}_{\ell_1}^T \mathbf{R}^2 \mathbf{e}_{\ell_1} + \beta_w\right].
\end{aligned}$$

Step 3: Thus we have that $[(\sigma_{ij0})_{i,j=1}^2]^{-1/2}(\mathrm{tr}\widehat{\mathbf{R}}\mathbf{D}_1 - \nu_1, \mathrm{tr}\widehat{\mathbf{R}}\mathbf{D}_1\widehat{\mathbf{R}}\mathbf{R}^{-1} - \nu_2)^T$ is asymptotically distributed as bivariate normal distribution with mean zero and identity covariance matrix. By the delta method, we have that $\sigma^{-1}p\sum_{k=1}^{K}\pi_k[\widehat{\theta}_k - (1+y_{n-2})^{-1}\theta_k]$ is asymptotically distributed as N(0, 1).

The proof of Theorem 4.1.2 is completed.

Proof of Theorem 4.1.3. Let the estimate of $\mathbf{R}^{-1}$ be

$$\widehat{\mathbf{R}^{-1}} = (\widehat{\theta}_1\mathbf{A}_1 + \ldots + \widehat{\theta}_K\mathbf{A}_K).$$

where $\widehat{\mathbf{R}^{-1}} = [\widehat{\mathbf{R}^{-1/2}}]^2$ and the estimate of $\mathbf{R}^{1/2}$ is $(\widehat{\mathbf{R}^{-1/2}})^{-1}$. Because the ex-

pression $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T[\mathrm{diag}(\mathbf{S})]^{-1/2}$ remains the same for any variances of $\mathbf{x}$, then we assume that $\mathrm{diag}(\mathbf{\Sigma}) = \mathbf{I}_p$ without loss of generality. Thus we have

$$\begin{aligned}
T_n &= (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T[\mathrm{diag}(\mathbf{S})]^{-1/2}\widehat{\mathbf{R}^{-1}}[\mathrm{diag}(\mathbf{S})]^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&= c\sum_{j=1}^{K}\theta_j(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{A}_j[\mathrm{diag}(\mathbf{S})]^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&\quad + \sum_{j=1}^{K}(\widehat{\theta}_j - c\theta_j)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{A}_j[\mathrm{diag}(\mathbf{S})]^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&= c\sum_{j=1}^{K}\theta_j(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{A}_j[\mathrm{diag}(\mathbf{S})]^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + O_p(p^{-1}) \\
&= c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T[\mathrm{diag}(\mathbf{S})]^{-1/2}\mathbf{R}^{-1}[\mathrm{diag}(\mathbf{S})]^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + O_p(p^{-1}) \\
&= c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{[\mathrm{diag}(\mathbf{S})]^{-1/2} - \mathbf{I}_p\}\mathbf{R}^{-1}[\mathrm{diag}(\mathbf{S})]^{-1/2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&\quad + c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}\{[\mathrm{diag}(\mathbf{S})]^{-1/2} - \mathbf{I}_p\}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + O_p(p^{-1}) \\
&= c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + 2c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{[\mathrm{diag}(\mathbf{S})]^{-1/2} - \mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&\quad + c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{[\mathrm{diag}(\mathbf{S})]^{-1/2} - \mathbf{I}_p\}\mathbf{R}^{-1}\{[\mathrm{diag}(\mathbf{S})]^{-1/2} - \mathbf{I}_p\}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + O_p(p^{-1}) \\
&= c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - c(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&\quad + \frac{3c}{4}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}^2\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&\quad + \frac{c}{4}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + O_p(p^{-1}).
\end{aligned}$$

Because

$$\mathrm{Var}(n(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)) = (2p + \kappa p n_1^{-1})\frac{n^2}{n_1^2} + (2p + \kappa p n_2^{-1})\frac{n^2}{n_2^2} + \frac{4pn^2}{n_1 n_2}, \quad (4.4.16)$$

then let

$$\begin{aligned}
&(2p)^{-1/2}nT_n \\
= \;& (2p)^{-1/2}cn(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&-(2p)^{-1/2}cn(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&+(2p)^{-1/2}n\frac{3c}{4}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}^2\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
&+(2p)^{-1/2}n\frac{c}{4}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) + O_p(p^{-1}). \quad (4.4.17)
\end{aligned}$$

Because $(2p)^{-1/2}cn(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)$ is the quadratic form of $\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2$, then $(2p)^{-1/2}cn(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)-\mathrm{E}(2p)^{-1/2}cn(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)$ follows a central limit theorem. In the following, we will prove that $-(2p)^{-1/2}cn(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)=\mathrm{E}[-(2p)^{-1/2}cn(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)]+o_p(1)$. Similarly, it can be proved that

$$\begin{aligned}&(2p)^{-1/2}n\frac{3c}{4}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}^2\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\\&-\mathrm{E}(2p)^{-1/2}n\frac{3c}{4}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}^2\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)=o_p(1)\end{aligned}$$

and

$$\begin{aligned}&(2p)^{-1/2}n\frac{c}{4}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\\&-\mathrm{E}(2p)^{-1/2}n\frac{c}{4}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)=o_p(1).\end{aligned}$$

Then we have

$$\begin{aligned}&c\sigma\Big[(2p)^{-1/2}nT_n-(2p)^{-1/2}cn\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\\&+(2p)^{-1/2}cn\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\qquad(4.4.18)\\&-(2p)^{-1/2}n\frac{3c}{4}\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}^2\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\\&-(2p)^{-1/2}n\frac{c}{4}\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\Big]\\\to\;&N(0,1)\qquad(4.4.19)\end{aligned}$$

where

$$\begin{aligned}\sigma^2\;=\;&\mathrm{Var}\Big[(2p)^{-1/2}n\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\Big]\\&+\mathrm{Var}\Big[(2p)^{-1/2}n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\Big].\end{aligned}$$

In fact, it will be proved that $\mathrm{Var}\Big[(2p)^{-1/2}cn(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-$

$\bar{\mathbf{x}}_2)\Big] = o(1)$.

