Large-Scale Simultaneous Testing of Cross-Covariance Matrix with Applications to PheWAS

Tianxi Cai\textsuperscript{1}, T. Tony Cai\textsuperscript{2}, Katherine Liao\textsuperscript{3}, and Weidong Liu\textsuperscript{4}

Abstract

The electronic medical record (EMR) system linked with biorepository has led to the availability of detailed patient level phenotypic data along with biological and genetic marker information. Such type of data expansion enables us to perform phenome-wide association studies (PheWAS) to examine the relationship between a set of genomic markers and a wide range of phenotypic outcomes by integrating multiple sources of information. Motivated by such PheWAS applications, we consider simultaneous testing of columns of high-dimensional cross-covariance matrices and develop a multiple testing procedure.

Theoretical results show that under mild regularity conditions the proposed testing procedure maintains a desired false discovery rate (FDR) and false discovery proportion (FDP). We also provide results on the magnitudes of the signals that can be detected with high power. Simulation studies demonstrate that the proposed methods can be substantially more powerful than existing FDR controlling procedures in the presence of correlation. We applied the proposed procedures to a PheWAS of two auto-immune genetic markers using a cohort of Rheumatoid Arthritis (RA) constructed from the EMR of the Partners Healthcare.

\textsuperscript{1}Department of Biostatistics, Harvard School of Public Health, Harvard University. The research of Tianxi Cai was supported in part by NIH Grants R01 GM079330 and U54 H6007963.

\textsuperscript{2}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.

\textsuperscript{3}Brigham and Women’s Hospital.

\textsuperscript{4}Department of Mathematics, Institute of Natural Sciences and MOE-LSC, Shanghai Jiao Tong University, Shanghai, China. Weidong Liu’s research was supported by NSFC Grant No.11201298 and No.11322107, Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, Shanghai Pujiang Program, Foundation for the Author of National Excellent Doctoral Dissertation of PR China and Program for New Century Excellent Talents in University.
Keywords: Correlation, false discovery rate, large-scale multiple testing, multiple responses, PheWAS.

1 Introduction

Simultaneously assessing the associations among a large number of variables is an important step in discovery research for a wide range of applications. Large-scale testing tools are frequently needed for important biomedical studies such as genome-wide association studies (GWAS) (Bush and Moore, 2012); high throughput gene expression profiling studies of microRNAs and mRNAs (Katagiri and Glazebrook, 2009; Duggan et al., 1999; Nelson et al., 2004); and phenome-wide association studies (PheWAS) of a large number of disease phenotypes (Denny et al., 2010). A critical step in performing a large-scale multiple testing is to control the false discovery rate. Standard FDR control procedures, such as the Benjamini-Hochberg (B-H) procedure (Benjamini and Hochberg, 1995), which are typically built under the independence assumption, would fail to provide desired error controls in the presence of strong correlation, particularly when the signals are sparse as in the PheWAS setting. Various alternative methods, including the Benjamini-Yekutieli (B-Y) procedure (Benjamini and Yekutieli, 2001), have been developed to allow for the general dependency by paying a logarithmic penalty on the FDR. Unfortunately, these methods tend to be overly conservative (see simulation results in Section 4). Furthermore, existing methods largely focus on the association between a single outcome variable with a large number of candidate covariates (Efron, 2004, 2007; Owen, 2005; Fan et al., 2012; Fan and Han, 2013, e.g.). However, multiple related outcomes are often of interest in many applications such as the PheWAS. These aforementioned methods are not directly applicable to such PheWAS settings.

In this paper, we translate the problem of assessing the association between a large number of variables and multiple outcomes into the statistical problem of simultaneous testing of columns of high-dimensional cross-covariance matrices; and develop a general framework for
such a testing problem without requiring strong assumptions on the correlation structure. We first discuss the motivating application of PheWAS with multiple outcomes and then present the general framework for large-scale multiple testing of columns of high-dimensional cross-covariance matrices.

