

**Statistica Sinica Preprint No: SS-2016-0032R1**

<b>Title</b>	Two-Sample Tests for High-Dimensional Linear Regression with an Application to Detecting Interactions
<b>Manuscript ID</b>	SS-2016-0032R1
<b>URL</b>	<a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a>
<b>DOI</b>	10.5705/ss.202016.0032
<b>Complete List of Authors</b>	Yin Xia Tianxi Cai and T. Tony Cai
<b>Corresponding Author</b>	Yin Xia
<b>E-mail</b>	xiayin@email.unc.edu

# Two-Sample Tests for High-Dimensional Linear Regression with an Application to Detecting Interactions

Yin Xia<sup>1</sup>, Tianxi Cai<sup>2</sup>, and T. Tony Cai<sup>3</sup>

## Abstract

Motivated by applications in genomics, we consider in this paper global and multiple testing for the comparisons of two high-dimensional linear regression models. A procedure for testing the equality of the two regression vectors globally is proposed and shown to be particularly powerful against sparse alternatives. We then introduce a multiple testing procedure for identifying unequal coordinates while controlling the false discovery rate and false discovery proportion. Theoretical justifications are provided to guarantee the validity of the proposed tests and optimality results are established under sparsity assumptions on the regression coefficients. The proposed testing procedures are easy to implement. Numerical properties of the procedures

---

<sup>1</sup>Department of Statistics, School of Management, Fudan University and Department of Statistics & Operations Research, University of North Carolina at Chapel Hill. The research of Yin Xia was supported in part by “The Recruitment Program of Global Experts” Youth Project from China, the startup fund from Fudan University and NSF Grant DMS-1612906.

<sup>2</sup>Department of Biostatistics, Harvard School of Public Health, Harvard University. The research of Tianxi Cai was supported in part by NIH Grants R01 GM079330, P50 MH106933, and U54 HG007963.

<sup>3</sup>Department of Statistics, The Wharton School, University of Pennsylvania. The research of Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334.

are investigated through simulation and data analysis. The results show that the proposed tests maintain the desired error rates under the null and have good power under the alternative at moderate sample sizes. The procedures are applied to the Framingham Offspring study to investigate the interactions between smoking and cardiovascular related genetic mutations important for an inflammation marker.

**Keywords:** False discovery proportion, false discovery rate, high-dimensional linear regression, hypothesis testing, multiple comparisons, sparsity, two-sample tests.

## 1 Introduction

As we enter a new era of data science, called by some the “information century”, research in several novel genomics and epigenomics fields are well underway. Large-scale genome-wide scans, such as genome-wide association studies, have become widely available tools for identifying common genetic variants that contribute to complex diseases and treatment responses (McCarthy et al. (2008); Venter et al. (2001)). However, there is growing evidence that genetic variants alone explain only a small proportion of variations in complex disease phenotypes. Most complex diseases are a result of interplay between genes and environment (Hunter (2005)). It is thus of substantial interest to rigorously study the effects of environment and its interaction with genetic predispositions on disease phenotypes.

When the environmental factor is a binary variable such as smoking status or gender, such interaction problems can be addressed through the two-sample high-dimensional regression framework. Specifically, interaction detection can be formulated based on comparing two

high-dimensional regression models

$$\mathbf{Y}_d = \boldsymbol{\mu}_d + \mathbf{X}_d \boldsymbol{\beta}_d + \boldsymbol{\epsilon}_d, \quad \text{for } d = 1, 2, \quad (1)$$

and identifying the nonzero components of  $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$ , where  $\boldsymbol{\beta}_d = (\beta_{1,d}, \dots, \beta_{p,d})^\top \in \mathbb{R}^p$ ,  $\boldsymbol{\mu}_d = (\mu_{1,d}, \dots, \mu_{n_d,d})^\top$ ,  $\mathbf{X}_d = (\mathbf{X}_{1,d}^\top, \dots, \mathbf{X}_{n_d,d}^\top)^\top$ ,  $\mathbf{Y}_d = (Y_{1,d}, \dots, Y_{n_d,d})^\top$ , and  $\boldsymbol{\epsilon}_d = (\epsilon_{1,d}, \dots, \epsilon_{n_d,d})^\top$ , with  $\{\epsilon_{k,d}\}$  being independent and identically distributed (i.i.d) random variables with mean zero and variance  $\sigma_{\epsilon_d}^2$  and independent of  $\mathbf{X}_{k,d}$ ,  $k = 1, \dots, n_d$ . Two-sample interaction detection problems arise in many other biomedical settings. For example, when the two samples represent diseased and non-diseased group and  $Y$  represents a diagnostic test, and the non-zero components of  $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$  represent the covariates that affect the diagnostic accuracy of  $Y$  (Pepe (2003)). When the two samples represent two treatment groups, the proposed testing procedures have important applications in personalized medicine. The non-zero components of  $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$  correspond to markers useful for individualized treatment selection since the rule that optimize the treatment selection for an individual patient with genomic markers  $\mathbf{X}$  can be formed based on  $(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^\top \mathbf{X}$  (Matsouaka et al. (2014)). However, the high dimensionality of the genomic data presents substantial statistical challenges in efficiently identifying gene-environment interactions and markers useful for personalized treatment selection.

There is a paucity of literature focusing on multiple testing of the regression coefficients in the high-dimensional two-sample setting while controlling the false discovery rate (FDR) and false discovery proportion (FDP). For example, Zhang and Zhang (2014), Van de Geer et al. (2014), and Javanmard and Montanari (2013, 2014) considered confidence intervals

and tests for a given coordinate of a high-dimensional linear regression vector. Procedures that are based on the “de-biased” Lasso estimators were proposed. The focus was solely on inference for a given coordinate and simultaneous testing of all coordinates was not considered. Recently, Liu and Luo (2014) investigated the one-sample version of the multiple testing problem, testing simultaneously

$$H'_{0,i} : \beta_{i,1} = 0 \text{ versus } H'_{1,i} : \beta_{i,1} \neq 0, \quad i = 1, \dots, p,$$

with the control of FDR. They constructed the test statistics based on bias-corrected sample covariances of the residuals and inverse regression, as explained in detail in Section 2.2. The one-sample setting is simpler than the two-sample multiple testing problem considered in the present paper. For example, their proposed test statistics have desirable theoretical properties due to the facts that (i) they are asymptotically normally distributed under  $H'_{0,i} : \beta_{i,1} = 0$ , and (ii) the correlation between two test statistics is equal to the partial correlation between two covariates, which is fully determined by the precision matrix. However, those properties no longer hold when we extend the hypothesis testing problem to two samples as described in (3).

In this paper, we are interested in developing efficient procedures for testing  $\beta_1 - \beta_2$ . The first goal is to develop a global test for

$$H_0 : \beta_1 = \beta_2 \text{ versus } H_1 : \beta_1 \neq \beta_2 \tag{2}$$

that is powerful against sparse alternatives. We then develop a procedure for simultaneously testing the hypotheses

$$H_{0,i} : \beta_{i,1} = \beta_{i,2} \text{ versus } H_{1,i} : \beta_{i,1} \neq \beta_{i,2}, \quad i = 1, \dots, p, \tag{3}$$

with FDR and FDP control. The test statistics are constructed using the covariances between the residuals of the fitted regression models and the inverse regression models. Although the techniques build on the inverse regression method developed in Liu and Luo (2014) for the one-sample case, the two-sample case poses significant additional difficulties in both methodology development and technical analyses. We point out here two such major challenges and more detailed discussion is given in Section 2.3.

- (a) The construction of test statistics is much more involved than the one-sample case. This is mainly due to the fact that the difference of regression coefficients can no longer be reduced to the difference of residual covariances as in the one-sample setting. Furthermore, corrections of the test statistics are essential in the two-sample case to establish the asymptotic normality.
- (b) The technical analyses of the two-sample case are much more challenging. This is because the one-sample case can be easily reduced to a weakly correlated testing problem provided that the precision matrix of the covariates is sparse or nearly sparse, while the two-sample case cannot as the correlation structure is much more complicated.

The properties of the proposed testing procedures are investigated theoretically as well as numerically through simulation and data analysis. Theoretical justifications are provided to ensure the validity of the proposed tests and optimality results are established under sparsity assumptions on the regression coefficients. A simulation study is carried out to demonstrate that the proposed tests maintain the desired error rates under the null and have good power under the alternative at moderate sample sizes. The simulation results

also show that the new multiple testing procedure outperforms the well known Benjamini-Yekutieli procedure (Benjamini and Yekutieli (2001)). In addition, the proposed testing procedures are illustrated by an application to the Framingham Offspring Study (Kannel et al., 1979) to study how smoking and its interaction with a genetic predisposition affect an inflammation marker which plays an important role in the risk of developing cardiovascular disease.

The rest of the paper is organized as follows. In Section 2, we introduce the construction of the new test statistics and discuss the technical differences and theoretical challenges of the two-sample testing problems. Section 3 develops a maximum-type statistic  $M_n$  and the corresponding test for the global hypothesis  $H_0 : \beta_1 = \beta_2$  through the inverse regression framework. We establish in this section the asymptotic null distribution of  $M_n$  and show the optimality results under sparse alternatives. Large-scale multiple testing with FDR and FDP control is presented in Section 4. Section 5 investigates the numerical performance of the proposed procedures by simulations. In Section 6, we apply the proposed procedures to the Framingham Offspring Study. The proofs of the main results are given in Section 8.

## 2 Methodology

### 2.1 Notation and Definitions

We first introduce the notation and definitions that will be used throughout the paper. For a vector  $\beta_d = (\beta_{1,d}, \dots, \beta_{p,d})^\top \in \mathbb{R}^p$ , define the  $\ell_q$  norm by  $|\beta_d|_q = (\sum_{i=1}^p |\beta_{i,d}|^q)^{1/q}$  for  $1 \leq q \leq \infty$ . For subscripts, we use the convention that  $i$  stands for the  $i^{\text{th}}$  entry of a vector

and  $(i, j)$  for the entry in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column of a matrix,  $k$  represents the  $k^{\text{th}}$  sample and  $d$  is the group indicator. Let  $\mathbf{X}_d = (\mathbf{X}_{1,d}^\top, \dots, \mathbf{X}_{n_d,d}^\top)^\top$  be the  $n_d \times p$  data matrix, and  $\mathbf{Y}_d = (Y_{1,d}, \dots, Y_{n_d,d})^\top$  be the  $n_d \times 1$  data matrix, for  $d = 1, 2$ . Throughout, suppose that we have i.i.d random samples  $\{Y_{k,d}, \mathbf{X}_{k,d}, 1 \leq k \leq n_d\}$  with  $\mathbf{X}_{k,d} = (X_{k,1,d}, \dots, X_{k,p,d})$  being a random vector with covariance matrix  $\Sigma_d$  for  $d = 1, 2$ . Define  $\Sigma_d^{-1} = \mathbf{\Omega}_d = (\omega_{i,j,d})$ .

For any vector  $\boldsymbol{\mu}_d \in \mathbb{R}^p$ , let  $\boldsymbol{\mu}_{-i,d}$  denote the  $(p-1)$ -dimensional vector formed by removing the  $i^{\text{th}}$  entry from  $\boldsymbol{\mu}_d$ . For a symmetric matrix  $\mathbf{A}_d$ , let  $\lambda_{\max}(\mathbf{A}_d)$  and  $\lambda_{\min}(\mathbf{A}_d)$  denote the largest and smallest eigenvalues of  $\mathbf{A}_d$ , respectively. For any  $n \times p$  matrix  $\mathbf{A}_d$ ,  $\mathbf{A}_{i,-j,d}$  denotes the  $i^{\text{th}}$  row of  $\mathbf{A}_d$  with its  $j^{\text{th}}$  entry removed and  $\mathbf{A}_{-i,j,d}$  denotes the  $j^{\text{th}}$  column of  $\mathbf{A}_d$  with its  $i^{\text{th}}$  entry removed.  $\mathbf{A}_{-i,-j,d}$  denotes the  $(n-1) \times (p-1)$  submatrix of  $\mathbf{A}_d$  with its  $i^{\text{th}}$  row and  $j^{\text{th}}$  column removed. Let  $\mathbf{A}_{-,j,d}$  denote the  $n \times (p-1)$  submatrix of  $\mathbf{A}_d$  with the  $j^{\text{th}}$  column removed,  $\mathbf{A}_{i,\cdot,d}$  denote the  $i^{\text{th}}$  row of  $\mathbf{A}_d$ ,  $\mathbf{A}_{\cdot,j,d}$  denote the  $j^{\text{th}}$  column of  $\mathbf{A}_d$  and  $\bar{\mathbf{A}}_{\cdot,j,d} = 1/n \sum_{i=1}^n \mathbf{A}_{i,j,d}$ . Let  $\bar{\mathbf{A}}_{-,j,d} = 1/n \sum_{i=1}^n \mathbf{A}_{i,-j,d}$ ,  $\bar{\mathbf{A}}_{\cdot,j,d} = (\bar{\mathbf{A}}_{\cdot,j,d}, \dots, \bar{\mathbf{A}}_{\cdot,j,d})_{n \times 1}^\top$ , and  $\bar{\mathbf{A}}_{(\cdot,-j,d)} = (\bar{\mathbf{A}}_{\cdot,-j,d}^\top, \dots, \bar{\mathbf{A}}_{\cdot,-j,d}^\top)_{n \times (p-1)}^\top$ . Let  $\bar{\mathbf{A}}_d = 1/n \sum_{i=1}^n \mathbf{A}_{i,\cdot,d}$ . For a matrix  $\mathbf{\Omega} = (\omega_{i,j})_{p \times p}$ , the matrix 1-norm is the maximum absolute column sum,  $\|\mathbf{\Omega}\|_{L_1} = \max_{1 \leq j \leq p} \sum_{i=1}^p |\omega_{i,j}|$ , the matrix elementwise infinity norm is defined to be  $\|\mathbf{\Omega}\|_\infty = \max_{1 \leq i,j \leq p} |\omega_{i,j}|$  and the elementwise  $\ell_1$  norm is  $\|\mathbf{\Omega}\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\omega_{i,j}|$ . For a set  $\mathcal{H}$ , let  $|\mathcal{H}|$  be the cardinality of  $\mathcal{H}$ . For two sequences of real numbers  $\{a_n\}$  and  $\{b_n\}$ , write  $a_n = O(b_n)$  if there exists a constant  $C$  such that  $|a_n| \leq C|b_n|$  holds for all  $n$ , write  $a_n = o(b_n)$  if  $\lim_{n \rightarrow \infty} a_n/b_n = 0$ , and write  $a_n \asymp b_n$  if there are positive constants  $c$  and  $C$  such that  $c \leq a_n/b_n \leq C$  for all  $n$ .