Step 1. We will show that under H_0, we have

$$\begin{aligned} & -(2p)^{-1/2}cn(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ = & -\frac{(2p)^{-1/2}cn}{n-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(2p+\beta_w\sum_{\ell=1}^{p}\sum_{k=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^3\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k\right)+o_p(1) \end{aligned}$$

$$\begin{aligned} & (2p)^{-1/2}n\frac{3c}{4}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}^2\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ = & \frac{3cn(n-1)}{4(n-2)^2\sqrt{2p}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)[2p+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4]+o_p(1) \end{aligned}$$

$$\begin{aligned} & (2p)^{-1/2}n\frac{c}{4}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ = & \frac{cn(n-1)}{4(n-2)^2\sqrt{2p}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathrm{tr}\mathbf{R}^{-1}\left(2(\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell)^3+\beta_w\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell\sum_{f=1}^{p}(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_h)^2(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\right)_{h,\ell=1}^{p} \\ & +o_p(1) \end{aligned}$$

and

$$\begin{aligned} & \mathrm{Var}\left[(2p)^{-1/2}nc(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\right] \\ = & \frac{nc^2}{2p}\left(\frac{1}{n_1^2}+\frac{1}{n_2^2}\right)\left[4\mathrm{tr}\mathbf{R}^2+2p\beta_w+2\sum_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^2\mathbf{e}_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1}+\beta_w\mathrm{tr}\mathbf{R}^{-1}\right]. \end{aligned}$$

We have

$$\begin{aligned} & (2p)^{-1/2}n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ = & (2p)^{-1/2}n\bar{\mathbf{x}}_1^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\bar{\mathbf{x}}_1 \\ & +(2p)^{-1/2}n\bar{\mathbf{x}}_2^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\bar{\mathbf{x}}_2 \\ & -2(2p)^{-1/2}n\bar{\mathbf{x}}_1^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\bar{\mathbf{x}}_2 \end{aligned}$$

$$
\begin{aligned}
= \quad & \frac{n^3}{n_1^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_1^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=n_1+1}^{n_1+n_2}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_2^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n_1+n_2}\sum_{k=n_1+1}^{n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_2^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=n_1+1}^{n_1+n_2}\sum_{j=n_1+1}^{n_1+n_2}\sum_{k=n_1+1}^{n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& - \frac{2n^3}{n_1n_2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=n_1+1}^{n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& - \frac{2n^3}{n_1n_2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=n_1+1}^{n_1+n_2}\sum_{j=1}^{n_1}\sum_{k=n_1+1}^{n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
= \quad & \frac{n^3}{n_1^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{1\leqslant i\leqslant n_1}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_1^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{1\leqslant i\neq j\leqslant n_1}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_1^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{1\leqslant i\neq j\neq k\leqslant n_1}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_1^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=n_1+1}^{n_1+n_2}\sum_{j=1}^{n_1}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_1^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=n_1+1}^{n_1+n_2}\sum_{1\leqslant j\neq k\leqslant n_1}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_2^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_2^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{i=1}^{n_1}\sum_{n_1+1\leqslant j\neq k\leqslant n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
& + \frac{n^3}{n_2^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{n_1+1\leqslant i\leqslant n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_i\mathbf{r}_i\mathbf{R}^{1/2}\mathbf{e}_\ell
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{n^3}{n_2^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{n_1+1\leqslant i\neq j\leqslant n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell-n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_\ell\\
&+\frac{n^3}{n_2^2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{n_1+1\leqslant i\neq j\neq k\leqslant n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell-n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell\\
&-\frac{2n^3}{n_1n_2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{1\leqslant i\leqslant n_1}\sum_{k=n_1+1}^{n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell-n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_i\mathbf{r}_k\mathbf{R}^{1/2}\mathbf{e}_\ell\\
&-\frac{2n^3}{n_1n_2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{1\leqslant i\neq j\leqslant n_1}\sum_{k=n_1+1}^{n_1+n_2}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell-n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell\\
&-\frac{2n^3}{n_1n_2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{n_1+1\leqslant i\leqslant n_1+n_2}\sum_{j=1}^{n_1}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell-n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_i\mathbf{R}^{1/2}\mathbf{e}_\ell\\
&-\frac{2n^3}{n_1n_2(n-2)}(2p)^{-1/2}\sum_{\ell=1}^{p}\sum_{n_1+1\leqslant i\neq k\leqslant n_1+n_2}\sum_{j=1}^{n_1}\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_\ell-n^{-1}\}\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&-(2p)^{-1/2}cn(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\\
=\;&-\frac{(2p)^{-1/2}cn}{n-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(2p+\beta_w\sum_{\ell=1}^{p}\sum_{k=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^3\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k\right)+o_p(1)
\end{aligned}
$$

Because

$$
\begin{aligned}
&\mathrm{Var}\left(\sum_{k=1}^{p}\sum_{1\leqslant i\neq j\leqslant n_1}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k-n^{-1})\mathbf{e}_k^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_k\right.\\
&\qquad\left.+\sum_{k=1}^{p}\sum_{1\leqslant i\neq j\neq\ell\leqslant n_1}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k-n^{-1})\mathbf{e}_k^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k\right)\\
=\;&\frac{n_1^3}{n^4}(2\mathrm{tr}\mathbf{R}^2+\beta_w p)\\
&+\frac{n_1^3}{n^4}\sum_{k_1,k_2}[2(\mathbf{e}_{k_1}^T\mathbf{R}\mathbf{e}_{k_2})^2+\beta_w\sum_{h=1}^{p}(\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{e}_{k_1})^2(\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{e}_{k_2})^2]\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1}\\
&+\frac{n_1^3}{n^4}(2\mathrm{tr}\mathbf{R}^2+\beta_w p)+o(1)\\
=\;&\frac{2n_1^3}{n^4}(2\mathrm{tr}\mathbf{R}^2+\beta_w p)+\frac{n_1^3}{n^4}\sum_{k_1}[2\mathbf{e}_{k_1}^T\mathbf{R}^2\mathbf{e}_{k_1}+\beta_w]\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1}+o(1)
\end{aligned}
$$