1.1 PheWAS of Multiple Outcomes

An important goal of GWAS is to explore genetic susceptibility to complex diseases. These studies have led to identification of many genomic regions as putatively harboring disease susceptibility alleles for a wide range of disorders (Billings and Florez, 2010; Gudmundsson et al., 2007; Hunter et al., 2007; Zanke et al., 2007; Stahl et al., 2010; Thorgeirsson et al., 2008, e.g). A common limitation of GWAS is the focus on a pre-defined and limited phenotypic domain. A complementary approach to GWAS is the PheWAS, in which the association between genomic markers and a diverse range of phenotypes are investigated (Denny et al., 2010). PheWAS enables the discovery of genetic markers with pleiotropic effects on multiple traits, and thus may provide a broader view of the relationship between genetic variation and networks of phenotypes (Pendergrass et al., 2013; Shameer et al., 2014; Hall et al., 2014). This is highly desirable since recent genetic studies have suggested that many genetic loci appear to harbor variants associated with multiple traits (Solovieff et al., 2013).

PheWAS has only recently become feasible due to the wide availability of the electronic medical record (EMR) systems linked with biorepositories. The EMR system provides detailed longitudinal patient level phenotypic data including codified variables (e.g. ICD9 codes, CPT results, lab results, medication prescription) and narrative variables, such as the number of mentions for clinical symptoms, extracted from clinical notes via natural language processing (NLP). Such a platform enables us to extract a wide range of phenotype information on each patient. In addition to compiling phenotype data, research institutions such as the Partners HeathCare also link the EMR to bio-specimen repositories that coordinate and collect discarded blood samples for biomarker and genetic research. Through anonymization of data,
patient level phenotype information linked with genotype information becomes available for extensive clinical research. For example, a rheumatoid arthritis (RA) cohort along with a matched control cohort have been established for discovery research (Liao et al., 2010). Genomic and biological markers have been measured for 1837 RA cases and 1574 non-RA controls whose phenotype information are available from the EMR (Kurreeman et al., 2011; Liao et al., 2012, 2013). Such data enables us to conduct PheWAS to rigorously study genome-phenome association networks and examine how such associations may differ across different patient populations.

In contrast to GWAS, the large number of phenotypic variables in PheWAS are often substantially correlated. When examining the association between a set of genomic or biological markers and these correlated phenotypes, the test statistics assessing all the associations of interest could have rather complicated correlation structure. This would contribute significantly to the difficulty of the multiple testing problem. Due to the significant correlations, statistical methods for simultaneously assessing a large number of correlated associations while controlling for a desired overall FDR remains elusive.

1.2 Multiple Testing of Columns of Cross-Covariance Matrices

In this paper we propose to translate the PheWAS problem into simultaneous testing of columns of high-dimensional cross-covariance matrices and develop a general framework for such a testing problem without requiring strong assumptions on the correlation structures. More specifically, let \( Y = (Y_1, \ldots, Y_d)' \) be a \( d \)-dimensional outcome vector and \( X = (X_1, \ldots, X_p)' \) be a \( p \)-dimensional phenotype vector. In the PheWAS setting, \( Y \) may be a vector of genomic markers and \( X \) represents all phenotypic disease conditions of interest. For \( 1 \leq i \leq p \), define the cross-covariance vector between \( Y \) and \( X_i \) by \( \sigma_i = (\sigma_1^{(i)}, \ldots, \sigma_d^{(i)})' \), where \( \sigma_j^{(i)} = \text{Cov}(Y_j, X_i) \). In other words, \( \sigma_i \) is the \( i \)th column of the cross-covariance matrix

\[
\Sigma_{YX} \equiv \text{Cov}(Y, X) = [\sigma_1, \ldots, \sigma_p]
\]
between the markers $\mathbf{Y}$ and the outcomes $\mathbf{X}$. We are interested in simultaneously testing the collection of $p$ hypotheses

$$H_{0i} : \sigma_i = 0 \quad \text{versus} \quad H_{1i} : \sigma_i \neq 0, \quad 1 \leq i \leq p$$

while controlling the overall error rate. The $i$th hypothesis examines whether the marker vector $\mathbf{Y}$ is associated with the $i$th disease condition. We are particularly interested in the setting where the number of the true alternative hypotheses is relatively small as in the case of the applications in PheWAS.