## 2.2 Test Statistics

To form the test statistics, we consider the *inverse* regression models obtained by regressing  $X_{k,i,d}$  on  $(Y_{k,d}, \mathbf{X}_{k,-i,d})$ , as introduced in Liu and Luo (2014)

$$X_{k,i,1} = \alpha_{i,1} + (Y_{k,1}, \mathbf{X}_{k,-i,1})\boldsymbol{\gamma}_{i,1} + \eta_{k,i,1}, \quad (k = 1, \dots, n_1)$$

$$X_{k,i,2} = \alpha_{i,2} + (Y_{k,2}, \mathbf{X}_{k,-i,2})\boldsymbol{\gamma}_{i,2} + \eta_{k,i,2}, \quad (k = 1, \dots, n_2)$$

where for  $d = 1, 2$ ,  $\eta_{k,i,d}$  has mean zero and variance  $\sigma_{\eta_{i,d}}^2$  and is uncorrelated with  $(Y_{k,d}, \mathbf{X}_{k,-i,d})$ , and  $\boldsymbol{\gamma}_{i,d} = (\gamma_{i,1,d}, \dots, \gamma_{i,p,d})^\top$  satisfies

$$\boldsymbol{\gamma}_{i,d} = -\sigma_{\eta_{i,d}}^2 (-\beta_{i,d}/\sigma_{\epsilon_d}^2, \beta_{i,d}\boldsymbol{\beta}_{-i,d}^\top/\sigma_{\epsilon_d}^2 + \boldsymbol{\Omega}_{i,-i,d})^\top, \quad (4)$$

where  $\sigma_{\eta_{i,d}}^2 = (\beta_{i,d}^2/\sigma_{\epsilon_d}^2 + \omega_{i,i,d})^{-1}$ , as provided in Liu and Luo (2014).

**Remark 1** Equation (4) can be obtained directly as follows. Denote the covariance matrix of  $\mathbf{Z} = (X_{k,i,d}, Y_{k,d}, \mathbf{X}_{k,-i,d})$  by  $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{Z})$ . Section 2.5 of Anderson (2003) shows that  $\boldsymbol{\gamma}_{i,d}$  can be obtained by  $\boldsymbol{\gamma}_{i,d} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ , where  $\boldsymbol{\Sigma}_{22} = \text{Cov}(\mathbf{Z}_1)$  with  $\mathbf{Z}_1 = (Y_{k,d}, \mathbf{X}_{k,-i,d})$  and  $\boldsymbol{\Sigma}_{21} = \text{Cov}(\mathbf{Z}_1, X_{k,i,d})$  is the covariance between  $\mathbf{Z}_1$  and  $X_{k,i,d}$ . Then (4) follows from the regression model  $\mathbf{Y}_d = \boldsymbol{\mu}_d + \mathbf{X}_d\boldsymbol{\beta}_d + \boldsymbol{\epsilon}_d$  and the fact that  $\mathbf{X}_d$  and  $\boldsymbol{\epsilon}_d$  are uncorrelated with each other.

Because  $r_{i,d} = \text{Cov}(\epsilon_{k,d}, \eta_{k,i,d})$  can be expressed as  $-\gamma_{i,1,d}\text{Cov}(\epsilon_{k,d}, Y_{k,d}) = -\gamma_{i,1,d}\sigma_{\epsilon_d}^2 = -\sigma_{\eta_{i,d}}^2\beta_{i,d}$ , the null hypotheses in global testing problem (2) and entry-wise testing problem (3) would be, respectively, equivalent to

$$H_0 : \max_{1 \leq i \leq p} |r_{i,1}/\sigma_{\eta_{i,1}}^2 - r_{i,2}/\sigma_{\eta_{i,2}}^2| = 0, \quad (5)$$

and

$$H_{0,i} : r_{i,1}/\sigma_{\eta_{i,1}}^2 = r_{i,2}/\sigma_{\eta_{i,2}}^2, \quad i = 1, \dots, p, \quad (6)$$

and we base the tests on the estimates of  $\{r_{i,d}/\sigma_{\eta_{i,d}}^2, i = 1, \dots, p; d = 1, 2\}$ .

Define the residuals

$$\begin{aligned} \hat{\epsilon}_{k,d} &= Y_{k,d} - \bar{Y}_d - (\mathbf{X}_{k,\cdot,d} - \bar{\mathbf{X}}_d)\hat{\boldsymbol{\beta}}_d \\ \hat{\eta}_{k,i,d} &= X_{k,i,d} - \bar{X}_{i,d} - (Y_{k,d} - \bar{Y}_d, (\mathbf{X}_{k,\cdot,-i,d} - \bar{\mathbf{X}}_{\cdot,-i,d}))\hat{\boldsymbol{\gamma}}_{i,d}, \end{aligned}$$

where  $\hat{\boldsymbol{\beta}}_d = (\hat{\beta}_{1,d}, \dots, \hat{\beta}_{p,d})$  and  $\hat{\boldsymbol{\gamma}}_{i,d} = (\hat{\gamma}_{i,1,d}, \dots, \hat{\gamma}_{i,p,d})$  are the respective estimators of  $\boldsymbol{\beta}_d$  and  $\boldsymbol{\gamma}_{i,d}$  satisfy

$$\begin{aligned} \max\{|\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d|_1, \max_{1 \leq i \leq p} |\hat{\boldsymbol{\gamma}}_{i,d} - \boldsymbol{\gamma}_{i,d}|_1\} &= O_{\mathbb{P}}(a_{n1}), \\ \max\{|\hat{\boldsymbol{\beta}}_d - \boldsymbol{\beta}_d|_2, \max_{1 \leq i \leq p} |\hat{\boldsymbol{\gamma}}_{i,d} - \boldsymbol{\gamma}_{i,d}|_2\} &= O_{\mathbb{P}}(a_{n2}), \end{aligned} \quad (7)$$

for some  $a_{n1}$  and  $a_{n2}$  such that

$$\max\{a_{n1}a_{n2}, a_{n2}^2\} = o\{(n \log p)^{-1/2}\}, \text{ and } a_{n1} = o(1/\log p). \quad (8)$$

Estimators  $\hat{\boldsymbol{\beta}}_d$  and  $\hat{\boldsymbol{\gamma}}_{i,d}$  that satisfy (7) and (8) can be obtained easily via standard methods such as the lasso and Danzig selector, see, for example, Xia et al. (2015) and Liu and Luo (2014).

Based on the residuals  $\hat{\epsilon}_{k,d}$  and  $\hat{\eta}_{k,i,d}$ , a natural estimator of  $r_{i,d}$  is the sample covariance between the residuals,

$$\tilde{r}_{i,d} = n_d^{-1} \sum_{k=1}^{n_d} \hat{\epsilon}_{k,d} \hat{\eta}_{k,i,d}.$$

Because  $\tilde{r}_{i,d}$  tends to be biased, we define a bias corrected estimator for  $r_{i,d}$  as

$$\hat{r}_{i,d} = \tilde{r}_{i,d} + \hat{\sigma}_{\epsilon_d}^2 \hat{\gamma}_{i,1,d} + \hat{\sigma}_{\eta_{i,d}}^2 \hat{\beta}_{i,d}, \quad (9)$$

where  $\hat{\sigma}_{\epsilon_d}^2 = n_d^{-1} \sum_{k=1}^{n_d} \hat{\epsilon}_{k,d}^2$  and  $\hat{\sigma}_{\eta_{i,d}}^2 = n_d^{-1} \sum_{k=1}^{n_d} \hat{\eta}_{k,i,d}^2$  are the sample variances satisfying

$$\max\{|\hat{\sigma}_{\epsilon_d}^2 - \sigma_{\epsilon_d}^2|, \max_{1 \leq i \leq p} |\hat{\sigma}_{\eta_{i,d}}^2 - \sigma_{\eta_{i,d}}^2|\} = O_P\{(\log p/n_d)^{1/2}\},$$

which can be obtained by Lemma 2 in Xia et al. (2015) under conditions (7) and (8). By Lemma 2, the bias of  $\hat{r}_{i,d}$  is then of order  $\max\{\beta_{i,d}(\log p/n_d)^{1/2}, (n_d \log p)^{-1/2}\}$ .

**Remark 2** The most straightforward way to estimate  $r_{i,d}$  is to use the sample covariance between the error terms,  $n_d^{-1} \sum_{k=1}^{n_d} \epsilon_{k,d} \eta_{k,i,d}$ . However, the error terms are unknown, and we can use the the sample covariance between the residuals  $\tilde{r}_{i,d}$  instead. The bias of  $\tilde{r}_{i,d}$  exceeds the desired rate  $(n_d \log p)^{-1/2}$ , and thus we calculate the difference of  $\tilde{r}_{i,d}$  and  $n_d^{-1} \sum_{k=1}^{n_d} \epsilon_{k,d} \eta_{k,i,d}$ , which up to order  $(n_d \log p)^{-1/2}$ , is equal to  $\hat{\sigma}_{\epsilon_d}^2 \hat{\gamma}_{i,1,d} + \hat{\sigma}_{\eta_{i,d}}^2 \hat{\beta}_{i,d}$ . Hence, we define  $\hat{r}_{i,d} = \tilde{r}_{i,d} + \hat{\sigma}_{\epsilon_d}^2 \hat{\gamma}_{i,1,d} + \hat{\sigma}_{\eta_{i,d}}^2 \hat{\beta}_{i,d}$  as in (9).

For  $i = 1, \dots, p$  and  $d = 1, 2$ , a natural estimator of  $r_{i,d}/\sigma_{\eta_{i,d}}^2$  can then be defined by

$$T_{i,d} = \hat{r}_{i,d}/\hat{\sigma}_{\eta_{i,d}}^2. \tag{10}$$

Subsequently, we may test the hypotheses (2) and (3) using the estimators  $\mathcal{T} = \{T_{i,1} - T_{i,2} : i = 1, \dots, p\}$ . However, since  $T_{i,1} - T_{i,2}$  in  $\mathcal{T}$  are heteroscedastic with possibly a wide range of variability, we instead consider a standardized version of  $T_{i,1} - T_{i,2}$ . Specifically, let

$$U_{i,d} = n_d^{-1} \sum_{k=1}^{n_d} \{\epsilon_{k,d} \eta_{k,i,d} - \mathbb{E}(\epsilon_{k,d} \eta_{k,i,d})\} \text{ and } \tilde{U}_{i,d} = (\beta_{i,d} + U_{i,d})/\sigma_{\eta_{i,d}}^2.$$

It can be shown in Lemma 2 that, uniformly in  $i = 1, \dots, p$ ,

$$|T_{i,d} - \tilde{U}_{i,d}| = O_P\{\beta_{i,d}(\log p/n_d)^{1/2}\} + o_P\{(n_d \log p)^{-1/2}\}.$$

Noting that  $\theta_{i,d} = \text{Var}(\tilde{U}_{i,d}) = \text{Var}(\epsilon_{k,d}\eta_{k,i,d}/\sigma_{\eta_{i,d}}^2)/n_d = (\sigma_{\epsilon_d}^2/\sigma_{\eta_{i,d}}^2 + \beta_{i,d}^2)/n_d$ , we estimate  $\theta_{i,d}$  by

$$\hat{\theta}_{i,d} = (\hat{\sigma}_{\epsilon_d}^2/\hat{\sigma}_{\eta_{i,d}}^2 + \hat{\beta}_{i,d}^2)/n_d.$$

and define the standardized statistics

$$W_i = \frac{T_{i,1} - T_{i,2}}{(\hat{\theta}_{i,1} + \hat{\theta}_{i,2})^{1/2}}, \quad i = 1, \dots, p. \quad (11)$$

We base the tests for (2) and (3) on  $\{W_i, i = 1, \dots, p\}$ , which will be studied in detail in Sections 3 and 4.

### 2.3 Discussion

We discuss here the substantial differences between the two-sample and one-sample cases and the necessity for significant adjustments and corrections in the two-sample setting.

The proposed tests are based on estimators of  $r_{i,1}/\sigma_{\eta_{i,1}}^2 - r_{i,2}/\sigma_{\eta_{i,2}}^2$ . Here we estimate  $r_{i,d} = \text{Cov}(\epsilon_{k,d}, \eta_{k,i,d})$  through constructing a bias-corrected sample covariance between the residuals,  $\hat{r}_{i,d}$ , as defined in (9). That is, we need to get an estimate of the difference between the naive estimate  $\tilde{r}_{i,d}$  and an unbiased estimate of  $r_{i,d}$ , which is  $n_d^{-1} \sum_{k=1}^{n_d} \epsilon_{k,d}\eta_{k,i,d}$ .

Liu and Luo (2014) considered the one-sample case of the multiple testing problem (3) so  $r_i/\sigma_{\eta_i}^2 = 0$  is equivalent to  $r_i = 0$  under the null hypothesis, and  $r_i$  is easier to estimate. The procedure in Liu and Luo (2014) is based on the estimation of  $r_i$  instead of  $r_i/\sigma_{\eta_i}^2$ . In the two-sample case,  $r_{i,1}/\sigma_{\eta_{i,1}}^2 = r_{i,2}/\sigma_{\eta_{i,2}}^2$  is not equivalent to  $r_{i,1} = r_{i,2}$ . Thus, it is necessary to construct testing procedures based directly on estimators of  $r_{i,1}/\sigma_{\eta_{i,1}}^2 - r_{i,2}/\sigma_{\eta_{i,2}}^2$ .