$$= \frac{n_1^3}{n^4}\left[4\mathrm{tr}\mathbf{R}^2 + \beta_w 2p + 2\sum_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^2\mathbf{e}_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1} + \beta_w\mathrm{tr}\mathbf{R}^{-1}\right] + o(1),$$

$$\begin{aligned}
&\mathrm{Var}\Bigg(\sum_{k=1}^{p}\sum_{n_1+1\leqslant i\neq j\leqslant n}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1})\mathbf{e}_k^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_k \\
&\qquad + \sum_{k=1}^{p}\sum_{n_1+1\leqslant i\neq j\neq \ell\leqslant n}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1})\mathbf{e}_k^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k\Bigg) \\
= &\frac{n_2^3}{n^4}\left[4\mathrm{tr}\mathbf{R}^2 + \beta_w 2p + 2\sum_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^2\mathbf{e}_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1} + \beta_w\mathrm{tr}\mathbf{R}^{-1}\right] + o(1),
\end{aligned}$$

$$\begin{aligned}
&\mathrm{Var}\Bigg(\sum_{k=1}^{p}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1})\mathbf{e}_k^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_k \\
&\qquad + \sum_{k=1}^{p}\sum_{i=1}^{n_1}\sum_{n_1+1\leqslant j\neq \ell\leqslant n}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1})\mathbf{e}_k^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k\Bigg) \\
= &\frac{n_1 n_2^2}{n^4}\left[4\mathrm{tr}\mathbf{R}^2 + \beta_w 2p + 2\sum_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^2\mathbf{e}_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1} + \beta_w\mathrm{tr}\mathbf{R}^{-1}\right] + o(1),
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Var}\Bigg(\sum_{k=1}^{p}\sum_{i=n_1+1}^{n}\sum_{j=1}^{n_1}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1})\mathbf{e}_k^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_j^T\mathbf{R}^{1/2}\mathbf{e}_k \\
&\qquad + \sum_{k=1}^{p}\sum_{i=n_1+1}^{n}\sum_{1\leqslant j\neq \ell\leqslant n_1}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_k - n^{-1})\mathbf{e}_k^T\mathbf{R}^{-1/2}\mathbf{r}_j\mathbf{r}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k\Bigg) \\
= &\frac{n_2 n_1^2}{n^4}\left[4\mathrm{tr}\mathbf{R}^2 + \beta_w 2p + 2\sum_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^2\mathbf{e}_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1} + \beta_w\mathrm{tr}\mathbf{R}^{-1}\right] + o(1),
\end{aligned}$$

then we have

$$\mathrm{Var}\Big[(2p)^{-1/2}cn(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\Big]$$

$$= \frac{nc^2}{2p}\left(\frac{1}{n_1^2}+\frac{1}{n_2^2}\right)\left[4\mathrm{tr}\mathbf{R}^2+2p\beta_w+2\sum_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^2\mathbf{e}_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1}+\beta_w\mathrm{tr}\mathbf{R}^{-1}\right]\to 0.$$

Thus we have

$$\begin{aligned}
&(2p)^{-1/2}n\frac{c}{4}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\\
=\;&\frac{cn(n-1)}{4(n-2)^2\sqrt{2p}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathrm{tr}\mathbf{R}^{-1}\left(2(\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell)^3+\beta_w\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell\sum_{f=1}^{p}(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_h)^2(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\right)_{h,\ell=1}^{p}\\
&+o_p(1). \qquad (4.4.20)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
& (2p)^{-1/2} n \frac{3c}{4} \mathrm{E} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}^2 \mathbf{R}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
= & \frac{3cn(n-1)}{4(n-2)^2\sqrt{2p}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) [2p + \beta_w \sum_{k=1}^{p} \sum_{\ell=1}^{p} (\mathbf{e}_k^T \mathbf{R}^{1/2} \mathbf{e}_\ell)^4] + o_p(1) \quad (4.4.21)
\end{aligned}$$