Large-scale multiple testing of the columns of a high-dimensional cross-covariance matrix is technically difficult due to the complex entrywise dependence structures. There are two main challenges, one is the construction of suitable test statistics for individual hypotheses and establishing the null distribution of these test statistics; another is in the construction of a good procedure to account for the multiplicity of the tests so that the overall FDR is controlled under dependency.

We propose in this paper a large-scale association testing procedure for $\{H_{0i}, i = 1, \ldots, p\}$ with theoretical guarantees for FDR control under flexible correlation structures. We consider a test statistic mimicking the Hotelling’s $T^2$ and demonstrate that in the presence of correlation induced by the observed data, the null distribution of the statistic can be well approximated by a $\chi^2$ within an appropriate range. Furthermore, we develop a multiple testing procedure by thresholding the test statistic and demonstrate that the thresholding rule based on the asymptotic null distribution controls both the FDR and the false discovery proportion (FDP).

No existing multiple testing procedures can be applied to the PheWAS setting and provide accurate FDR or FDP control. Available methods and theoretical results largely focus on $z$-tests or $t$-tests for individual hypotheses and commonly require the knowledge of the null distribution. For example, the methods of Fan et al. (2012) and Fan and Han (2013) considered multiple testing for normal means and required the covariance matrix to be known or well estimated. Efron (2004, 2007) developed FDR controlling procedures for multiple $t$-tests. The cross-covariance testing considered here is far more complicated due to the involvement
of multiple outcomes and the unknown and complicated dependence structure of the entries of
the cross-covariance matrix. The PheWAS setting necessitates the use of test statistics only
with approximate null distributions. No existing theoretical results can ensure the control
of FDR and FDP under such complex dependence structure without assuming that the null
distribution of the test statistics is known.

The choice of the null distribution may substantially affect the simultaneous inference
procedure (Efron, 2004; Liu and Shao, 2014). In fact, Liu and Shao (2014) showed that in
multiple t-tests with the dimension being much larger than the sample size, if the p-values
are calculated from the asymptotic distribution, such as normal distribution or t-distribution,
then the FDR and FDP of the B-H method can converge to one. It is thus critical to justify the
use of asymptotic null distribution for proper FDR control, which is non-trivial in the current
setting. Under the assumption that \( p_1 \geq cp \) for some \( c > 0 \), the method in Storey et al. (2004)
icorporates the estimation of the proportion of the true nulls into the B-H method, where \( p_1 \)
is the number of true signals. In the PheWAS setting, the signal is sparse and \( p_1 \) is of order
\( o(p) \), which leads to the equivalence of the Storey et al. (2004) and B-H procedures. Due to
the sparsity, the strong and complex dependence between the test statistics as well as the
need to estimate the null distribution, neither of the Storey et al. (2004) and B-H procedures
is able to control FDR or FDP in our setting.

Owen (2005) investigated the variance of the number of falsely rejected hypotheses under
the assumption that all \( \rho_j = 0, 1 \leq j \leq p \), where \( \rho_j \) is the correlation coefficient between
the univariate response \( Y \) and the covariates \( X_i \). So his work is related to the control of
false discovery number (FDN) which is different from the control of FDR. In addition, Owen
(2005) used the sample correlation coefficients as the test statistics so that the dependence
structure between the test statistics can be calculated explicitly. Our test statistics are more
complicated. It is difficult to calculate the correlation between \( T_i \) and \( T_j \). Hence, the results
in Owen (2005) are not applicable in our setting.
1.3 Structure of the Paper

The rest of the paper is organized as follows. The proposed methodology detailing the test statistics as well as the FDR control procedure is introduced in Section 2. A theoretical analysis of the multiple testing procedure, presented in Section 3 and proved in Section 6, shows that the proposed method controls the FDR and the false discovery proportion (FDP) at a desired nominal level asymptotically. We discuss how the proposed procedure relates to and differs from existing FDR controlling methods in Section 3.3. Results from simulation studies are given in Section 4 where we also apply our proposed procedures to the aforementioned RA cohort to comprehensively evaluate how two important genetic markers for auto-immune diseases may contribute to comorbidities of RA. Section 5 discusses a few related issues.