Furthermore, in the one-sample case, the asymptotic normality of  $T_i$  can be established

because  $\beta_{i,1} = 0$  under the null, which is shown in Lemma 2. Thus the theoretical properties of the individual test statistics are easier to obtain. In the two-sample case,  $\beta_{i,1}$  and  $\beta_{i,2}$  are not necessary equal to 0 under the null, and corrections are thus essential in order to show  $W_i$  is close to a normal random variable; the technical details are much more complicated.

More importantly, in the one-sample case, under the null hypothesis  $\beta_{i,1} = 0$ , and thus  $\text{Corr}(\epsilon_k \eta_{k,i}, \epsilon_k \eta_{k,j}) = \omega_{i,j} / (\omega_{i,i} \omega_{j,j})$ , which is fully determined by the precision matrix of the covariates and thus simplifies the calculations. In the two-sample version,  $\beta_{i,1} = \beta_{i,2}$  under the null hypothesis and they are not necessary equal to zero. The calculation of  $\text{Corr}(\epsilon_{k,d} \eta_{k,i,d}, \epsilon_{k,d} \eta_{k,j,d})$ , which determines the correlation between  $W_i$  and  $W_j$ , is much more involved, and it can be shown in the proof of Theorem 4 that

$$\tilde{\xi}_{i,j,d} = \text{Corr}(\epsilon_{k,d} \eta_{k,i,d}, \epsilon_{k,d} \eta_{k,j,d}) = \frac{(\omega_{i,j,d} \sigma_{\epsilon_d}^2 + 2\beta_{i,d} \beta_{j,d})}{\{(\omega_{i,i,d} \sigma_{\epsilon_d}^2 + 2\beta_{i,d}^2)(\omega_{j,j,d} \sigma_{\epsilon_d}^2 + 2\beta_{j,d}^2)\}^{1/2}}. \quad (12)$$

The technical analysis for establishing the theoretical results in Sections 3 and 4 is thus much more challenging.

### 3 Global Test

In this section, we wish to test the global hypothesis

$$H_0 : \beta_1 = \beta_2 \text{ versus } H_1 : \beta_1 \neq \beta_2.$$

We propose a procedure based on the standardized statistics  $\{W_i, i = 1, \dots, p\}$

$$M_n = \max_{1 \leq i \leq p} W_i^2 = \max_{1 \leq i \leq p} \frac{(T_{i,1} - T_{i,2})^2}{\hat{\theta}_{i,1} + \hat{\theta}_{i,2}}. \quad (13)$$

It is shown in Section 3.1 that, under certain regularity conditions,  $M_n - 2 \log p + \log \log p$  converges to a Gumbel distribution under the null, and the asymptotic  $\alpha$ -level test can thus be defined as

$$\Psi_\alpha = I(M_n \geq q_\alpha + 2 \log p - \log \log p), \quad (14)$$

where  $q_\alpha$  is the  $1 - \alpha$  quantile of the Gumbel distribution with the cumulative distribution function  $\exp(-\pi^{-1/2}e^{-t/2})$ ,

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

We reject the null hypothesis  $H_0$  whenever  $\Psi_\alpha = 1$ .

### 3.1 Asymptotic Null Distribution

We first introduce some regularity conditions, under which,  $M_n - 2 \log p + \log \log p$  converges weakly to a Gumbel random variable with distribution function  $\exp(-\pi^{-1/2}e^{-t/2})$ .

(C1)  $\log p = o(n^{1/5})$ ,  $n_1 \asymp n_2$ , and for some constants  $C_0, C_1, C_2 > 0$ ,  $C_0^{-1} \leq \lambda_{\min}(\mathbf{\Omega}_d) \leq \lambda_{\max}(\mathbf{\Omega}_d) \leq C_0$ ,  $C_1^{-1} \leq \sigma_{\epsilon_d}^2 \leq C_1$ , and  $|\boldsymbol{\beta}_d|_\infty \leq C_2$  for  $d = 1, 2$ . There exists some  $\tau > 0$  such that  $|A_\tau| = O(p^r)$  with  $r < 1/4$ , where  $A_\tau = \{i : |\beta_{i,d}| \geq (\log p)^{-2-\tau}, 1 \leq i \leq p, \text{ for } d = 1 \text{ or } 2\}$ .

(C2) Let  $\mathbf{D}_d$  be the diagonal of  $\mathbf{\Omega}_d$  and let  $(\xi_{i,j,d}) = \mathbf{R}_d = \mathbf{D}_d^{-1/2} \mathbf{\Omega}_d \mathbf{D}_d^{-1/2}$ , for  $d = 1, 2$ .  $\max_{1 \leq i < j \leq p} |\xi_{i,j,d}| \leq \xi_d < 1$  for some constant  $0 < \xi_d < 1$ .

(C3) There exists some constant  $K > 0$  such that  $\max_{\text{Var}(\mathbf{a}^\top \mathbf{X}_{k,\cdot,d}^\top)=1} \mathbf{E} \exp(K(\mathbf{a}^\top \mathbf{X}_{k,\cdot,d}^\top)^2)$  and  $\mathbf{E} \exp(K\epsilon_{k,d}^2)$  are finite.

Condition (C1) on the eigenvalues is commonly used in the high-dimensional setting and implies that most of the variables are not highly correlated with each other. Condition (C2) is also mild. For example, if  $\max_{1 \leq i \leq j \leq p} |\xi_{i,j,d}| = 1$ , then  $\mathbf{\Omega}_d$  is singular. (C3) is a sub-Gaussian tail condition, and it can be weakened to a polynomial tail condition if  $p < n^c$  for some constant  $c > 0$ .

**Theorem 1** *Suppose (C1), (C2), (C3), (7), and (8) hold. Then under  $H_0$ , for any  $t \in \mathbb{R}$ ,*

$$P(M_n - 2 \log p + \log \log p \leq t) \rightarrow \exp\{-\pi^{-1/2} \exp(-t/2)\}, \text{ as } n_1, n_2, p \rightarrow \infty, \quad (15)$$

where  $M_n$  is defined in (13). Under  $H_0$ , the convergence in (15) is uniform for all  $\{Y_{k,d}, \mathbf{X}_{k,d} : k = 1, 2, \dots, n_d\}$  satisfying (C1), (C2), (C3), (7), and (8).

**Remark 3** The analysis can be extended to test  $H_0 : \boldsymbol{\beta}_{G,1} = \boldsymbol{\beta}_{G,2}$  versus  $H_1 : \boldsymbol{\beta}_{G,1} \neq \boldsymbol{\beta}_{G,2}$  for a given index set  $G$ . We can construct the test statistic as  $M_{G,n} = \max_{i \in G} W_i^2$ , and obtain a similar Gumbel limiting null distribution by replacing  $p$  with  $|G|$ , as  $n_1, n_2, |G| \rightarrow \infty$ . The condition (C1) will be slightly different, with  $A_\tau$  being replaced by  $A_{G,\tau} = \{i : |\beta_{i,d}| \geq (\log p)^{-2-\tau}, i \in G, \text{ for } d = 1 \text{ or } 2\}$ .

**Remark 4** Condition (C1) is slightly stronger than the conditions in Liu and Luo (2014) as we need  $|A_\tau| = O(p^r)$  with  $r < 1/4$ . This is due to the major difference between the one-sample and two-sample cases that the global null  $H_0 : \boldsymbol{\beta} = 0$  is a simple null in the one-sample case and the null  $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$  is composite in the two-sample case. In the one-sample case,  $T_i$  is a nearly unbiased estimate of  $\beta_i$  because  $\beta_i = 0$  under the global null. However, in the two-sample case, as stated in Lemma 2, additional correction terms

involving  $\beta_{i,d}$  are needed in order to make  $T_{i,d}$  nearly unbiased because  $\beta_{i,1}$  and  $\beta_{i,2}$  are not necessary equal to 0 under the null. Thus, slightly stronger conditions on  $A_\tau$  are needed.

### 3.2 Asymptotic Power

We now analyze the asymptotic power of the test  $\Psi_\alpha$  given in (14). The test is shown to be particularly powerful against a large class of sparse alternatives and the power is minimax rate optimal. We first define a class of regression coefficients:

$$\mathcal{U}(c) = \left\{ (\beta_1, \beta_2) : \max_{1 \leq i \leq p} \frac{|\beta_{i,1} - \beta_{i,2}|}{(\theta_{i,1} + \theta_{i,2})^{1/2}} \geq c(\log p)^{1/2} \right\}. \quad (16)$$

We show that the null hypothesis  $H_0$  can be rejected by the test  $\Psi_\alpha$  with overwhelming probability, if  $(\beta_1, \beta_2) \in \mathcal{U}(2\sqrt{2})$ .

**Theorem 2** *Let the test  $\Psi_\alpha$  be given in (14). Suppose (C1), (C3), (7) and (8) hold. Then*

$$\inf_{(\beta_1, \beta_2) \in \mathcal{U}(2\sqrt{2})} P(\Psi_\alpha = 1) \rightarrow 1, \quad n, p \rightarrow \infty.$$

Theorem 2 shows that the null parameter set in which  $\beta_1 = \beta_2$  is asymptotically distinguishable from  $\mathcal{U}(2\sqrt{2})$  by the test  $\Psi_\alpha$ .

We further show that the lower bound in (16) is rate optimal. Let  $\mathcal{T}_\alpha$  be the set of all  $\alpha$ -level tests,  $P(T_\alpha = 1) \leq \alpha$  under  $H_0$  for all  $T_\alpha \in \mathcal{T}_\alpha$ . If  $c$  in (16) is sufficiently small, then any  $\alpha$  level test is unable to reject the null hypothesis correctly uniformly over  $(\beta_1, \beta_2) \in \mathcal{U}(c)$  with probability tending to one.

**Theorem 3** *Suppose that  $\log p = o(n)$ . Let  $\alpha, \beta > 0$  and  $\alpha + \beta < 1$ . Then there exists a constant  $c_0 > 0$  such that for all sufficiently large  $n$  and  $p$ ,*

$$\inf_{(\beta_1, \beta_2) \in \mathcal{U}(c_0)} \sup_{T_\alpha \in \mathcal{T}_\alpha} P(T_\alpha = 1) \leq 1 - \beta.$$

Theorem 3 shows that the order  $(\log p)^{1/2}$  in the lower bound of  $\max_{1 \leq i \leq p} \{|\beta_{i,1} - \beta_{i,2}| / (\theta_{i,1} + \theta_{i,2})^{1/2}\}$  in (16) cannot be further improved.

## 4 Multiple Testing with False Discovery Rate Control

### 4.1 Multiple Testing Procedure

If the global null hypothesis is rejected, it is then of interest to identify the subset of variables in  $\mathbf{X}$  that interact with the group indicator. This can be achieved by simultaneously testing on the entries of  $\beta_1 - \beta_2$  with FDR and FDP control,

$$H_{0,i} : \beta_{i,1} = \beta_{i,2} \text{ versus } H_{1,i} : \beta_{i,1} \neq \beta_{i,2}, \quad 1 \leq i \leq p. \quad (17)$$

The standardized differences of  $T_{i,1} - T_{i,2}$  are defined by the test statistics  $W_i = (T_{i,1} - T_{i,2}) / (\hat{\theta}_{i,1} + \hat{\theta}_{i,2})^{1/2}$  as in (11). Let  $t$  be the threshold such that  $H_{0,i}$  is rejected if  $|W_i| \geq t$ . Let  $\mathcal{H}_0 = \{i : \beta_{i,1} = \beta_{i,2}, 1 \leq i \leq p\}$  be the set of true nulls. Let  $R_0(t) = \sum_{i \in \mathcal{H}_0} I(|W_i| \geq t)$  and  $R(t) = \sum_{1 \leq i \leq p} I(|W_i| \geq t)$ , respectively, denote the total number of false positives and the total number of rejections. The FDP and FDR are defined as

$$\text{FDP}(t) = \frac{R_0(t)}{R(t) \vee 1}, \quad \text{FDR}(t) = \mathbf{E}\{\text{FDP}(t)\}.$$

Ideally, we select the threshold level as

$$t_0 = \inf \{0 \leq t \leq (2 \log p)^{1/2} : \text{FDP}(t) \leq \alpha\}.$$

However,  $\mathcal{H}_0$  is unknown, and we estimate  $\sum_{i \in \mathcal{H}_0} I\{|W_i| \geq t\}$  by  $2p\{1 - \Phi(t)\}$  due to the sparsity of  $\beta_1 - \beta_2$ , where  $\Phi(t)$  is the standard normal cumulative distribution function.

This leads to the following multiple testing procedure.

1. Calculate the test statistics  $W_i = (T_{i,1} - T_{i,2})/(\hat{\theta}_{i,1} + \hat{\theta}_{i,2})^{1/2}$  as in (11).
2. For a given  $0 \leq \alpha \leq 1$ , calculate

$$\hat{t} = \inf \left\{ 0 \leq t \leq (2 \log p)^{1/2} : \frac{2p\{1 - \Phi(t)\}}{R(t) \vee 1} \leq \alpha \right\}.$$

If  $\hat{t}$  does not exist, set  $\hat{t} = (2 \log p)^{1/2}$ .

3. For  $1 \leq i \leq p$ , reject  $H_{0,i}$  if and only if  $|W_i| \geq \hat{t}$ .

## 4.2 Theoretical Properties

We now investigate the theoretical properties of this multiple testing procedure. For any  $1 \leq i \leq p$ , define

$$\Gamma_i(\gamma) = \{j : 1 \leq j \leq p, |\xi_{i,j,d}| \geq (\log p)^{-2-\gamma}, d = 1, 2\},$$

where  $\xi_{i,j,d}$  is defined in Condition (C2). Under regularity conditions, this procedure controls the FDP and FDR at the pre-specified level  $\alpha$ , asymptotically.