where

$$\begin{aligned}
& (2p)^{-1/2} n \frac{3c}{4} \mathrm{E} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}^2 \mathbf{R}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \\
= & (2p)^{-1/2} n \frac{3c}{4} \mathrm{E} \bar{\mathbf{x}}_1^T \{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}^2 \mathbf{R}^{-1} \bar{\mathbf{x}}_1 \\
& + (2p)^{-1/2} n \frac{3c}{4} \mathrm{E} \bar{\mathbf{x}}_2^T \{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}^2 \mathbf{R}^{-1} \bar{\mathbf{x}}_2 \\
& - 2(2p)^{-1/2} n \frac{3c}{4} \mathrm{E} \bar{\mathbf{x}}_1^T \{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}^2 \mathbf{R}^{-1} \bar{\mathbf{x}}_2 \\
= & \frac{n^4 (2p)^{-1/2}}{n_1^2 (n-2)^2} \frac{3c}{4} \mathrm{E} \sum_{\ell=1}^{p} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \sum_{k=1}^{n_1} \{\mathbf{e}_\ell^T \mathbf{R}^{1/2} \mathbf{r}_i^T \mathbf{r}_i \mathbf{R}^{1/2} \mathbf{e}_\ell - n^{-1}\}^2 \mathbf{e}_\ell^T \mathbf{R}^{-1/2} \mathbf{r}_j \mathbf{r}_k \mathbf{R}^{1/2} \mathbf{e}_\ell \\
& + \frac{n^4 (2p)^{-1/2}}{n_1^2 (n-2)^2} \frac{3c}{4} \mathrm{E} \sum_{\ell=1}^{p} \sum_{i=n_1+1}^{n_1+n_2} \sum_{j=1}^{n_1} \sum_{k=1}^{n_1} \{\mathbf{e}_\ell^T \mathbf{R}^{1/2} \mathbf{r}_i^T \mathbf{r}_i \mathbf{R}^{1/2} \mathbf{e}_\ell - n^{-1}\}^2 \mathbf{e}_\ell^T \mathbf{R}^{-1/2} \mathbf{r}_j \mathbf{r}_k \mathbf{R}^{1/2} \mathbf{e}_\ell \\
& + \frac{n^4 (2p)^{-1/2}}{n_2^2 (n-2)^2} \frac{3c}{4} \mathrm{E} \sum_{\ell=1}^{p} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n_1+n_2} \sum_{k=n_1+1}^{n_1+n_2} \{\mathbf{e}_\ell^T \mathbf{R}^{1/2} \mathbf{r}_i^T \mathbf{r}_i \mathbf{R}^{1/2} \mathbf{e}_\ell - n^{-1}\}^2 \mathbf{e}_\ell^T \mathbf{R}^{-1/2} \mathbf{r}_j \mathbf{r}_k \mathbf{R}^{1/2} \mathbf{e}_\ell \\
& + \frac{n^4 (2p)^{-1/2}}{n_2^2 (n-2)^2} \frac{3c}{4} \mathrm{E} \sum_{\ell=1}^{p} \sum_{i=n_1+1}^{n_1+n_2} \sum_{j=n_1+1}^{n_1+n_2} \sum_{k=n_1+1}^{n_1+n_2} \{\mathbf{e}_\ell^T \mathbf{R}^{1/2} \mathbf{r}_i^T \mathbf{r}_i \mathbf{R}^{1/2} \mathbf{e}_\ell - n^{-1}\}^2 \mathbf{e}_\ell^T \mathbf{R}^{-1/2} \mathbf{r}_j \mathbf{r}_k \mathbf{R}^{1/2} \mathbf{e}_\ell \\
& - \frac{2n^4 (2p)^{-1/2}}{n_1 n_2 (n-2)^2} \frac{3c}{4} \mathrm{E} \sum_{\ell=1}^{p} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \sum_{k=n_1+1}^{n_1+n_2} \{\mathbf{e}_\ell^T \mathbf{R}^{1/2} \mathbf{r}_i^T \mathbf{r}_i \mathbf{R}^{1/2} \mathbf{e}_\ell - n^{-1}\}^2 \mathbf{e}_\ell^T \mathbf{R}^{-1/2} \mathbf{r}_j \mathbf{r}_k \mathbf{R}^{1/2} \mathbf{e}_\ell \\
& - \frac{2n^4 (2p)^{-1/2}}{n_1 n_2 (n-2)^2} \frac{3c}{4} \mathrm{E} \sum_{\ell=1}^{p} \sum_{i=n_1+1}^{n_1+n_2} \sum_{j=1}^{n_1} \sum_{k=n_1+1}^{n_1+n_2} \{\mathbf{e}_\ell^T \mathbf{R}^{1/2} \mathbf{r}_i^T \mathbf{r}_i \mathbf{R}^{1/2} \mathbf{e}_\ell - n^{-1}\}^2 \mathbf{e}_\ell^T \mathbf{R}^{-1/2} \mathbf{r}_j \mathbf{r}_k \mathbf{R}^{1/2} \mathbf{e}_\ell + o(1) \\
= & \frac{3cn(n-1)}{4(n-2)^2\sqrt{2p}} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) [2p + \beta_w \sum_{k=1}^{p} \sum_{\ell=1}^{p} (\mathbf{e}_k^T \mathbf{R}^{1/2} \mathbf{e}_\ell)^4] + o(1)
\end{aligned}$$

Moreover, we have

$$(2p)^{-1/2} n \frac{c}{4} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\} \mathbf{R}^{-1} \{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$
\begin{aligned}
&= (2p)^{-1/2} n \frac{c}{4} \mathrm{tr}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S}) - \mathbf{I}_p\} \\
&= \frac{(2p)^{-1/2} n^4}{(n-2)^2 n_1^2} \frac{c}{4} \mathrm{tr}\mathbf{R}^{-1} \sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\sum_{m=1}^{n_1} \Bigg(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
&\quad \{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\} \Bigg)_{h,\ell=1}^{p} \\
&\quad + \frac{(2p)^{-1/2} n^4}{(n-2)^2 n_1^2} \frac{c}{4} \mathrm{tr}\mathbf{R}^{-1} \sum_{i=n_1+1}^{n}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\sum_{m=1}^{n_1} \Bigg(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
&\quad \{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\} \Bigg)_{h,\ell=1}^{p} \\
&\quad + \frac{(2p)^{-1/2} n^4}{(n-2)^2 n_1 n_2} \frac{c}{4} \mathrm{tr}\mathbf{R}^{-1} \sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}\sum_{k=1}^{n_1}\sum_{m=1}^{n_1} \Bigg(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
&\quad \{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\} \Bigg)_{h,\ell=1}^{p} \\
&\quad + \frac{(2p)^{-1/2} n^4}{(n-2)^2 n_1 n_2} \frac{c}{4} \mathrm{tr}\mathbf{R}^{-1} \sum_{i=n_1+1}^{n}\sum_{j=n_1+1}^{n}\sum_{k=1}^{n_1}\sum_{m=1}^{n_1} \Bigg(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
&\quad \{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\} \Bigg)_{h,\ell=1}^{p} \\
&\quad + \frac{(2p)^{-1/2} n^4}{(n-2)^2 n_1 n_2} \frac{c}{4} \mathrm{tr}\mathbf{R}^{-1} \sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=n_1+1}^{n}\sum_{m=1}^{n_1} \Bigg(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
&\quad \{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\} \Bigg)_{h,\ell=1}^{p} \\
&\quad + \frac{(2p)^{-1/2} n^4}{(n-2)^2 n_1 n_2} \frac{c}{4} \mathrm{tr}\mathbf{R}^{-1} \sum_{i=n_1+1}^{n}\sum_{j=1}^{n_1}\sum_{k=n_1+1}^{n}\sum_{m=1}^{n_1} \Bigg(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
&\quad \{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\} \Bigg)_{h,\ell=1}^{p} \\
&\quad + \frac{(2p)^{-1/2} n^4}{(n-2)^2 n_2^2} \frac{c}{4} \mathrm{tr}\mathbf{R}^{-1} \sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}\sum_{k=n_1+1}^{n}\sum_{m=1}^{n_1} \Bigg(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell \\
&\quad \{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\} \Bigg)_{h,\ell=1}^{p} \\
&\quad + \frac{c(2p)^{-1/2} n^4}{4(n-2)^2 n_2^2} \mathrm{tr}\mathbf{R}^{-1} \sum_{i=n_1+1}^{n}\sum_{j=n_1+1}^{n}\sum_{k=n_1+1}^{n}\sum_{m=1}^{n_1} \Big(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell
\end{aligned}
$$