2 Methodology

We develop the proposed multiple testing procedure in detail in this section. Suppose data for analysis consists of \( n \) independent and identically distributed copies of \((\mathbf{Y}, \mathbf{X})\), denoted by \( \mathcal{D} = \{(\mathbf{Y}_k', \mathbf{X}_k')', 1 \leq k \leq n\} \), where \( \mathbf{Y}_k = (Y_{k1}, \ldots, Y_{kd})' \) and \( \mathbf{X}_k = (X_{k1}, \ldots, X_{kp})' \). We wish to simultaneously test the hypotheses \( \{H_{0i}, i = 1, \ldots, p\} \) given in (1.1) based on the observed random samples in \( \mathcal{D} \). Note that here we are interested in multiple testing of the columns, not individual entries, of the cross-covariance matrix \( \Sigma_{YX} \).

We first introduce the test statistics for testing the individual hypothesis \( H_{0i} : \sigma_i = 0 \). Let
\[
\hat{Z}_{ki} = (\mathbf{Y}_k - \bar{\mathbf{Y}})(X_{ki} - \bar{X}_i)
\]
where \( \bar{\mathbf{Y}} = n^{-1} \sum_{k=1}^{n} \mathbf{Y}_k \) and \( \bar{X}_i = n^{-1} \sum_{k=1}^{n} X_{ki} \) for \( 1 \leq i \leq p \) and \( 1 \leq k \leq n \). Then
\[
\hat{\sigma}_i = n^{-1} \sum_{k=1}^{n} \hat{Z}_{ki}
\]
is a consistent and asymptotic unbiased estimator of \( \sigma_i \), the covariance vector between \( \mathbf{Y} \) and \( X_i \). A test for \( H_{0i} \) may be constructed based on the observed \( n^{1/2} \hat{\sigma}_i \) along with its covariance
matrix, which can be approximated using the sample covariance matrix of \( \{ \hat{Z}_{ki}, k = 1, ..., n \} \),
\[
\Sigma_{Zi} = n^{-1} \sum_{k=1}^{n} (\hat{Z}_{ki} - \hat{\sigma}_i)(\hat{Z}_{ki} - \hat{\sigma}_i)'.
\]

Inspired by the Hotelling’s \( T^2 \) statistic for testing a multivariate normal mean vector, we propose to test \( H_{0i} \) using the test statistic
\[
T_i = n(\hat{\sigma}_i)'(\Sigma_{Zi})^{-1}\hat{\sigma}_i.
\] (2.2)

Let \( \mathcal{H}_0 = \{ i : \sigma_i = 0 \} \) be the index set of all null hypotheses, \( \mathcal{H}_1 = \{ i : \sigma_i \neq 0 \} \), and \( p_0 = \text{Card}(\mathcal{H}_0) \). As indicated in Lemma 6.1 in Section 6, under the conditions of Theorem 3.1,
\[
\max_{i \in \mathcal{H}_0} \left| \frac{\mathbb{P}(T_i \geq t)}{G(t)} - 1 \right| \to 0
\]
uniformly in \( t \in [0, a_p] \), where \( G(t) = \mathbb{P}(\chi^2_d \geq t) \) and
\[
a_p = 2 \log p + (d - 1) \log \log p.
\] (2.3)
Hence, when \( n \) is large, the chi-squared distribution provides an accurate approximation to the null distribution of \( T_i \) in the range \( [0, a_p] \).