**Theorem 4** *Let*

$$\mathcal{S}_\rho = \left\{ i : 1 \leq i \leq p, \frac{|\beta_{i,1} - \beta_{i,2}|}{(\theta_{i,1} + \theta_{i,2})^{1/2}} \geq (\log p)^{1/2+\rho} \right\}.$$

Suppose for some  $\rho > 0$  and some  $\delta > 0$ ,  $|\mathcal{S}_\rho| \geq [1/(\pi^{1/2}\alpha) + \delta](\log p)^{1/2}$ . Suppose that  $|A_\tau \cap \mathcal{H}_0| = o(p^\nu)$  for any  $\nu > 0$ , where  $A_\tau$  is given in Condition (C1). Assume that  $p_0 = |\mathcal{H}_0| \geq cp$  for some  $c > 0$ , and (7) and (8) hold. If there exists some  $\gamma > 0$  such that  $\max_{1 \leq i \leq p} |\Gamma_i(\gamma)| = o(p^\nu)$  for any  $\nu > 0$ , then under (C1) - (C3) with  $p \leq cn^r$  for some  $c > 0$  and  $r > 0$ , we have

$$\lim_{(n,p) \rightarrow \infty} \frac{\text{FDR}(\hat{t})}{\alpha p_0/p} = 1,$$

$$\frac{\text{FDP}(\hat{t})}{\alpha p_0/p} \rightarrow 1$$

in probability, as  $(n, p) \rightarrow \infty$ .

The condition on  $|\mathcal{S}_\rho|$  is mild, because among  $p$  hypotheses in total, it only requires a few number of entries with the standardized difference exceeding  $(\log p)^{1/2+\rho}/n^{1/2}$  for some constant  $\rho > 0$ . The technical condition  $|A_\tau \cap \mathcal{H}_0| = o(p^\nu)$  for any  $\nu > 0$  is to ensure that most of the regression residuals are not highly correlated with each other under the null hypotheses  $H_{0,i} : \beta_{i,1} = \beta_{i,2}$ .

## 5 Simulation Study

We consider the numerical performance, including the sizes and powers of both the global and the multiple testing procedures, through simulation studies. We investigated the performance of both procedures under two sets of simulations. For the first, we generated the data by considering two constructions of regression coefficients under three matrix models, with covariates being a combination of continuous and discrete random variables. For the second set, we studied the numerical performance of the proposed multiple testing procedure

in a setting that is similar to the data application described in Section 6. We compared the proposed multiple testing procedure with Benjamini-Yekutieli (B-Y) procedure, as considered in Benjamini and Yekutieli (2001), and show that the B-Y procedure is much more conservative and has lower power in all cases.

## 5.1 Implementation Details

The proposed testing procedures required the estimation of the regression coefficients  $\beta_d$  and  $\gamma_{i,d}$ , for  $i = 1, \dots, p$  and  $d = 1, 2$ . One may use the Lasso to estimate these parameters, as follows.

$$\beta_d = \mathbf{D}_X^{-1/2} \arg \min_{\mathbf{u}} \left\{ \frac{1}{2n_d} \left| (\mathbf{X}_d - \bar{\mathbf{X}}_d) \mathbf{D}_X^{-1/2} \mathbf{u} - (\mathbf{Y}_d - \bar{\mathbf{Y}}_d) \right|_2^2 + \lambda_n |\mathbf{u}|_1 \right\}, \quad (18)$$

and

$$\gamma_{i,d} = \mathbf{D}_{i,d}^{-1/2} \arg \min_{\mathbf{v}} \left\{ \frac{1}{2n_d} \left| ((\mathbf{Y}_d, \mathbf{X}_{\cdot, -i, d}) - (\bar{\mathbf{Y}}_d, \bar{\mathbf{X}}_{\cdot, -i, d})) \mathbf{D}_{i,d}^{-1/2} \mathbf{v} - (\mathbf{X}_{\cdot, i, d} - \bar{\mathbf{X}}_{\cdot, i, d}) \right|_2^2 + \lambda_{i,n} |\mathbf{v}|_1 \right\}, \quad (19)$$

where  $\mathbf{D}_X = \text{diag}(\hat{\Sigma})$ ,  $\mathbf{D}_{i,d} = \text{diag}(\hat{\sigma}_{\mathbf{Y}_d}, \hat{\Sigma}_{-i, -i})$ ,  $\lambda_n = \kappa \sqrt{\hat{\sigma}_{\mathbf{Y}_d} \log p / n_d}$  and  $\lambda_{i,n} = \kappa \sqrt{\hat{\sigma}_{i,i} \log p / n_d}$ ,

in which  $\hat{\sigma}_{\mathbf{Y}_d}$  is the sample variance of  $\mathbf{Y}_d$  and  $\hat{\Sigma} = (\hat{\sigma}_{i,j})$  is the sample covariance matrix of

$\mathbf{X}_d$ . In the global testing of  $H_0 : \beta_1 = \beta_2$ , we chose the tuning parameter  $\kappa = 2$ .

For multiple testing of  $H_{0,i} : \beta_{i,1} = \beta_{i,2}$ , we selected the tuning parameters  $\lambda_n$  and  $\lambda_{i,n}$  in (18) and (19) adaptively by the data with the principle of making  $\sum_{i \in \mathcal{H}_0} I\{|W_i| \geq t\}$  and  $2\{1 - \Phi(t)\}|\mathcal{H}_0|$  as close as possible. That is, a good choice of the tuning parameters should minimize the error

$$\int_c^1 \left( \frac{\sum_i I(|W_i^{(b)}| \geq \Phi^{-1}(1 - \alpha/2))}{\alpha p} - 1 \right)^2 d\alpha,$$

where  $c > 0$  and  $W_i^{(b)}$  is the statistic of the corresponding tuning parameter. Step 2 below is a discretization of the above integral. The algorithm is summarized as follows.

1. Let  $\lambda_n = b/20\sqrt{\hat{\sigma}_{\mathbf{Y}_d} \log p/n_d}$  and  $\lambda_{i,n} = b/20\sqrt{\hat{\sigma}_{i,i} \log p/n_d}$  for  $b = 1, \dots, 40$ . For each  $b$ , calculate  $\hat{\beta}_d^{(b)}$  and  $\hat{\gamma}_{i,d}^{(b)}$ ,  $i = 1, \dots, p$ ,  $d = 1, 2$ . Based on the estimation of regression coefficients, construct the corresponding statistics  $W_i^{(b)}$  for each  $b$ .
2. Choose  $\hat{b}$  as the minimizer of

$$\hat{b} = \arg \min \sum_{s=1}^{10} \left( \frac{\sum_{1 \leq i \leq p} I\{|W_i^{(b)}| \geq \Phi^{-1}(1 - s[1 - \Phi\{(\log p)^{1/2}\}]/10)\}}{2ps[1 - \Phi\{(\log p)^{1/2}\}]/10} - 1 \right)^2.$$

The tuning parameters  $\lambda_n$  and  $\lambda_{i,n}$  are then chosen to be

$$\lambda_n = \hat{b}/20\sqrt{\hat{\sigma}_{\mathbf{Y}_d} \log p/n_d} \text{ and } \lambda_{i,n} = \hat{b}/20\sqrt{\hat{\sigma}_{i,i} \log p/n_d} \quad (20)$$

## 5.2 Simulation Under Different Matrix Models

We first generated the design matrices  $\mathbf{X}_{k,\cdot,d}$ , for  $k = 1, \dots, n_d$  and  $d = 1, 2$ , with some of the covariates being continuous and the others being discrete. For simplicity, we generated  $\mathbf{X}_{k,\cdot,d}$  from the same distribution for  $d = 1, 2$ . As a first step, for three different matrix models, we obtained i.i.d samples  $\mathbf{X}_{k,\cdot,d} \sim N(0, \Sigma^{(m)})$ , for  $k = 1, \dots, n_d$ , with  $m = 1, 2$  and

3. We then replaced  $l$  covariates of  $\mathbf{X}_{k,\cdot,d}$  by one of three discrete values 0, 1 or 2, with probability  $1/3$  each, where  $l$  is a random integer between  $\lfloor p/2 \rfloor$  and  $p$ . We first introduce the matrix models  $\Sigma^{(m)}$  used in the simulations. Let  $\mathbf{D} = (D_{i,j})$  be a diagonal matrix with  $D_{i,i} = \text{Unif}(1, 3)$  for  $i = 1, \dots, p$ . The following models were used to generate the design matrices.

Model 1:  $\mathbf{\Omega}^{*(1)} = (\omega_{i,j}^{*(1)})$ , where  $\omega_{i,i}^{*(1)} = 1$ ,  $\omega_{i,i+1}^{*(1)} = \omega_{i+1,i}^{*(1)} = 0.6$ ,  $\omega_{i,i+2}^{*(1)} = \omega_{i+2,i}^{*(1)} = 0.3$  and  $\omega_{i,j}^{*(1)} = 0$  otherwise.  $\mathbf{\Omega}^{(1)} = \mathbf{D}^{1/2} \mathbf{\Omega}^{*(1)} \mathbf{D}^{1/2}$ .

Model 2:  $\mathbf{\Omega}^{*(2)} = (\omega_{i,j}^{*(2)})$ , where  $\omega_{i,j}^{*(2)} = \omega_{j,i}^{*(2)} = 0.5$  for  $i = 10(k-1) + 1$  and  $10(k-1) + 2 \leq j \leq 10(k-1) + 10$ ,  $1 \leq k \leq p/10$ .  $\omega_{i,j}^{*(2)} = 0$  otherwise.  $\mathbf{\Omega}^{(2)} = \mathbf{D}^{1/2} (\mathbf{\Omega}^{*(2)} + \delta \mathbf{I}) / (1 + \delta) \mathbf{D}^{1/2}$  with  $\delta = |\lambda_{\min}(\mathbf{\Omega}^{*(2)})| + 0.05$ .

Model 3:  $\mathbf{\Omega}^{*(3)} = (\omega_{i,j}^{*(3)})$ , where  $\omega_{i,i}^{*(3)} = 1$ ,  $\omega_{i,j}^{*(3)} = 0.8 \times \text{Bernoulli}(1, 0.05)$  for  $i < j$  and  $\omega_{j,i}^{*(3)} = \omega_{i,j}^{*(3)}$ .  $\mathbf{\Omega}^{(3)} = \mathbf{D}^{1/2} (\mathbf{\Omega}^{*(3)} + \delta \mathbf{I}) / (1 + \delta) \mathbf{D}^{1/2}$  with  $\delta = |\lambda_{\min}(\mathbf{\Omega}^{*(3)})| + 0.05$ .

**Global Test** For the global testing of  $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ , the sample sizes were taken to be  $n = n_1 = n_2 = 100$ , while the dimension  $p$  varied over the values 100, 200, 400, and 1000. Under the global null hypothesis, we have  $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}$ , and two scenarios of generating  $\boldsymbol{\beta}$  were considered. For case 1, 10 nonzero locations  $\{k_1, \dots, k_{10}\}$  of  $\boldsymbol{\beta}$  were randomly generated with magnitudes  $\beta_{k_i,1} = 2i^{0.5}n_1^{-0.15}$ ,  $i = 1, \dots, 10$ . For case 2,  $s$  nonzero locations for  $\boldsymbol{\beta}$  were randomly selected, with  $s = 5, 8, 10$ , and 15 for  $p = 100, 200, 400$  and 1000, respectively. The nonzero locations had magnitudes with any values between -10 and 10. The error terms  $\epsilon_{k,d}$  were generated as normal random variables with mean 0 and variances having any values between 0.5 and 2.5. The nominal significance level for all the tests was set at  $\alpha_1 = 0.05$ .

Table 1 shows that the sizes of the global test  $\Psi_{\alpha_1}$  are close to the nominal level for both cases under all matrix models. This reflects the fact that the null distribution of the test statistics  $M_n$  is well approximated by its limiting null distribution, as shown in Theorem 1. The empirical sizes are slightly below the nominal level in some cases for lower dimensions, as similarly observed in Xia et al. (2015), due to correlation among the variables. It is also

shown in Table 1 that the proposed test is powerful in all settings, though  $\beta_1$  and  $\beta_2$  only differ in five or fewer locations with magnitudes of the order  $\sqrt{\log p/n}$ .

To evaluate the power of the global test, we selected five locations,  $\{k_1, \dots, k_5\}$ , among the nonzero locations of  $\beta_1$ , with magnitudes  $\beta_{k_j,2} = \beta_{k_j,1} + u_j$ ,  $j = 1, \dots, m$ , where  $u_j$  has magnitude randomly and uniformly from the set  $[-2\beta(\log p/n)^{1/2}, -\beta(2 \log p/n)^{1/2}] \cup [\beta(2 \log p/n)^{1/2}, 2\beta(\log p/n)^{1/2}]$ , with  $\beta = \max_{1 \leq i \leq p} |\beta_{i,1}|$ . The actual sizes and powers in percentage for each case under three matrix models, reported in Table 1, are estimated from 1000 replications. For each replication, the nonzero locations and magnitudes of the regression coefficients could vary.

**Multiple Testing** For simultaneous testing of  $\{H_{0,i} : \beta_{i,1} - \beta_{i,2} = 0, \text{ for } 1 \leq i \leq p\}$  with FDR control, we first generated  $\beta_1$  according to the above two cases. For case 1, ten nonzero locations  $\{k'_1, \dots, k'_{10}\}$  for  $\beta_2$  were randomly generated and the locations could vary for these two vectors. The magnitudes were generated with values  $\beta_{k'_i,2} = 4i^{0.5}n_2^{-0.15}$ ,  $i = 1, \dots, 10$ . For case 2,  $s$  nonzero locations for  $\beta_2$  were randomly selected, again with  $s = 5, 8, 10$ , and 15 for  $p = 100, 200, 400$  and 1000, respectively, also with magnitudes having any values between -10 and 10.