$$\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\Big)_{h,\ell=1}^{p}$$
$$+\frac{c(2p)^{-1/2}n^4}{4(n-2)^2n_1^2}\mathrm{tr}\mathbf{R}^{-1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\sum_{m=n_1+1}^{n}\Big(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell$$
$$\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\Big)_{h,\ell=1}^{p}$$
$$+\frac{c(2p)^{-1/2}n^4}{4(n-2)^2n_1^2}\mathrm{tr}\mathbf{R}^{-1}\sum_{i=n_1+1}^{n}\sum_{j=1}^{n_1}\sum_{k=1}^{n_1}\sum_{m=n_1+1}^{n}\Big(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell$$
$$\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\Big)_{h,\ell=1}^{p}$$
$$+\frac{c(2p)^{-1/2}n^4}{4(n-2)^2n_1n_2}\mathrm{tr}\mathbf{R}^{-1}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}\sum_{k=1}^{n_1}\sum_{m=n_1+1}^{n}\Big(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell$$
$$\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\Big)_{h,\ell=1}^{p}$$
$$+\frac{c(2p)^{-1/2}n^4}{4(n-2)^2n_1n_2}\mathrm{tr}\mathbf{R}^{-1}\sum_{i=n_1+1}^{n}\sum_{j=n_1+1}^{n}\sum_{k=1}^{n_1}\sum_{m=n_1+1}^{n}\Big(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell$$
$$\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\Big)_{h,\ell=1}^{p}$$
$$+\frac{(2p)^{-1/2}n^4}{(n-2)^2n_1n_2}\frac{c}{4}\mathrm{tr}\mathbf{R}^{-1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\sum_{k=n_1+1}^{n}\sum_{m=n_1+1}^{n}\Big(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell$$
$$\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\Big)_{h,\ell=1}^{p}$$
$$+\frac{(2p)^{-1/2}n^4}{(n-2)^2n_1n_2}\frac{c}{4}\mathrm{tr}\mathbf{R}^{-1}\sum_{i=n_1+1}^{n}\sum_{j=1}^{n_1}\sum_{k=n_1+1}^{n}\sum_{m=n_1+1}^{n}\Big(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell$$
$$\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\Big)_{h,\ell=1}^{p}$$
$$+\frac{(2p)^{-1/2}n^4}{(n-2)^2n_2^2}\frac{c}{4}\mathrm{tr}\mathbf{R}^{-1}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n}\sum_{k=n_1+1}^{n}\sum_{m=n_1+1}^{n}\Big(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell$$
$$\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\Big)_{h,\ell=1}^{p}$$
$$\frac{(2p)^{-1/2}n^4c}{4(n-2)^2n_2^2}\mathrm{tr}\mathbf{R}^{-1}\sum_{i=n_1+1}^{n}\sum_{j=n_1+1}^{n}\sum_{k=n_1+1}^{n}\sum_{m=n_1+1}^{n}\Big(\{\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_i\mathbf{r}_i^T\mathbf{R}^{1/2}\mathbf{e}_h - n^{-1}\}\mathbf{e}_h^T\mathbf{R}^{1/2}\mathbf{r}_j\mathbf{r}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell$$

$$\{\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{r}_m\mathbf{r}_m^T\mathbf{R}^{1/2}\mathbf{e}_\ell - n^{-1}\}\Big)_{h,\ell=1}^{p} + o_p(1).$$

Similar to (4.4.20), it is obtained that

$$\begin{aligned}
& (2p)^{-1/2}n\frac{c}{4}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\
= & \frac{cn(n-1)}{4(n-2)^2\sqrt{2p}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathrm{tr}\mathbf{R}^{-1}\left(2(\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell)^3+\beta_w\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell\sum_{f=1}^{p}(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_h)^2(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\right)_{h,\ell=1}^{p} \\
& +o_p(1). \qquad (4.4.22)
\end{aligned}$$

Step 2. Under H_0, we have $(2p)^{-1/2}cn\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) = cnp(2p)^{-1/2}(n_1^{-1}+n_2^{-1})$, and

$$\mathrm{Var}(cn(2p)^{-1/2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)) = \frac{c^2(2p+\kappa pn_1^{-1})}{2p}\frac{n^2}{n_1^2}+\frac{c^2(2p+\kappa pn_2^{-1})}{2p}\frac{n^2}{n_2^2}+\frac{4c^2n^2p}{2pn_1n_2}.$$