We next propose an FDR controlled multiple testing procedure by thresholding the test statistics \( \{ T_i, i = 1, ..., p \} \). Specifically, let \( t > 0 \) be a rejection threshold so that \( H_{0i} \) is rejected if and only if \( T_i \geq t \). For any given threshold \( t > 0 \), the false discovery proportion (FDP) based on the random sample \( \mathcal{D} \) is
\[
\text{FDP}(t) = \frac{\sum_{i \in \mathcal{H}_0} I(T_i \geq t)}{\max\{\sum_{i=1}^{p} I(T_i \geq t), 1\}}.
\] (2.4)
To maximize the power of the test or equivalently the rejection rate among \( \mathcal{H}_1 \) while maintaining an FDP level of \( \alpha \), the optimal threshold \( t \) is then
\[
\hat{t}_0 = \inf\{ t : \text{FDP}(t) \leq \alpha \}.
\]
Since the denominator of $FDP(t)$ in (2.4) is observable, the key to empirically control the FDP is to find a good estimate of the numerator $\sum_{i \in \mathcal{H}_0} I(T_i \geq t)$. We will show that $\sum_{i \in \mathcal{H}_0} I(T_i \geq t)$ can be approximated by $p_0 G(t)$. More precisely,

$$\sup_{0 \leq t \leq b_p} \left| \frac{\sum_{i \in \mathcal{H}_0} I(T_i \geq t)}{p_0 G(t)} - 1 \right| \to 0 \quad \text{in probability},$$

where

$$b_p = 2 \log p + (d - 3) \log(\log p).$$

The range $[0, b_p]$ is nearly optimal. When $t \geq b_p + \log(\log p)$, $G(t)$ may not consistently estimate $p_0^{-1} \sum_{i \in \mathcal{H}_0} I(T_i \geq t)$. Note that $p_0$ can be further estimated by $p$ due to the sparsity in the number of alternative hypotheses in many real data applications. Based on the above analysis, we propose the following multiple testing procedure for simultaneously testing the hypotheses in (1.1):

**FDR control procedure.** Calculate $T_i$ in (2.2) and for any given nominal FDR level $\alpha \in (0, 1)$, reject $H_{0i}$ whenever $T_i \geq \hat{t}$, where

$$\hat{t} = \inf \left\{ 0 \leq t \leq b_p : G(t) \leq \frac{\alpha \max \{ \sum_{1 \leq i \leq p} I(T_i \geq t), 1 \} }{p} \right\}$$

(2.5)

if the righthand side of (2.5) exists; and $\hat{t} = a_p$ otherwise.

Note that, when $\hat{t}$ in (2.5) does not exist, the FDR control procedure simply thresholds the test statistics at the value $a_p$ given in (2.3). Unlike the conventional B-H procedure, the proposed thresholding rule enables the corresponding multiple testing procedure to control both the FDR and the FDP. As shown numerically in Section 4, if the number of alternatives $p_1$ is small, then $\hat{t}$ in (2.5) may not exist and the true FDP of the B-H method can be much higher than $\alpha$. In this case, the proposed method uses $a_p$ as the thresholding level adaptively and is able to control the FDP efficiently. See more discussion in Section 5.
3 Theoretical results

We investigate in this section the theoretical properties of the proposed multiple testing procedure and discuss its relation to existing methods. Numerical performance of the procedure will be studied in Section 4. Let

$$FDP = \frac{\sum_{i \in H_0} I(T_i \geq \hat{t})}{\max\{\sum_{1 \leq i \leq p} I(T_i \geq \hat{t}), 1\}}$$

and

$$FDR = \mathbb{E}(FDP).$$

We first state some conditions on the correlation structure. Let

$$Z_i = (Y - \mu_Y)(X_i - \mu_i),$$

$$\Sigma_{Z_i} = \text{Cov}(Z_i)$$ and

$$\xi_i = \Sigma_{Z_i}^{-1/2}Z_i.$$ As in canonical correlation analysis, define the maximum correlation coefficients

$$\rho_{ij}^* = \max_{\|a\|=1,\|b\|=1} |\text{corr}(a^T \xi_i, b^T \xi_j)|,$$

where $\|\cdot\|$ denotes the Euclidean norm. $\rho_{ij}^*$ characterizes the dependence between $T_i$ and $T_j$.

Define

$$\mathcal{B}(\delta) = \{(i, j): i \in H_0, j \in H_0, \rho_{ij}^* \geq \delta, i \neq j\}$$ with $\delta \in (0, 1)$.