In Table 2, we present the empirical FDR and true discovery rate (power) of the proposed procedure (NEW) and the B-Y procedure at the FDR level of  $\alpha_2 = 0.1$ , based on 100 replications, where the power is summarized based on

$$\frac{1}{100} \sum_{l=1}^{100} \frac{\sum_{i \in \mathcal{H}_1} I(|W_{i,l}| \geq \hat{t})}{|\mathcal{H}_1|},$$

where  $W_{i,l}$  denotes standardized difference for the  $l^{\text{th}}$  replication and  $\mathcal{H}_1$  denotes the nonzero

$p$	Case 1			Case 2		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
	Size					
100	4.1	3.2	2.9	4.4	2.9	2.8
400	4.8	3.8	3.7	4.0	4.1	3.5
1000	6.1	4.4	5.4	5.9	4.6	6.4
	Power					
100	71.9	64.3	67.4	95.1	97.1	96.6
400	88.3	86.2	83.5	82.3	77.0	82.1
1000	95.1	92.6	97.9	47.3	42.0	48.1

Table 1: Empirical sizes and powers (%) for global testing with  $\alpha_1 = 0.05$ ,  $n_1 = n_2 = 100$ , and 1000 replications.

locations of  $\beta_1 - \beta_2$ . The results suggest that across all configurations, the FDRs are well controlled under the nominal level  $\alpha$  by both FDR control procedures. However, the B-Y procedure is extremely conservative in all scenarios. For the new FDR procedure, the empirical FDRs are also conservative, due to the correlations among the regression residuals under the nulls  $\mathcal{H}_{0,i}$ , and also due to the fact that we use  $|\mathcal{H}|$  to estimate  $|\mathcal{H}_0|$  because the latter is usually unknown. Furthermore, the total number of true signals is small in all cases due to the sparsity of the regression coefficients; for example, when the total number of true signals is ten, the FDP for each replication tends to be either 0 or some number close to 0.1, which will also cause the conservatism of the FDR estimation. In case 2, we can see that the empirical FDR gets closer to the nominal level as dimension increases, because the number of true signals increases when  $p$  grows. In summary, the new procedure has empirical FDR much closer to the nominal level than B-Y procedure in all cases. Table 2 also reflects that the FDR control procedure introduced in Section 4 is more powerful than the B-Y procedure for different scenarios.

### 5.3 Simulation by Mimicking Data

We now consider a simulation setting mimicking the data considered in Section 6, where we have  $p = 119$ ,  $n_1 = 46$  and  $n_2 = 417$ . We investigated both cases of the construction of the regression coefficients as considered in Section 5.2, with ten nonzero locations, under all three matrix models, with covariates as a combination of continuous and discrete random variables. The nominal level was set at  $\alpha_3 = 0.1$ , and the empirical FDR's and powers for both FDR procedures, as reported in Table 3, were evaluated based on 100 replications. As

$p$	Method	Case 1			Case 2		
		Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
		Size					
100	NEW	5.9	5.8	6.8	3.8	4.5	3.6
	B-Y	0.3	1.0	0.7	0.1	0.3	0.7
400	NEW	6.7	7.4	6.8	6.2	5.5	5.5
	B-Y	0.4	0.6	0.4	0.2	0.7	0.5
1000	NEW	6.2	6.0	6.1	9.4	9.4	9.8
	B-Y	0.6	1.0	0.4	1.5	1.6	1.4
		Power					
100	NEW	95.3	94.7	94.7	93.3	92.1	90.4
	B-Y	91.5	88.1	88.5	88.6	90.3	88.3
400	NEW	92.7	88.2	90.8	84.3	82.9	83.6
	B-Y	86.1	82.2	84.3	81.5	78.7	81.3
1000	NEW	84.7	82.7	85.1	71.7	70.4	71.9
	B-Y	77.7	75.0	77.6	66.2	64.5	66.1

Table 2: Empirical FDRs and powers (%) for the new FDR procedure and B-Y procedure with  $\alpha_2 = 0.1$ ,  $n_1 = n_2 = 100$ , and 100 replications.

in Section 5.2, the empirical FDRs are well controlled under the data setting by the new FDR procedure, while the B-Y procedure is again very conservative. For case 1, the empirical FDR's of the new procedure are slightly larger than the nominal level, due to the fact that  $n_1$  is much smaller than  $n_2$  in this setting, and thus  $\beta_1$  and  $\beta_2$  have magnitudes much closer to each other based on their construction. The performance of the new method for case 2 is less conservative than in Section 5.2 due to the fact we have ten nonzero locations for the regression coefficients when the dimension is 119 in the data setting. Table 3 also indicates that the new procedure is more powerful than the B-Y procedure under the data setting in all scenarios.

$p$	Method	Case 1			Case 2		
		Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
		Size					
119	NEW	9.4	11.2	11.0	8.7	8.9	8.8
	B-Y	2.2	3.0	2.9	1.7	1.4	1.6
		Power					
119	NEW	83.6	81.7	83.9	79.6	78.2	80.3
	B-Y	76.2	72.1	74.8	73.7	72.6	74.6

Table 3: Empirical FDRs and powers (%) for the new FDR procedure and B-Y procedure under the data setting with  $\alpha_3 = 0.1$ ,  $p = 119$ ,  $n_1 = 46$ ,  $n_2 = 417$ , and 100 replications.

## 6 Data Analysis

We illustrate our proposed methods using the Framingham Offspring Study (Kannel et al. (1979)) of coronary artery disease (CAD). Over the past three decades, various risk prediction models for CAD have been developed (Wilson et al. (1998); Ridker et al. (2007)). Unlike those for many other diseases, the risk models such as the Framingham Risk Score have been incorporated into clinical practice guidelines (Lloyd-Jones et al. (2004); D'Agostino Sr et al. (2008)). However, these models, largely based on traditional clinical risk factors, have recognized limitations in their clinical utilities. It is thus important to explore avenues beyond the routine clinical measures to improve prediction. One potential approach is to fully understand the roles of intermediate phenotypes, such as the C- reactive protein (CRP) and genomic markers. In recent years, many genome-wide association studies (GWAS) have been conducted to identify CAD-related single-nucleotide polymorphism (SNP) mutations. The newly identified SNPs, while significantly associated with CAD risk or the intermediate phenotypes of CAD, explain very little of the genetic risk for the trait (Humphries et al. (2008); Paynter et al. (2009)). This coincides with the growing awareness that the failure to identify genetic scores that significantly improve risk prediction for complex traits may be in part due to failure to account for the interplay of genes and environment. It is thus of substantial interests to study environment and its interaction with a genetic predisposition in causing human diseases.

Here, we use data from Framingham Offspring Study to examine how the interaction between smoking and genetic risk factors affect the inflammation marker CRP, since the in-

inflammation system plays a vital role in the atherosclerotic process (Ross (1999)). We focus on the 463 female participants with complete information on CRP, 116 SNP's previously reported as associated with CAD intermediate phenotypes, two leading principal components that adjust for population stratification, as well as age and smoking status at exam seven. Smoking is known to roughly double life-time risk of CAD and is thought to increase cardiovascular risk via a few different mechanisms. We examine the interaction between smoking and the genetic markers, as well as other risk factors based on the proposed method. We fit linear regression models for smokers and for non-smokers and the variables with significantly different coefficients between smokers and non-smokers are deemed as having an interactive effect.

The effects of top eight SNPs including rs11585329, rs17583120, rs17132534, rs11214606, rs17529477, rs10891552, rs4293, and rs4351, on CRP are considered as significantly modified by smoking. Interestingly, the smoking and rs11585329 interaction has been reported as important contributor to the risk of colorectal cancer whereas inflammation is a hallmark of cancer (Liu et al. (2013)). SNP rs17132534 belongs to the UCP2 gene whose main function is the control of mitochondria-derived reactive oxygen species. A variant in the UCP2 has been previously shown to interact with smoking to influence plasma markers of oxidative stress and hence likely to be associated with prospective CHD risk (Stephens et al. (2008)). SNPs rs10891552, rs17529477, and rs11214606 all belong to the DRD2 gene, which is linked to addictive behaviors, including alcoholism and smoking. Smoking was found to modify the effects of polymorphism in DRD2 gene on gastric cancer risk (Ikeda et al. (2008)). SNPs rs4293 and rs 4351 belong to the ACE gene, linked with hypertension and CAD among

other disorders. Interactions between smoking and polymorphism in the ACE gene have been reported for blood pressure and coronary atherosclerosis (Hibi et al. (1997); Sayed-Tabatabaei et al. (2004); Schut et al. (2004)).

## 7 Extension to Non-Binary Environmental Variable

Motivated by applications in genomics, we have proposed hypothesis testing procedures for detecting the interactions between environment and genomic markers when the environmental variable is binary, such as smoking status, as illustrated in Section 6. Our testing approach can be extended to detect the interactions when the environmental variable is discrete and finite, but non-binary. Specifically, suppose the environmental variable takes  $K$  possible values. Interaction detection can then be formulated based on comparing  $K$  high-dimensional regression models

$$\mathbf{Y}_d = \boldsymbol{\mu}_d + \mathbf{X}_d \boldsymbol{\beta}_d + \boldsymbol{\epsilon}_d, \quad \text{for } d = 1, \dots, K.$$

One wishes to develop a global test for

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \dots = \boldsymbol{\beta}_K \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_l \neq \boldsymbol{\beta}_k \quad \text{for some } 1 \leq l < k \leq K, \quad (21)$$

as well as develop a procedure for simultaneously testing the hypotheses

$$H_{0,i} : \beta_{i,1} = \beta_{i,2} = \dots = \beta_{i,K} \quad \text{versus} \quad H_{1,i} : \beta_{i,l} \neq \beta_{i,k} \quad \text{for some } 1 \leq l < k \leq K, \quad i = 1, \dots, p, \quad (22)$$

with FDR and FDP control.

The test statistics for each model can be formulated similarly as in Section 2.2. For  $d = 1, \dots, K$ , we let

$$T_{i,d} = \hat{r}_{i,d} / \hat{\sigma}_{\eta_{i,d}}^2$$

and estimate  $\theta_{i,d}$  by

$$\hat{\theta}_{i,d} = (\hat{\sigma}_{\epsilon_d}^2 / \hat{\sigma}_{\eta_{i,d}}^2 + \hat{\beta}_{i,d}^2) / n_d.$$

Then the pairwise standardized statistics can be defined by

$$W_i^{(l,k)} = \frac{T_{i,l} - T_{i,k}}{(\hat{\theta}_{i,l} + \hat{\theta}_{i,k})^{1/2}}, \quad 1 \leq l < k \leq K, \quad i = 1, \dots, p.$$

Then if  $K$  is finite, we construct the sum of square type test statistic by

$$S_i = \sum_{1 \leq l < k \leq K} (W_i^{(l,k)})^2.$$

As in Cai and Xia (2014), it can be shown that the limiting null distribution of  $S_i$  is a mixture chi-square distribution. Based on this fact, we can further develop global and multiple testing procedures. When the environmental variable is binary, the test statistics  $S_i$  reduce to (11) in Section 2.2. On the other hand, if the environmental variable is continuous, the testing problem is significantly different, and out of the scope of the current paper. We leave it to future research.

## 8 Proofs

We prove the main results in this section. We begin by collecting technical lemmas that will be used in the proof of the main theorems.

## 8.1 Technical Lemmas

The first lemma is the classical Bonferroni inequality.

**Lemma 1 (Bonferroni inequality)** *Let  $B = \cup_{t=1}^p B_t$ . For any  $k < [p/2]$ , we have*

$$\sum_{t=1}^{2k} (-1)^{t-1} F_t \leq P(B) \leq \sum_{t=1}^{2k-1} (-1)^{t-1} F_t,$$

where  $F_t = \sum_{1 \leq i_1 < \dots < i_t \leq p} P(B_{i_1} \cap \dots \cap B_{i_t})$ .

For  $d = 1, 2$ , let  $U_{i,d} = n_d^{-1} \sum_{k=1}^{n_d} \{\epsilon_{k,d} \eta_{k,i,d} - \mathbb{E}(\epsilon_{k,d} \eta_{k,i,d})\}$  and  $\tilde{U}_{i,d} = \beta_{i,d} + U_{i,d} / \sigma_{\eta_{i,d}}^2$ . The following lemma is essentially proved in Liu and Luo (2014).

**Lemma 2** *Suppose that Conditions (C1), (C3), (7) and (8) hold. Then*

$$T_{i,d} = \tilde{U}_{i,d} + (\tilde{\sigma}_{\epsilon_d}^2 / \sigma_{\epsilon_d}^2 + \tilde{\sigma}_{\eta_{i,d}}^2 / \sigma_{\eta_{i,d}}^2 - 2) \beta_{i,d} + o_P\{(n_d \log p)^{-1/2}\},$$

where  $\tilde{\sigma}_{\epsilon_d}^2 = n_d^{-1} \sum_{k=1}^{n_d} (\epsilon_{k,d} - \bar{\epsilon}_{k,d})^2$  and  $\tilde{\sigma}_{\eta_{i,d}}^2 = n_d^{-1} \sum_{k=1}^{n_d} (\eta_{k,i,d} - \bar{\eta}_{k,i,d})^2$  with  $\bar{\epsilon}_{k,d} = n_d^{-1} \sum_{k=1}^{n_d} \epsilon_{k,d}$  and  $\bar{\eta}_{k,i,d} = n_d^{-1} \sum_{k=1}^{n_d} \eta_{k,i,d}$ . Consequently, uniformly in  $i = 1, \dots, p$ ,

$$|T_{i,d} - \tilde{U}_{i,d}| = O_P\{\beta_{i,d} (\log p / n_d)^{1/2}\} + o_P\{(n_d \log p)^{-1/2}\}.$$

**Lemma 3** *Let  $\mathbf{X}_k \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$  for  $k = 1, \dots, n_1$  and  $\mathbf{Y}_k \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$  for  $k = 1, \dots, n_2$ .*

*Define*

$$\tilde{\boldsymbol{\Sigma}}_1 = (\tilde{\sigma}_{i,j,1})_{p \times p} = \frac{1}{n_1} \sum_{k=1}^{n_1} (\mathbf{X} - \boldsymbol{\mu}_1)(\mathbf{X} - \boldsymbol{\mu}_1)^\top, \quad \tilde{\boldsymbol{\Sigma}}_2 = (\tilde{\sigma}_{i,j,2})_{p \times p} = \frac{1}{n_2} \sum_{k=1}^{n_2} (\mathbf{Y} - \boldsymbol{\mu}_2)(\mathbf{Y} - \boldsymbol{\mu}_2)^\top.$$

*Then, for some constant  $C > 0$ ,  $\tilde{\sigma}_{i,j,1} - \tilde{\sigma}_{i,j,2}$  satisfies the large deviation bound*

$$P \left[ \max_{(i,j) \in \mathcal{S}} \frac{(\tilde{\sigma}_{i,j,1} - \tilde{\sigma}_{i,j,2} - \sigma_{i,j,1} + \sigma_{i,j,2})^2}{\text{Var}\{(X_{k,i} - \mu_{1,i})(X_{k,j} - \mu_{1,j})\} / n_1 + \text{Var}\{(Y_{k,i} - \mu_{2,i})(Y_{k,j} - \mu_{2,j})\} / n_2} \geq x^2 \right] \leq C |\mathcal{S}| \{1 - \Phi(x)\} + O(p^{-1})$$

*uniformly for  $0 \leq x \leq (8 \log p)^{1/2}$  and any subset  $\mathcal{S} \subseteq \{(i, j) : 1 \leq i \leq j \leq p\}$ .*

The complete proof of this lemma can be found in the supplementary material of Xia et al. (2015).