Then by (4.4.16), (4.4.21), (4.4.22), (4.4.20) and (4.4.19),we have

$$\frac{nT_n - c\mu_0}{c\sigma_0} \xrightarrow{H_0} N(0,1)$$

where

$$\begin{aligned}
\mu_0 = & n\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\
& -n\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\
& +n\frac{3}{4}\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}^2\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\
& +n\frac{1}{4}\mathrm{E}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\
= & np(n_1^{-1}+n_2^{-1}) - \frac{n}{n-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(2p+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^3\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k\right) \\
& +\frac{3n(n-1)}{4(n-2)^2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)[2p+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4] \\
& +\frac{n(n-1)}{4(n-2)^2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathrm{tr}\mathbf{R}^{-1}\mathbf{A}_0
\end{aligned}$$

and

$$\begin{aligned}\sigma_0^2 &= \mathrm{Var}(n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)) \\ &\quad +\mathrm{Var}\left[n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\right] \\ &= (2p+\kappa p n_1^{-1})\frac{n^2}{n_1^2}+(2p+\kappa p n_2^{-1})\frac{n^2}{n_2^2}+\frac{4n^2p}{n_1n_2} \\ &\quad +n\left(\frac{1}{n_1^2}+\frac{1}{n_2^2}\right)\left[4\mathrm{tr}\mathbf{R}^2+2p\beta_w+2\sum_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^2\mathbf{e}_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1}+\beta_w\mathrm{tr}\mathbf{R}^{-1}\right].\end{aligned}$$

with $\mathbf{A}_0$ being a $p\times p$ matrix with the (h,ℓ) element $a_{h,\ell}$ being $a_{h,\ell}=2(\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell)^3+\beta_w\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell\sum_{f=1}^p(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_h)^2(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2$. In fact, we have

$$\frac{nT_n-\widehat{c}\widehat{\mu}_0}{\widehat{c}\widehat{\sigma}_0}\xrightarrow{H_0}N(0,1)$$

where $\widehat{\mu}_0$, $\widehat{\sigma}_0^2$ and $\widehat{\mathbf{A}}_0$ are obtained by replacing $\mathbf{R}$ in μ_0, σ_0^2 and $\mathbf{A}_0$ by $\tilde{\mathbf{R}}$, $\widehat{c}=\frac{1}{1+p/(n-2)}$ and

$$\tilde{\mathbf{R}}=\left\{\mathrm{diag}\left[\left(\sum_{k=1}^K\widehat{\theta}_k\mathbf{A}_k\right)^{-1}\right]\right\}^{-1/2}\left(\sum_{k=1}^K\widehat{\theta}_k\mathbf{A}_k\right)^{-1}\left\{\mathrm{diag}\left[\left(\sum_{k=1}^K\widehat{\theta}_k\mathbf{A}_k\right)^{-1}\right]\right\}^{-1/2}.$$

The proof of Theorem 4.1.3 is completed.

Proof of Theorem 4.1.4. We have

$$\begin{aligned}&(2p)^{-1/2}nT_n \\ =\ &(2p)^{-1/2}n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T[\mathrm{diag}(\mathbf{S})]^{-1/2}\widehat{\mathbf{R}^{-1}}[\mathrm{diag}(\mathbf{S})]^{-1/2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ =\ &cn(2p)^{-1/2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ &-cn(2p)^{-1/2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathrm{diag}(\mathbf{\Sigma})\}\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ &+(2p)^{-1/2}\frac{3c}{4}n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathrm{diag}(\mathbf{\Sigma})\}^2\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ &+(2p)^{-1/2}\frac{c}{4}n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathrm{diag}(\mathbf{\Sigma})\}\mathbf{\Sigma}^{-1}\{\mathrm{diag}(\mathbf{S})-\mathbf{\Sigma}\}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)+O_p(p^{-1}).\end{aligned}$$

Under $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu_2}$, we have

$$\mathrm{E}[(2p)^{-1/2}cn(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)] = \frac{cn}{\sqrt{2p}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\frac{cnp}{\sqrt{2p}}(n_1^{-1}+n_2^{-1})$$

and

$$\begin{aligned} & \mathrm{Var}[c(2p)^{-1/2}n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)] \\ = & \frac{c^2}{2p}(2p+\kappa p n_1^{-1})\frac{n^2}{n_1^2}+\frac{c^2}{2p}(2p+\kappa p n_2^{-1})\frac{n^2}{n_2^2}+\frac{4c^2n^2p}{n_1n_2} \\ & +4c^2n^2(n_1^{-1}+n_2^{-1})(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2). \end{aligned}$$

Moreover, similar to the proofs of Theorem 4.1.3, we have

$$\begin{aligned} & cn(2p)^{-1/2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathrm{diag}(\boldsymbol{\Sigma})\}\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ = & -\frac{(2p)^{-1/2}cn}{n-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(2p+\beta_w\sum_{\ell=1}^{p}\sum_{k=1}^{p}(\mathbf{e}_\ell^T\mathbf{R}^{1/2}\mathbf{e}_k)^3\mathbf{e}_\ell^T\mathbf{R}^{-1/2}\mathbf{e}_k\right)+o_p(1), \end{aligned}$$

$$\begin{aligned} & (2p)^{-1/2}\frac{3c}{4}n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathrm{diag}(\boldsymbol{\Sigma})\}^2\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ = & \frac{3cn(n-1)}{4(n-2)^2\sqrt{2p}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)[2p+\beta_w\sum_{k=1}^{p}\sum_{\ell=1}^{p}(\mathbf{e}_k^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^4]+o_p(1), \end{aligned}$$

$$\begin{aligned} & (2p)^{-1/2}\frac{c}{4}n(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathrm{diag}(\boldsymbol{\Sigma})\}\boldsymbol{\Sigma}^{-1}\{\mathrm{diag}(\mathbf{S})-\boldsymbol{\Sigma}\}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2) \\ = & \frac{cn(n-1)}{4(n-2)^2\sqrt{2p}}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\mathrm{tr}\mathbf{R}^{-1}\left(2(\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell)^3+\beta_w\mathbf{e}_h^T\mathbf{R}\mathbf{e}_\ell\sum_{f=1}^{p}(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_h)^2(\mathbf{e}_f^T\mathbf{R}^{1/2}\mathbf{e}_\ell)^2\right)_{h,\ell=1}^{p} \\ & +o_p(1). \end{aligned}$$

and

$$\mathrm{Var}\left[(2p)^{-1/2}nc(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^T\{\mathrm{diag}(\mathbf{S})-\mathbf{I}_p\}\mathbf{R}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\right]$$

$$= \frac{nc^2}{2p}\left(\frac{1}{n_1^2}+\frac{1}{n_2^2}\right)\left[4\mathrm{tr}\mathbf{R}^2+2p\beta_w+2\sum_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^2\mathbf{e}_{k_1}\mathbf{e}_{k_1}^T\mathbf{R}^{-1}\mathbf{e}_{k_1}+\beta_w\mathrm{tr}\mathbf{R}^{-1}\right].$$