The set $\mathcal{A}(\varepsilon) = \mathcal{B}\{(\log p)^{-2-\varepsilon}\}$ includes the pairs $\xi_i$ and $\xi_j$ that are strongly correlated for $i, j \in H_0$. The first condition requires the number of strongly correlated pairs to be not too large.

(C1). There exist some $\varepsilon > 0$ and some $\delta > 0$ such that

$$\sum_{(i, j) \in \mathcal{A}(\varepsilon)} \frac{2^{2\rho_{ij}^*}}{p^{1+\rho_{ij}^*} + \delta} = O\{p^2(\log p)^{-2}\}. \quad (3.6)$$

Remark 1. When $(Y', X')$ has the elliptically contoured distribution, it is easy to see that $\rho_{ij}^* \leq |\rho_{ij}|$ for $i, j \in H_0$, where $(\rho_{ij})_{1 \leq i, j \leq p}$ is the correlation matrix of $X$. In this case, (3.6) is reduced to

$$\sum_{(i, j) \in \mathcal{A}_1(\varepsilon)} \frac{2|\rho_{ij}|}{p^{1+|\rho_{ij}|} + \delta} = O\{p^2(\log p)^{-2}\}, \quad (3.7)$$

where $\mathcal{A}_1(\varepsilon)$ is defined as $\mathcal{A}(\varepsilon)$ with $\rho_{ij}^*$ being replaced by $|\rho_{ij}|$. Condition (3.6) holds if

Card$\{\mathcal{B}(\delta)\} = O(p^\rho)$ for any $0 < \delta < 1$ and some $\rho < 2/(1+\delta)$ and Card$\{\mathcal{A}(\varepsilon)\} = O(p^\rho)$ for
some $\rho < 2$ and $\varepsilon > 0$. The correlation condition (3.6) can be further weakened if the number of signals becomes larger and can be easily satisfied in many applications. For example, in the scale free network, only a few variables are associated with many variables and most of variables are only associated with a few others. In PheWAS settings, most diseases only have a few co-morbidities and hence these assumptions for Card\{$B(\delta)$\} and Card\{$A(\varepsilon)$\} are reasonable.

**Remark 2.** Note that when $i \in H_0$ and $j \in H_0$, we have $Y$ is uncorrelated with $(X_i, X_j)$. If we assume that $Y$ is independent from $(X_i, X_j)$ for $i \in H_0$ and $j \in H_0$, then $\rho^*_{ij} = |\rho_{ij}|$. In this case, (C1) is reduced to a correlation condition on $X$ which is quite natural.

### 3.1 FDR and FDP Control

The next two theorems suggest that the proposed multiple testing procedure controls FDR and FDP asymptotically. Furthermore, the actual FDR and FDP converge to $\alpha p_0/p$ asymptotically when the number of non-trivial signals,

$$m_1(c) = \text{Card}\{i : 1 \leq i \leq p, \ \sigma_i^2 \Sigma^{-1} \sigma_i \geq c (\log p)/n\},$$

is sufficiently large for some $c$. Let $\lambda_{i1}$ and $\lambda_{id}$ be the largest and smallest eigenvalues of $\Sigma_{Zi}$, respectively.

**Theorem 3.1.** Suppose that $p \leq n^\beta$ for some $\beta > 0$, $\mathbb{E}\|Y - \mu_Y\|^{8\beta + 4 + \varepsilon} \leq K$, $\max_{1 \leq i \leq p} \mathbb{E}|X_i - \mu_i|^{8\beta + 4 + \varepsilon} \leq K$ and $c_1 \leq \lambda_{id} \leq \lambda_{i1} \leq c_2$ for all $1 \leq i \leq p$ and some $\varepsilon > 0$, $K > 0$, $c_1 > 0$ and $c_2 > 0$. Under (C1), we have for any $\varepsilon > 0$,

$$\lim_{(n,p) \to \infty} \mathbb{P}(\text{FDP} \leq \alpha + \varepsilon) = 1 \quad \text{and} \quad \lim\sup_{(n,p) \to \infty} \text{FDR} \leq \alpha.$$