## 8.2 Proof of Theorem 1

To prove Theorem 1, we first show that the terms in  $A_\tau$  are negligible. Then we focus on the terms in  $\mathcal{H} \setminus A_\tau$ , where  $\mathcal{H} = \{1, \dots, p\}$ , and show that  $\mathbf{P}(\max_{i \in \mathcal{H} \setminus A_\tau} W_i^2 - 2 \log p + \log \log p \leq t) \rightarrow \exp(-\pi^{-1/2} \exp(-t/2))$ , where  $W_i$  is defined in (11).

Define

$$V_i = \frac{U_{i,1}/\sigma_{\eta_{i,1}}^2 - U_{i,2}/\sigma_{\eta_{i,2}}^2}{(\theta_{i,1} + \theta_{i,2})^{1/2}},$$

where  $\theta_{i,d} = \text{Var}(\tilde{U}_{i,d}) = \text{Var}(\epsilon_{k,d} \eta_{k,i,d} / \sigma_{\eta_{i,d}}^2) / n_d = (\sigma_{\epsilon_d}^2 / \sigma_{\eta_{i,d}}^2 + \beta_{i,d}^2) / n_d$ , for  $d = 1, 2$ . By Lemma 2 in Xia et al. (2015), under conditions (7) and (8), we have

$$|\hat{\sigma}_{\epsilon_d}^2 - \sigma_{\epsilon_d}^2| = O_{\mathbf{P}}\left(\sqrt{\frac{\log p}{n_d}}\right), \text{ and } \max_i |\hat{\sigma}_{\eta_{i,d}}^2 - \sigma_{\eta_{i,d}}^2| = O_{\mathbf{P}}\left(\sqrt{\frac{\log p}{n_d}}\right). \quad (23)$$

Thus we have

$$\max_i |\hat{\theta}_{i,d} - \theta_{i,d}| = o_{\mathbf{P}}(1/(n_d \log p)). \quad (24)$$

By Lemma 2, we have

$$W_i = V_i + b_i + o_{\mathbf{P}}\{(\log p)^{-1/2}\},$$

where  $b_i = \{(\tilde{\sigma}_{\epsilon_1}^2 / \sigma_{\epsilon_1}^2 + \tilde{\sigma}_{\eta_{i,1}}^2 / \sigma_{\eta_{i,1}}^2) \beta_{i,1} - (\tilde{\sigma}_{\epsilon_2}^2 / \sigma_{\epsilon_2}^2 + \tilde{\sigma}_{\eta_{i,2}}^2 / \sigma_{\eta_{i,2}}^2) \beta_{i,2}\} / (\hat{\theta}_{i,1} + \hat{\theta}_{i,2})^{1/2}$ . Note that for  $i \in \mathcal{H} \setminus A_\tau$ ,  $\beta_{i,d} = o\{(\log p)^{-1}\}$ . Thus we have  $\max_{i \in \mathcal{H} \setminus A_\tau} |W_i - V_i| = o_{\mathbf{P}}\{(\log p)^{-1/2}\}$ . For  $i \in A_\tau$ ,

$$b_i \leq \left| \frac{\tilde{\sigma}_{\epsilon_1}^2 \beta_{i,1} / \sigma_{\epsilon_1}^2 - \tilde{\sigma}_{\epsilon_2}^2 \beta_{i,2} / \sigma_{\epsilon_2}^2}{\{\text{Var}(\epsilon_{k,1}^2) \beta_{i,1}^2 / (\sigma_{\epsilon_1}^4 n_1) + \text{Var}(\epsilon_{k,2}^2) \beta_{i,2}^2 / (\sigma_{\epsilon_2}^4 n_2)\}^{1/2}} \right|$$

$$+ \left| \frac{\tilde{\sigma}_{\eta_{i,1}}^2 \beta_{i,1} / \sigma_{\eta_{i,1}}^2 - \tilde{\sigma}_{\eta_{i,2}}^2 \beta_{i,2} / \sigma_{\eta_{i,2}}^2}{\{\text{Var}(\eta_{k,i,1}^2) \beta_{i,1}^2 / (\sigma_{\eta_{i,1}}^4 n_1) + \text{Var}(\eta_{k,i,2}^2) \beta_{i,2}^2 / (\sigma_{\eta_{i,2}}^4 n_2)\}^{1/2}} \right| + o_{\mathbb{P}}\{(\log p)^{-1/2}\}.$$

Due to the fact that the indices of the random variables only show up in the second term here, by Lemma 3 and the condition that  $|A_\tau| = O(p^r)$  with  $r < 1/4$ , we have

$$\mathbb{P}(\max_{i \in A_\tau} W_i^2 \geq 2 \log p - \log \log p + t) \leq |A_\tau| \{\mathbb{P}(V_i^2 \geq 2r \log p) + \mathbb{P}(\tilde{b}_i^2 \geq 2r \log p)\} + o(1) = o(1),$$

where  $\tilde{b}_i = \left| \frac{\tilde{\sigma}_{\eta_{i,1}}^2 \beta_{i,1} / \sigma_{\eta_{i,1}}^2 - \tilde{\sigma}_{\eta_{i,2}}^2 \beta_{i,2} / \sigma_{\eta_{i,2}}^2}{\{\text{Var}(\eta_{k,i,1}^2) \beta_{i,1}^2 / (\sigma_{\eta_{i,1}}^4 n_1) + \text{Var}(\eta_{k,i,2}^2) \beta_{i,2}^2 / (\sigma_{\eta_{i,2}}^4 n_2)\}^{1/2}} \right|$ . Thus, it suffices to show that

$$\mathbb{P}(\max_{i \in \mathcal{H} \setminus A_\tau} V_i^2 - 2 \log p + \log \log p \leq t) \rightarrow \exp(-\pi^{-1/2} \exp(-t/2)).$$

Let  $q = |\mathcal{H} \setminus A_\tau|$  and let  $n_2/n_1 \leq K_1$  with  $K_1 \geq 1$ . Define  $Z_{k,i} = (n_2/n_1) \{\epsilon_{k,1} \eta_{k,i,1} - \mathbb{E}(\epsilon_{k,1} \eta_{k,i,1})\} / \sigma_{\eta_{i,1}}^2$  for  $1 \leq k \leq n_1$  and  $Z_{k,i} = -\{\epsilon_{k,2} \eta_{k,i,2} - \mathbb{E}(\epsilon_{k,2} \eta_{k,i,2})\} / \sigma_{\eta_{i,2}}^2$  for  $n_1 + 1 \leq k \leq n_2$ . Thus we have

$$V_i = \frac{\sum_{k=1}^{n_1+n_2} Z_{k,i}}{(n_2^2 \theta_{k,1} / n_1 + n_2 \theta_{k,2})^{1/2}}.$$

Without loss of generality, we assume  $\sigma_{\epsilon_d}^2 = \sigma_{\eta_{i,d}}^2 = 1$ . Define

$$\hat{V}_i = \frac{\sum_{k=1}^{n_1+n_2} \hat{Z}_{k,i}}{(n_2^2 \theta_{k,1} / n_1 + n_2 \theta_{k,2})^{1/2}},$$

where  $\hat{Z}_{k,i} = Z_{k,i} I(|Z_{k,i}| \leq \tau_n) - \mathbb{E}\{Z_{k,i} I(|Z_{k,i}| \leq \tau_n)\}$ , and  $\tau_n = (4K_1/K) \log(p+n)$ . Note that  $\max_{i \in \mathcal{H} \setminus A_\tau} V_i^2 = \max_{1 \leq i \leq q} \hat{V}_i^2$ , and that

$$\begin{aligned} & \max_{1 \leq i \leq q} n^{-1/2} \sum_{k=1}^{n_1+n_2} \mathbb{E}[|Z_{k,i}| I\{|Z_{k,i}| \geq (4K_1/K) \log(p+n)\}] \\ & \leq C n^{1/2} \max_{1 \leq k \leq n_1+n_2} \max_{1 \leq i \leq q} \mathbb{E}[|Z_{k,i}| I\{|Z_{k,i}| \geq (4K_1/K) \log(p+n)\}] \\ & \leq C n^{1/2} (p+n)^{-2} \max_{1 \leq k \leq n_1+n_2} \max_{1 \leq i \leq q} \mathbb{E}[|Z_{k,i}| \exp\{(K/2)|Z_{k,i}|\}] \\ & \leq C n^{1/2} (p+n)^{-2}. \end{aligned}$$

Hence,  $\mathbb{P}\left\{\max_{1 \leq i \leq q} |V_i - \hat{V}_i| \geq (\log p)^{-1}\right\} \leq \mathbb{P}\left(\max_{1 \leq i \leq q} \max_{1 \leq k \leq n_1+n_2} |Z_{k,i}| \geq \tau_n\right) = O(p^{-1})$ . By the fact that  $\left|\max_{1 \leq i \leq q} V_i^2 - \max_{1 \leq i \leq q} \hat{V}_i^2\right| \leq 2 \max_{1 \leq i \leq q} |\hat{V}_i| \max_{1 \leq i \leq q} |V_i - \hat{V}_i| + \max_{1 \leq i \leq q} |V_i - \hat{V}_i|^2$ , it suffices to prove that for any  $t \in \mathbb{R}$ , as  $n, p \rightarrow \infty$ ,

$$\mathbb{P}\left(\max_{1 \leq i \leq q} \hat{V}_i^2 - 2 \log p + \log \log p \leq t\right) \rightarrow \exp(-\pi^{-1/2} \exp(-t/2)). \quad (25)$$

By Lemma 1, for any integer  $l$  with  $0 < l < q/2$ ,

$$\begin{aligned} \sum_{d=1}^{2l} (-1)^{d-1} \sum_{1 \leq i_1 < \dots < i_d \leq q} \mathbb{P}\left(\bigcap_{j=1}^d F_{i_j}\right) &\leq \mathbb{P}\left(\max_{1 \leq i \leq q} \hat{V}_i^2 \geq y_p\right) \\ &\leq \sum_{d=1}^{2l-1} (-1)^{d-1} \sum_{1 \leq i_1 < \dots < i_d \leq q} \mathbb{P}\left(\bigcap_{j=1}^d F_{i_j}\right), \end{aligned} \quad (26)$$

where  $y_p = 2 \log p - \log \log p + t$  and  $F_{i_j} = (\hat{V}_{i_j}^2 \geq y_p)$ . Let  $\tilde{Z}_{k,i} = \hat{Z}_{k,i}/(n_2 \theta_{i,1}/n_1 + \theta_{i,2})^{1/2}$  for  $i = 1, \dots, q$  and  $\mathbf{W}_k = (\tilde{Z}_{k,i_1}, \dots, \tilde{Z}_{k,i_d})$ , for  $1 \leq k \leq n_1 + n_2$ . Define  $|\mathbf{a}|_{\min} = \min_{1 \leq i \leq d} |a_i|$  for any vector  $\mathbf{a} \in R^d$ . Then we have

$$\mathbb{P}\left(\bigcap_{j=1}^d F_{i_j}\right) = \mathbb{P}\left(\left|n_2^{-\frac{1}{2}} \sum_{k=1}^{n_1+n_2} \mathbf{W}_k\right|_{\min} \geq y_p^{\frac{1}{2}}\right).$$

Then it follows from Theorem 1 in Zaitsev (1987) that

$$\begin{aligned} \mathbb{P}\left(\left|n_2^{-1/2} \sum_{k=1}^{n_1+n_2} \mathbf{W}_k\right|_{\min} \geq y_p^{1/2}\right) &\leq \mathbb{P}\left\{|\mathbf{N}_d|_{\min} \geq y_p^{1/2} - \epsilon_n (\log p)^{-1/2}\right\} \\ &\quad + c_1 d^{\frac{5}{2}} \exp\left\{-\frac{n^{1/2} \epsilon_n}{c_2 d^3 \tau_n (\log p)^{1/2}}\right\}, \end{aligned} \quad (27)$$

where  $c_1 > 0$  and  $c_2 > 0$  are constants,  $\epsilon_n \rightarrow 0$  which will be specified later, and  $\mathbf{N}_d = (N_{m_1}, \dots, N_{m_d})$  is a normal random vector with  $\mathbf{E}(\mathbf{N}_d) = 0$  and  $\text{Cov}(\mathbf{N}_d) = n_1/n_2 \text{Cov}(\mathbf{W}_1) + \text{Cov}(\mathbf{W}_{n_1+1})$ . Here  $d$  is a fixed integer that does not depend on  $n, p$ . Because  $\log p = o(n^{1/5})$ , we can let  $\epsilon_n \rightarrow 0$  sufficiently slowly that, for any large  $M > 0$ ,

$$c_1 d^{5/2} \exp\left\{-\frac{n^{1/2} \epsilon_n}{c_2 d^3 \tau_n (\log p)^{1/2}}\right\} = O(p^{-M}). \quad (28)$$

Combining (26), (27), and (28) we have

$$\mathbb{P}\left(\max_{1 \leq i \leq q} \hat{V}_i^2 \geq y_p\right) \leq \sum_{d=1}^{2l-1} (-1)^{d-1} \sum_{1 \leq i_1 < \dots < i_d \leq q} \mathbb{P}\left\{|\mathbf{N}_d|_{\min} \geq y_p^{1/2} - \epsilon_n (\log p)^{-1/2}\right\} + o(1). \quad (29)$$

Similarly, using Theorem 1 in Zaitsev (1987) again, we can get

$$\mathbb{P}\left(\max_{1 \leq i \leq q} \hat{V}_i^2 \geq y_p\right) \geq \sum_{d=1}^{2l} (-1)^{d-1} \sum_{1 \leq i_1 < \dots < i_d \leq q} \mathbb{P}\left\{|\mathbf{N}_d|_{\min} \geq y_p^{1/2} + \epsilon_n (\log p)^{-1/2}\right\} - o(1). \quad (30)$$

The following lemma is shown in the supplementary material of Cai et al. (2013) with  $q \asymp p$  and  $y_p = 2 \log p - \log \log p + t$ .