Then we have

$$\frac{nT_n-c\mu_0-cn\delta_n}{c\sqrt{\sigma_0^2+4n^2(n_1^{-1}+n_2^{-1})\delta_n}}\to N(0,1).$$

where $\delta_n=\boldsymbol{\mu}_d^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_d$ with $\boldsymbol{\mu}_d=\boldsymbol{\mu}_1-\boldsymbol{\mu}_2$. The asymptotic power function for $H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ is as follows

$$Q(\delta_n,\sigma_0|\alpha)=\Phi\left(\frac{-z_\alpha\sigma_0+n\delta_n}{\sqrt{\sigma_0^2+4n^2(n_1^{-1}+n_2^{-1})\delta_n}}\right)$$

where $-z_\alpha$ is the critical value of $N(0,1)$ at the level α and Φ is the cdf of $N(0,1)$. Then we have $Q(\delta_n,\sigma_0|\alpha)\geqslant\Phi(-z_\alpha)$. That is, the test T_n is asymptotically unbiased.

Power Comparison: We consider the Gaussian case. The power function of the proposed test in this paper is approximately equal to

$$Q(\delta_n,\sigma_0|\alpha)\approx\Phi\left(\frac{-z_\alpha\sqrt{2pn^2(n_1^{-1}+n_2^{-1})^2}+n\delta_n}{\sqrt{2pn^2(n_1^{-1}+n_2^{-1})^2+4n^2(n_1^{-1}+n_2^{-1})\delta_n}}\right).$$

Bai and Saranadasa (1996)'s test has the power as follows

$$Q_{bs}(\boldsymbol{\mu}_d,\boldsymbol{\Sigma}|\alpha)=\Phi\left(-z_\alpha+\frac{\|\boldsymbol{\mu}_d\|^2}{\sqrt{2(n_1^{-1}+n_2^{-1})^2\mathrm{tr}\boldsymbol{\Sigma}^2}}\right).$$

If $\delta_n=o(1)$, then $\delta_n/\sqrt{2p(n_1^{-1}+n_2^{-1})^2+4(n_1^{-1}+n_2^{-1})\delta_n}\approx\delta_n/[\sqrt{2p}(n_1^{-1}+n_2^{-1})]$. When $\boldsymbol{\Sigma}$ is the identity matrix, we approximately have $Q(\delta_n,\sigma_0|\alpha)=Q_{bs}(\boldsymbol{\mu}_d,\boldsymbol{\Sigma}|\alpha)$. Moreover, $Q_{bs}(\boldsymbol{\mu}_d,\boldsymbol{\Sigma}|\alpha)$ is the increasing function of $\|\boldsymbol{\mu}_d\|^2/[\sqrt{2\mathrm{tr}\boldsymbol{\Sigma}^2}(n_1^{-1}+n_2^{-1})]$. Let $\boldsymbol{\Sigma}=\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$, the eigen-decomposition of $\boldsymbol{\Sigma}$, where $\boldsymbol{\Gamma}$ is a $p\times p$ orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix with k-th element λ_k. We represent $\boldsymbol{\mu}_d$ in the

coordinate system of $\mathbf{\Gamma}$ as $\boldsymbol{\mu}_d = \mathbf{\Gamma}\mathbf{a}$. Then

$$\delta_n/\sqrt{p} = \boldsymbol{\mu}_d^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_d = p^{-1/2} \sum_{k=1}^{p} a_k^2/\lambda_k,$$

where $\mathbf{a} = (a_1, \cdots, a_p)^T$ and

$$\|\boldsymbol{\mu}_d\|^2/\sqrt{\mathrm{tr}\{\mathbf{\Sigma}^2\}} = \sum_{k=1}^{p} a_k^2/(\sum_{k=1}^{p} \lambda_k^2)^{1/2}.$$

By Cauchy-Schwarz inequality $(p^{-1}\sum_{k=1}^{p} a_k^2/\lambda_k)^2(p^{-1}\sum_{k=1}^{p}\lambda_k)^2 \geqslant (p^{-1}\sum_{k=1}^{p} a_k^2)^2$ and $p^{-1}\sum_{k=1}^{p}\lambda_k^2 \geqslant (p^{-1}\sum_{k=1}^{p}\lambda_k)^2$, we have

$$(p^{-1}\sum_{k=1}^{p} a_k^2/\lambda_k)^2(p^{-1}\sum_{k=1}^{p}\lambda_k^2) \geqslant (p^{-1}\sum_{k=1}^{p} a_k^2)^2.$$

Then we have

$$\sum_{k=1}^{p} a_k^2/\lambda_k \geqslant \frac{\sum_{k=1}^{p} a_k^2}{\sqrt{\sum_{k=1}^{p}\lambda_k^2/p}}$$

or

$$\delta_n/[\sqrt{2p}(n_1^{-1}+n_2^{-1})] \geqslant \|\boldsymbol{\mu}_d\|^2/[\sqrt{2\mathrm{tr}\mathbf{\Sigma}^2}(n_1^{-1}+n_2^{-1})].$$

Thus we have

$$Q(\delta_n, \sigma_0|\alpha) \geqslant Q_{bs}(\boldsymbol{\mu}_d, \mathbf{\Sigma}|\alpha).$$

The proof of Theorem 4.1.4 is completed.