**Lemma 4** For any fixed integer  $d \geq 1$  and real number  $t \in \mathbb{R}$ ,

$$\sum_{1 \leq i_1 < \dots < i_d \leq q} \mathbb{P}\left\{|\mathbf{N}_d|_{\min} \geq y_p^{1/2} \pm \epsilon_n (\log p)^{-1/2}\right\} = \frac{1}{d!} \{(\pi)^{-1/2} \exp(-t/2)\}^d \{1 + o(1)\}. \quad (31)$$

It then follows from Lemma 4, (29), and (30) that

$$\begin{aligned} \limsup_{n, p \rightarrow \infty} \mathbb{P}\left(\max_{1 \leq i \leq q} \hat{V}_i^2 \geq y_p\right) &\leq \sum_{d=1}^{2l} (-1)^{d-1} \frac{1}{d!} \{(\pi)^{-1/2} \exp(-t/2)\}^d \\ \liminf_{n, p \rightarrow \infty} \mathbb{P}\left(\max_{1 \leq i \leq q} \hat{V}_i^2 \geq y_p\right) &\geq \sum_{d=1}^{2l-1} (-1)^{d-1} \frac{1}{d!} \{(\pi)^{-1/2} \exp(-t/2)\}^d \end{aligned}$$

for any positive integer  $l$ . By letting  $l \rightarrow \infty$ , we obtain (25) and Theorem 1 is proved.  $\blacksquare$

### 8.3 Proof of Theorem 2

Let  $M_n^1 = \max_{1 \leq i \leq j \leq p} \{T_{i,1} - T_{i,2} - (\beta_{i,1} - \beta_{i,2})\}^2 / (\hat{\theta}_{i,1} + \hat{\theta}_{i,2})$ . It follows from the proof of Theorem 1 that  $\mathbb{P}(M_n^1 \leq 2 \log p - 2^{-1} \log \log p) \rightarrow 1$ , as  $n, p \rightarrow \infty$ . By (23), (24), and the inequalities

$$\max_{1 \leq i \leq p} \frac{(\beta_{i,1} - \beta_{i,2})^2}{(\hat{\theta}_{i,1} + \hat{\theta}_{i,2})} \leq 2M_n^1 + 2M_n,$$

$$\max_{1 \leq i \leq p} \frac{|\beta_{i,1} - \beta_{i,2}|}{(\theta_{i,1} + \theta_{i,2})^{1/2}} \geq 2\sqrt{2}(\log p)^{1/2},$$

we have  $P(M_n \geq q_\alpha + 2 \log p - \log \log p) \rightarrow 1$  as  $n, p \rightarrow \infty$ . ■

## 8.4 Proof of Theorem 3

To prove the lower bound, we first construct a worst case scenario to test between  $\beta_1$  and  $\beta_2$ . We apply the arguments in Baraud (2002) to prove the result.

Without loss of generality, we assume  $\sigma_{\epsilon_d}^2 = 1$ ,  $\sigma_{i,i,d} = 1$ ,  $\sigma_{i,j,d} = 0$ ,  $i \neq j$  for  $d = 1, 2$ , and  $n_1 = n_2$ . Let  $\hat{m}$  be a random entry uniformly drawn from  $\mathcal{H} = \{1, \dots, p\}$ . We construct a class of  $\beta_1$ ,  $\mathcal{N} = \{\beta^{(\hat{m})}, \hat{m} \in \mathcal{H}\}$ , such that,  $\beta_{\hat{m},1} = \rho$  and  $\beta_{i,1} = 0$  for  $i \neq \hat{m}$ , with  $\rho = c(\log p/n)^{1/2}$ , where  $c < 1/2$  is a constant. Let  $\beta_2 = 0$  and  $\beta_1$  be uniformly distributed on  $\mathcal{N}$ . Let  $\mu_\rho$  be the distribution on  $\beta_1 - \beta_2$ . Note that  $\mu_\rho$  is a probability measure on  $\{\delta \in \mathcal{S}_1 : |\delta|_2^2 = \rho^2\}$ , where  $\mathcal{S}_1$  is a class of  $p$ -dimensional vectors with one nonzero entry. Then the likelihood ratio between samples  $\{Y_{k,1}, \mathbf{X}_{k,1}\}$  and  $\{Y_{k,2}, \mathbf{X}_{k,2}\}$  can be calculated as

$$L_{\mu_\rho} = \mathbb{E}_{\hat{m}} \left[ \prod_{k=1}^n \frac{1}{|\Sigma^{(\hat{m})}|^{1/2}} \exp \left\{ -\frac{1}{2} \mathbf{Z}_k^\top (\Omega^{(\hat{m})} - \mathbf{I}) \mathbf{Z}_k \right\} \right],$$

where  $\Sigma^{(\hat{m})} = \Omega^{(\hat{m})^{-1}}$  is the covariance matrix of  $\{Y_{k,1}, \mathbf{X}_{k,1}\}$  and  $\{\mathbf{Z}_1, \dots, \mathbf{Z}_n\}$  are i.i.d samples generated from  $N(0, \mathbf{I})$ . Because  $\text{Var}(Y_{k,1}) = \sigma_{\hat{m},\hat{m},1} \beta_{\hat{m},1}^2 + 1$ ,  $\text{Var}(Y_{k,2}) = 1$  and  $\text{Cov}(Y_{k,d}, X_{k,i,d}) = \beta_{i,d} \sigma_{i,i,d}$ . It can be easily calculated that  $|\Sigma^{(\hat{m})}| = 1$  and  $\Omega^{(\hat{m})} = (\omega_{i,j}^{(\hat{m})})$  with  $\omega_{1,1}^{(\hat{m})} = 1$ ,  $\omega_{1,\hat{m}+1}^{(\hat{m})} = \omega_{\hat{m}+1,1}^{(\hat{m})} = -\rho$ ,  $\omega_{\hat{m}+1,\hat{m}+1}^{(\hat{m})} = 1 + \rho^2$ , and  $\omega_{i,j}^{(\hat{m})} = 0$  otherwise. Hence

$$L_{\mu_\rho}^2 = p^{-2} \sum_{m,m' \in \mathcal{H}} \mathbb{E} \left[ \prod_{k=1}^n \exp \left\{ -\frac{1}{2} \mathbf{Z}_k^\top (\Omega^{(m)} + \Omega^{(m')} - 2\mathbf{I}) \mathbf{Z}_k \right\} \right].$$

With  $\Omega^{(m)} + \Omega^{(m')} - 2\mathbf{I} = (a_{i,j})$ , it is easy to see that, when  $m \neq m'$ ,  $a_{i,i} = \rho^2$  and  $a_{1,i} = -\rho$  for  $i = m + 1$  or  $m' + 1$ ,  $a_{j,i} = a_{i,j}$  and  $a_{i,j} = 0$  otherwise; when  $m = m'$ ,  $a_{i,i} = 2\rho^2$  and  $a_{1,i} = -2\rho$  for  $i = m + 1$ ,  $a_{j,i} = a_{i,j}$  and  $a_{i,j} = 0$  otherwise. Thus we have

$$\mathbb{E}(L_{\mu\rho}^2) = [\mathbb{E}(\exp\{\rho(x_1x_2 + x_2x_3) - \rho^2(x_2^2 + x_3^2)/2\})^n + p^{-1}[\mathbb{E}(\exp\{2\rho x_1x_2 - \rho^2x_2^2\})^n],$$

where  $x_1, x_2, x_3$  are independent standard normal random variables. Because  $\mathbb{E}(\exp\{\rho(x_1x_2 + x_2x_3)\}) = 1 + \rho^2$ ,  $\mathbb{E}(\exp\{-\rho^2x_2^2/2\}) = (1 + \rho^2)^{-1/2}$  and  $\mathbb{E}(\exp\{2\rho x_1x_2\}) = 1 + 2\rho^2$ , we have

$$\mathbb{E}(L_{\mu\rho}^2) = 1 + p^{2c-1} + o(1) = 1 + o(1).$$

Theorem 3 is thus proved by Baraud (2002). ■

## 8.5 Proof of Theorem 4

We first show that  $\hat{t}$ , as defined in Section 4.1, is attained in the interval  $[0, (2 \log p)^{1/2}]$ . We then show that  $A_\tau$  is negligible and we focus on the set  $\mathcal{H} \setminus A_\tau$ . We then show the FDP result by dividing the null set into small subsets and controlling the variance of  $R_0(t)$  for each subset, and the FDR result will thus also be proved.

Under the condition of Theorem 4, we have

$$\sum_{1 \leq i \leq p} I\{|W_i| \geq (2 \log p)^{1/2}\} \geq \{1/(\pi^{1/2}\alpha) + \delta\}(\log p)^{1/2},$$

with probability going to one. Hence, with probability tending to one, we have

$$\frac{2p}{\sum_{1 \leq i \leq p} I\{|W_i| \geq (2 \log p)^{1/2}\}} \leq 2p\{1/(\pi^{1/2}\alpha) + \delta\}^{-1}(\log p)^{-1/2}.$$

Let  $t_p = (2 \log p - 2 \log \log p)^{1/2}$ . Because  $1 - \Phi(t_p) \sim 1/\{(2\pi)^{1/2}t_p\} \exp(-t_p^2/2)$ , we have  $\mathbb{P}(1 \leq \hat{t} \leq t_p) \rightarrow 1$  according to the definition of  $\hat{t}$  in Section 4.1. For  $0 \leq \hat{t} \leq t_p$ ,

$$\frac{2p\{1 - \Phi(\hat{t})\}}{\max\{\sum_{1 \leq i \leq p} I\{|W_i| \geq \hat{t}\}, 1\}} = \alpha.$$

Thus, to prove Theorem 4, it suffices to prove

$$\left| \frac{\sum_{i \in \mathcal{H}_0} I\{|W_i| \geq t\} - p_0 G(t)}{pG(t)} \right| \rightarrow 0,$$

in probability, uniformly for  $0 \leq t \leq t_p$ , where  $G(t) = 2(1 - \Phi(t))$  and  $p_0 = |\mathcal{H}_0|$ . We will show that it suffices to show

$$\left| \frac{\sum_{i \in \mathcal{H}_0 \setminus A_\tau} I\{|V_i| \geq t\} - p_0 G(t)}{pG(t)} \right| \rightarrow 0, \quad (32)$$

in probability. We now consider two cases.

1). If  $t = \{2 \log p + o(\log p)\}^{1/2}$ , the proof of Theorem 1 yields that  $\mathbb{P}(\max_{i \in A_\tau} W_i^2 \geq t^2) = o(1)$ . Thus, it suffices to prove

$$\left| \frac{\sum_{i \in \mathcal{H}_0 \setminus A_\tau} I\{|W_i| \geq t\} - p_0 G(t)}{pG(t)} \right| \rightarrow 0,$$

in probability. We show in Theorem 1 that  $\max_{i \in \mathcal{H}_0 \setminus A_\tau} |W_i - V_i| = o_{\mathbb{P}}\{(\log p)^{-1/2}\}$ . Thus it suffices to show (32).

2). If  $t \leq (C \log p)^{1/2}$  for some  $C < 2$ , we have

$$\left| \frac{\sum_{i \in A_\tau \cap \mathcal{H}_0} I\{|W_i| \geq t\}}{pG(t)} \right| \leq \frac{|A_\tau \cap \mathcal{H}_0|}{O(p^{1-C/2})} \rightarrow 0$$

in probability. Thus, it is again enough to show (32).

Let  $0 \leq t_0 < t_1 < \dots < t_b = t_p$  such that  $t_\iota - t_{\iota-1} = v_p$  for  $1 \leq \iota \leq b-1$  and  $t_b - t_{b-1} \leq v_p$ , where  $v_p = 1/\sqrt{\log p(\log_4 p)}$ . Thus we have  $b \sim t_p/v_p$ . For any  $t$  such that  $t_{\iota-1} \leq t \leq t_\iota$ , by

the fact that  $G(t + o((\log p)^{-1/2}))/G(t) = 1 + o(1)$  uniformly in  $0 \leq t \leq c(\log p)^{1/2}$  for any constant  $c$ , we have

$$\begin{aligned} \frac{\sum_{i \in \mathcal{H}_0 \setminus A_\tau} I(|V_i| \geq t_\iota)}{p_0 G(t_\iota)} \frac{G(t_\iota)}{G(t_{\iota-1})} &\leq \frac{\sum_{i \in \mathcal{H}_0 \setminus A_\tau} I(|V_i| \geq t)}{p_0 G(t)} \\ &\leq \frac{\sum_{i \in \mathcal{H}_0 \setminus A_\tau} I(|V_i| \geq t_{\iota-1})}{p_0 G(t_{\iota-1})} \frac{G(t_{\iota-1})}{G(t_\iota)}. \end{aligned}$$

Thus it suffices to prove

$$\max_{0 \leq \iota \leq b} \left| \frac{\sum_{i \in \mathcal{H}_0 \setminus A_\tau} \{I(|V_i| \geq t_\iota) - G(t_\iota)\}}{p_0 G(t_\iota)} \right| \rightarrow 0,$$

in probability. Define  $\tilde{\mathcal{H}}_0 = \mathcal{H}_0 \setminus A_\tau$ . Note that

$$\begin{aligned} \mathbb{P} \left[ \max_{0 \leq \iota \leq b} \left| \frac{\sum_{i \in \tilde{\mathcal{H}}_0} \{I(|V_i| \geq t_\iota) - G(t_\iota)\}}{p_0 G(t_\iota)} \right| \geq \epsilon \right] &\leq \sum_{\iota=1}^b \mathbb{P} \left[ \left| \frac{\sum_{i \in \tilde{\mathcal{H}}_0} \{I(|V_i| \geq t_\iota) - G(t_\iota)\}}{p_0 G(t_\iota)} \right| \geq \epsilon \right] \\ &\leq \frac{1}{v_p} \int_0^{t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in \tilde{\mathcal{H}}_0} I(|V_i| \geq t)}{p_0 G(t)} - 1 \right| \geq \epsilon \right\} dt + \sum_{\iota=b-1}^b \mathbb{P} \left[ \left| \frac{\sum_{i \in \tilde{\mathcal{H}}_0} \{I(|V_i| \geq t_\iota) - G(t_\iota)\}}{p_0 G(t_\iota)} \right| \geq \epsilon \right]. \end{aligned}$$

Thus, it suffices to show, for any  $\epsilon > 0$ ,

$$\int_0^{t_p} \mathbb{P} \left\{ \left| \frac{\sum_{i \in \tilde{\mathcal{H}}_0} \{I(|V_i| \geq t) - \mathbb{P}(I(|V_i| \geq t))\}}{p_0 G(t)} \right| \geq \epsilon \right\} dt = o(v_p). \quad (33)$$

Note that

$$\begin{aligned} &\mathbb{E} \left| \frac{\sum_{i \in \tilde{\mathcal{H}}_0} \{I(|V_i| \geq t) - \mathbb{P}(I(|V_i| \geq t))\}}{p_0 G(t)} \right|^2 \\ &= \frac{\sum_{i, j \in \tilde{\mathcal{H}}_0} \{\mathbb{P}(|V_i| \geq t, |V_j| \geq t) - \mathbb{P}(|V_i| \geq t)\mathbb{P}(|V_j| \geq t)\}}{p_0^2 G^2(t)}. \end{aligned}$$

We divide the indices  $i, j \in \tilde{\mathcal{H}}_0$  into the subsets:  $\tilde{\mathcal{H}}_{01} = \{i, j \in \tilde{\mathcal{H}}_0, i = j\}$ ,  $\tilde{\mathcal{H}}_{02} = \{i, j \in \tilde{\mathcal{H}}_0, i \in \Gamma_j(\gamma), \text{ or } j \in \Gamma_i(\gamma)\}$  and  $\tilde{\mathcal{H}}_{03} = \tilde{\mathcal{H}}_0 \setminus (\tilde{\mathcal{H}}_{01} \cup \tilde{\mathcal{H}}_{02})$ . Then we have

$$\frac{\sum_{i, j \in \tilde{\mathcal{H}}_{01}} \{\mathbb{P}(|V_i| \geq t, |V_j| \geq t) - \mathbb{P}(|V_i| \geq t)\mathbb{P}(|V_j| \geq t)\}}{p_0^2 G^2(t)} \leq \frac{C}{p_0 G(t)}. \quad (34)$$

We now show the equation (12). Note that  $\text{Cov}(\epsilon_{k,d}\eta_{k,i,d}, \epsilon_{k,d}\eta_{k,j,d}) = \mathbb{E}(\epsilon_{k,d}^2\eta_{k,i,d}\eta_{k,j,d}) - \mathbb{E}(\epsilon_{k,d}\eta_{k,i,d})\mathbb{E}(\epsilon_{k,d}\eta_{k,j,d})$ . Because  $\text{Cov}(\epsilon_{k,d}, \eta_{k,i,d}) = -\sigma_{\eta_{i,d}}^2\beta_{i,d}$ , we have  $\mathbb{E}(\epsilon_{k,d}\eta_{k,i,d})\mathbb{E}(\epsilon_{k,d}\eta_{k,j,d}) = \sigma_{\eta_{i,d}}^2\sigma_{\eta_{j,d}}^2\beta_{i,d}\beta_{j,d}$ . Note that

$$\begin{aligned} & \mathbb{E}(\epsilon_{k,d}^2\eta_{k,i,d}\eta_{k,j,d}) \\ &= \mathbb{E}\{\epsilon_{k,d}^2(\eta_{k,i,d} + \epsilon_{k,d}\gamma_{i,1,d})(\eta_{k,j,d} + \epsilon_{k,d}\gamma_{j,1,d})\} - \mathbb{E}\{\epsilon_{k,d}^2(\eta_{k,i,d} + \epsilon_{k,d}\gamma_{i,1,d})\epsilon_{k,d}\gamma_{j,1,d}\} - \mathbb{E}(\epsilon_{k,d}^3\gamma_{i,1,d}\eta_{k,j,d}). \end{aligned}$$

By definition, we have  $\epsilon_{k,d}$  independent with  $\eta_{k,i,d} + \epsilon_{k,d}\gamma_{i,1,d}$ . Thus, we have

$$\mathbb{E}(\epsilon_{k,d}^2\eta_{k,i,d}\eta_{k,j,d}) = \sigma_{\epsilon_d}^2 \mathbb{E}\{(\eta_{k,i,d} + \epsilon_{k,d}\gamma_{i,1,d})(\eta_{k,j,d} + \epsilon_{k,d}\gamma_{j,1,d})\} - \mathbb{E}(\epsilon_{k,d}^3\gamma_{i,1,d}\eta_{k,j,d}).$$

Note that

$$\mathbb{E}(\epsilon_{k,d}^3\gamma_{i,1,d}\eta_{k,j,d}) = \mathbb{E}\{\epsilon_{k,d}^3\gamma_{i,1,d}(\eta_{k,j,d} + \epsilon_{k,d}\gamma_{j,1,d})\} - \mathbb{E}(\epsilon_{k,d}^4\gamma_{i,1,d}\gamma_{j,1,d}) = -3\gamma_{i,1,d}\gamma_{j,1,d}\sigma_{\epsilon_d}^4,$$

and that

$$\begin{aligned} & \mathbb{E}\{(\eta_{k,i,d} + \epsilon_{k,d}\gamma_{i,1,d})(\eta_{k,j,d} + \epsilon_{k,d}\gamma_{j,1,d})\} \\ &= \text{Cov}(\eta_{k,i,d}, \eta_{k,j,d}) + \gamma_{i,1,d}\text{Cov}(\epsilon_{k,d}, \eta_{k,j,d}) + \gamma_{j,1,d}\text{Cov}(\epsilon_{k,d}, \eta_{k,i,d}) + \gamma_{i,1,d}\gamma_{j,1,d}\sigma_{\epsilon_d}^2. \end{aligned}$$

We have  $\text{Cov}(\epsilon_{k,d}\eta_{k,i,d}, \epsilon_{k,d}\eta_{k,j,d}) = (\omega_{i,j}\sigma_{\epsilon_d}^2 + 2\beta_{i,d}\beta_{j,d})\sigma_{\eta_{i,d}}^2\sigma_{\eta_{j,d}}^2$ . Thus

$$\tilde{\xi}_{i,j,d} = \text{Corr}(\epsilon_{k,d}\eta_{k,i,d}, \epsilon_{k,d}\eta_{k,j,d}) = \frac{(\omega_{i,j,d}\sigma_{\epsilon_d}^2 + 2\beta_{i,d}\beta_{j,d})}{\{(\omega_{i,i,d}\sigma_{\epsilon_d}^2 + 2\beta_{i,d}^2)(\omega_{j,j,d}\sigma_{\epsilon_d}^2 + 2\beta_{j,d}^2)\}^{1/2}}.$$

Note that, for  $i \in \tilde{\mathcal{H}}_0$ , we have  $\beta_{i,d} = O((\log p)^{-2-\tau})$  and so  $|\text{Corr}(V_i, V_j)| \leq \xi < 1$ , where  $\xi = \max\{\xi_1, \xi_2\} + \epsilon$  with  $\xi_d$  defined in (C2) and  $\epsilon < 1 - \max\{\xi_1, \xi_2\}$ , for  $i, j \in \tilde{\mathcal{H}}_{02}$ . Hence

$$\begin{aligned} & \frac{\sum_{i,j \in \tilde{\mathcal{H}}_{02}} \{\mathbb{P}(|V_i| \geq t, |V_j| \geq t) - \mathbb{P}(|V_i| \geq t)\mathbb{P}(|V_j| \geq t)\}}{p^2 G(t)} \\ & \leq C \frac{p^{1+\nu} t^{-2} \exp\{-t^2/(1+\xi)\}}{p^2 G(t)} \leq \frac{C}{p^{1-\nu} \{G(t)\}^{2\xi/(1+\xi)}}. \end{aligned} \quad (35)$$

It remains to consider the subset  $\tilde{\mathcal{H}}_{03}$ , in which  $V_i$  and  $V_j$  are weakly correlated. It is easy to check that  $\max_{i,j \in \tilde{\mathcal{H}}_{03}} \mathbf{P}(|V_i| \geq t, |V_j| \geq t) = (1 + O\{(\log p)^{-1-\gamma}\})G^2(t)$ . Hence,

$$\frac{\sum_{i,j \in \tilde{\mathcal{H}}_{03}} \{\mathbf{P}(|V_i| \geq t, |V_j| \geq t) - \mathbf{P}(|V_i| \geq t)\mathbf{P}(|V_j| \geq t)\}}{p_0^2 G^2(t)} = O\{(\log p)^{-1-\gamma}\}. \quad (36)$$

Equation (33) and the FDP result then follow by combining (34), (35), and (36), and the FDR result is also proved. ■

## References

- Anderson, T. W. (2003). *An Introduction To Multivariate Statistical Analysis*. Wiley-Interscience, 3rd ed, New York.
- Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.
- Cai, T., Liu, W., and Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Am. Statist. Assoc.*, 108(501):265–277.
- Cai, T. T. and Xia, Y. (2014). High-dimensional sparse manova. *Journal of Multivariate Analysis*, 131:174–196.
- D’Agostino Sr, R., Vasan, R., Pencina, M., Wolf, P., Cobain, M., Massaro, J., and Kannel, W. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117(6):743.

- Hibi, K., Ishigami, T., Kimura, K., Nakao, M., Iwamoto, T., Tamura, K., Nemoto, T., Shimizu, T., Mochida, Y., Ochiai, H., et al. (1997). Angiotensin-converting enzyme gene polymorphism adds risk for the severity of coronary atherosclerosis in smokers. *Hypertension*, 30(3):574–579.
- Humphries, S., Yiannakouris, N., and Talmud, P. (2008). Cardiovascular disease risk prediction using genetic information (gene scores): is it really informative? *Current Opinion in Lipidology*, 19(2):128.
- Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nature Reviews Genetics*, 6(4):287–298.
- Ikeda, S., Sasazuki, S., Natsukawa, S., Shaura, K., Koizumi, Y., Kasuga, Y., Ohnami, S., Sakamoto, H., Yoshida, T., Iwasaki, M., et al. (2008). Screening of 214 single nucleotide polymorphisms in 44 candidate cancer susceptibility genes: a case–control study on gastric and colorectal cancers in the japanese population. *The American journal of gastroenterology*, 103(6):1476–1487.
- Javanmard, A. and Montanari, A. (2013). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory.
- Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.
- Kannel, W., Feinleib, M., McNamara, P., Garrison, R., and Castelli, W. (1979). An investigation of coronary heart disease in families The Framingham Offspring Study. *American Journal of Epidemiology*, 110(3):281–290.

- Liu, L., Zhong, R., Wei, S., Xiang, H., Chen, J., Xie, D., Yin, J., Zou, L., Sun, J., Chen, W., et al. (2013). The leptin gene family and colorectal cancer: interaction with smoking behavior and family history of cancer. *PloS one*, 8(4):e60777.
- Liu, W. and Luo, S. (2014). Hypothesis testing for high-dimensional regression models. Technical report.
- Lloyd-Jones, D., Wilson, P., Larson, M., Beiser, A., Leip, E., D'Agostino, R., and Levy, D. (2004). Framingham risk score and prediction of lifetime risk for coronary heart disease\*  
1. *The American Journal of Cardiology*, 94(1):20–24.
- Matsouaka, R. A., Li, J., and Cai, T. (2014). Evaluating marker-guided treatment selection strategies. *Biometrics*, 70(3):489–499.
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369.
- Paynter, N., Chasman, D., Buring, J., Shiffman, D., Cook, N., and Ridker, P. (2009). Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21. 3. *Annals of internal medicine*, 150(2):65.
- Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.
- Ridker, P., Buring, J., Rifai, N., and Cook, N. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score. *Journal of American Medical Association*, 297(6):611.

- Ross, R. (1999). Atherosclerosis is an inflammatory disease. *American Heart Journal*, 138(5):S419–S420.
- Sayed-Tabatabaei, F., Schut, A., Hofman, A., Bertoli-Avella, A., Vergeer, J., Witteman, J., and van Duijn, C. (2004). A study of gene–environment interaction on the gene for angiotensin converting enzyme: a combined functional and population based approach. *Journal of medical genetics*, 41(2):99–103.
- Schut, A. F., Sayed-Tabatabaei, F. A., Witteman, J. C., Avella, A. M., Vergeer, J. M., Pols, H. A., Hofman, A., Deinum, J., and van Duijn, C. M. (2004). Smoking-dependent effects of the angiotensin-converting enzyme gene insertion/deletion polymorphism on blood pressure. *Journal of hypertension*, 22(2):313–319.
- Stephens, J. W., Bain, S. C., and Humphries, S. E. (2008). Gene–environment interaction and oxidative stress in cardiovascular disease. *Atherosclerosis*, 200(2):229–238.
- Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351.
- Wilson, P., D’Agostino, R., Levy, D., Belanger, A., Silbershatz, H., and Kannel, W. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837.
- Xia, Y., Cai, T., and Cai, T. T. (2015). Testing differential network with applications to

detecting gene by gene interactions. *Biometrika*, 102:247–266.

Zaitsev, A. Y. (1987). On the gaussian approximation of convolutions under multidimensional analogues of sn bernstein's inequality conditions. *Probab. Theory Rel.*, 74(4):535–566.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242.