Chapter 5

Conclusion and Extension

5.1 Conclusion

In the first project, we propose a new joint feature screening method for generalized time varying coefficient model. The new method considers the joint effect among the predictors, so it can outperform the existing marginal methods when there exists some important variables that are marginally independent of the response. We also propose an efficient algorithm to carry on the whole screening process, and we also show that this algorithm possesses the accent property. The sure screening property of the new method is established which guarantees the important variables could be selected with an overwhelming probability. The numerical studies show that the new method can perform very well even if some important variables are marginally independent of the response.

In the second project, we proposed a new testing method for two sample mean problem for high dimension data which involves a new estimation method for the covariance matrix. We first prove the accuracy of our estimation method, and then the asymptotic distributions of the test statistic under both the null and alternative hypothesis are derived. Based on both the simulation studies and the real data example, we can see that the new method can outperform other existing methods when the variables are strongly correlated.

5.2 Extension

For the first project, we propose a new joint feature screening method for generalized time varying coefficient model. The new method considers the joint effect among the predictors, it does not possesses the ranking consistency. That means the new method would select some wrong predictors in the screening stage. In the future, we can focus on developing another joint screening method considering the effect among the predictors, and it also possesses the ranking consistency so the unimportant variables can be separated from important predictors provided an ideal cutoff.

For the second one, we proposed a new testing method for two sample mean problem for high dimension data which involves a new estimation method for the covariance matrix. In the portfolio risk estimation and optimization problems, a good estimator for the volatility matrix is very important. We may extend our new estimation method in the application of portfolio risk estimation and optimization.

Bibliography

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica*, 311–329.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

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

Chen, L. S., Paul, D., Prentice, R. L., and Wang, P. (2011). A regularized hotellings t2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association*, 106(496):1345–1360.

Chen, S. X., Qin, Y.-L., et al. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38(2):808–835.

Dempster, A. P. (1958). A high dimensional two sample significance test. *The Annals of Mathematical Statistics*, 995–1010.

Dempster, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics*, 16(1):41–50.

Donoho, D. L. (2005). Neighborly polytopes and sparse solutions of underdetermined linear equations. Technical report.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

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. PMID: 22279246.

Fan, J., Han, F., and Liu, H. (2014a). Challenges of big data analysis. *National science review*, 1(2):293–314.

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. and Li, R. (2007). Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006*, 595–622.

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., Ma, Y., and Dai, W. (2014b). 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. *The Journal of Machine Learning Research*, 10:2013–2038.

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

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

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.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, 1356–1378.

Lee, S. H., Lim, J., Li, E., Vannucci, M., and Petkova, E. (2012). Order test for high-dimensional two-sample means. *Journal of Statistical Planning and Inference*, 142(9):2719–2725.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a). Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877.

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. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

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.

Lopes, M., Jacob, L., and Wainwright, M. J. (2011a). A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, 1206–1214.

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

Pan, G. and Zhou, W. (2011). Central limit theorem for hotelling's t 2 statistic under large dimension. *The Annals of Applied Probability*, 1860–1910.

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.

Srivastava, M. S., Katayama, S., and Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis*, 114:349–358.

Thulin, M. (2014). A high-dimensional two-sample test for the mean using random subspaces. *Computational Statistics & Data Analysis*, 74:26–38.

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

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 224–244.

Xu, C. and Chen, J. (2014). The sparse mle for ultrahigh-dimensional feature screening. *Journal of the American Statistical Association*, 109(507):1257–1269.

Yang, G., Yu, Y., Li, R., and Buu, A. (2016). Feature screening in ultrahigh dimensional coxs model. *Statistical Sinica*.

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

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509.

Songshan Yang

songshan.yang@gmail.com
814-321-7332

EDUCATION

PhD Candidate August 2013-August 2018
- Methodology Center, Department of Statistics, Pennsylvania State University
  Advisor: Verne M. Willaman Professor Runze Li

Bachelor of Science in Statistics Aug. 2013
- Department of Statistics, Beijing Normal University, China
- Rank 1/51

RESEARCH INTERESTS

- Variable selection and feature screening for high dimensional data
- Statistical inference for high dimensional data

EXPERIENCE

Data Science Intern Jun. 2017 - Aug. 2017
Liberty Mutual Insurance, Marketing Analytics (Boston, MA)

PUBLICATIONS

Adjacency Matrix Comparison for Stochastic Block Models
Guangren Yang, **Songshan Yang**, Shenping Yang, Wang Zhou
Minor revision at ***Random Matrices: Theory and Applications***, 2017.

Feature Screening in Ultra-high Dimensional Generalized Varying-coefficient Models
Guangren Yang, **Songshan Yang**, Runze Li
Revision for ***Statistica Sinica***, 2017.

A Time-varying Effect Model for Examining Group Differences in Trajectories of Zero-inflated Count Outcomes with Applications in Substance Abuse Research.
Songshan Yang, James A. Cranford, Jennifer M. Jester, Runze Li, Robert A. Zucker, Anne Buu
Statistics in Medicine, **36**, 827-837, 2017.

Time-varying Effect Model for Studying Gender Difference in Health Behavior
Songshan Yang, James A. Cranford, Runze Li, Robert A. Zucker, Anne Buu
Statistical Methods in Medical Research, **26(6)**, 2812-2820, 2017.

New Optimal Weight Combination Model for Forecasting Precipitation
Songshan Yang, Xiaohua Yang, Rong Jiang, Yiche Zhang
Mathematical Problems in Engineering,
2012(1), 1-13, 2012.

New Refining Stratification Method for Logging Curve
Songshan Yang, Xiaohua Yang, Guodong Lv
Nonlinear Science Letter C: Nano, Biology and Environment, **2(1)**, 33-40, 2012.

Fuzzy Decision-making Model Based on Dynamic Programming and Its Application in Flood Control
Qilong Gu, Xiaohua Yang, **Songshan Yang**, Tianbai Zhao
Nonlinear Science Letter C: Nano, Biology and Environment, **2(1)**, 15-22, 2012.