**Theorem 3.2.** Suppose the conditions in Theorem 3.1 hold. Assume that

$$m_1(c) \geq \log p \quad \text{for some } c > 2. \quad (3.8)$$

Then we have

$$\lim_{(n,p) \to \infty} \frac{\text{FDR}}{\alpha p_0/p} \to 1 \quad \text{and} \quad \lim_{(n,p) \to \infty} \frac{\text{FDP}}{\alpha p_0/p} \to 1 \quad \text{in probability as } (n,p) \to \infty. \quad (3.9)$$
When additional assumptions are imposed on the sparsity and strength of the signals, the condition on the correlation matrix $R$ can also be further relaxed. For example, under

$$m_1(c) \geq p^\theta \quad \text{for some } 0 < \theta < 1 \text{ and } c > 2(1-\theta),$$

(3.10)

the condition (C1) can be weakened as follows.

\text{(C1').} Suppose that

$$\sum_{(i,j) \in \mathcal{A}(\varepsilon)} p^{2(1-\theta)R_{ij}^* + \delta} = O\{p^2/(\log p)^2\}, \quad \text{for some } \varepsilon > 0 \text{ and } \delta > 0.$$

Under these alternative conditions, we have similar results on the FDR and FDP control:

\textbf{Theorem 3.3.} Suppose that $p \leq n^\beta$ for some $\beta > 0$, $\mathbb{E}\|Y - \mu_Y\|^{8\beta+4+\epsilon} \leq K$, $\max_{1 \leq i \leq p} \mathbb{E}|X_i - \mu_i|^{8\beta+4+\epsilon} \leq K$ and $c_1 \leq \lambda_{id} \leq \lambda_i \leq c_2$ for all $1 \leq i \leq p$ and some $\epsilon > 0$, $K > 0$, $c_1 > 0$ and $c_2 > 0$. Under (C1') and (3.10), we have (3.9) holds.

When the number of non-trivial signals increases to the magnitude of $m_1(c) = p/(\log p)^\lambda$ for some $\lambda > 0$ and $c > 0$, we only require $\text{Card}\{\mathcal{A}(\varepsilon)\} \leq p^{2-\delta}$ for some $\delta > 0$. The number of pairs $(\xi_i, \xi_j)$ with non-trivial correlations can be as large as $p^{2-\delta}$.

When $d = 1$, Owen (2005) considered multiple tests assessing the association between $Y$ and $X_i$ for $i = 1, ..., p$ based on the sample correlation coefficients. He developed the variance of the number of falsely rejected hypotheses $\sum_{i \in \mathcal{H}_0} I\{|\hat{\rho}_i| \geq t\}$ for a fixed $t > 0$. He showed that the variance can be affected significantly by the correlations between $X_i$, $1 \leq i \leq p$. However, no FDR controlling procedures were provided. When considering FDR control, our results obtained under different sets of conditions suggest that, the effect of the correlations from $X$ is related to the number of signals. As $\theta$ in (3.10) increases, the total number of signals increases and the condition on the correlation becomes weaker.

\subsection*{3.2 Power Properties}

We now consider the power of the procedure. Define the power by

$$\widehat{PO} = \frac{\sum_{i \in \mathcal{H}_1} I(T_i \geq \hat{i})}{\text{Card}(\mathcal{H}_1)}.$$
Suppose (3.10) holds and the magnitude of all signals in \( \mathcal{H}_1 \) satisfies

\[
\sigma_i' \Sigma_{Z_i}^{-1} \sigma_i \geq \frac{c \log p}{n} \quad \text{for} \quad i \in \mathcal{H}_1.
\] (3.11)

**Theorem 3.4.** Suppose that \( p \leq n^\beta \) for some \( \beta > 0 \), \( \mathbb{E}\| \mathbf{Y} - \mathbf{\mu}_Y \|^8 \beta + 4 + \epsilon \leq K \), \( \max_{1 \leq i \leq p} \mathbb{E}|X_i - \mu_i|^{8 \beta + 4 + \epsilon} \leq K \) and \( c_1 \leq \lambda_{id} \leq \lambda_{i1} \leq c_2 \) for all \( 1 \leq i \leq p \) and some \( \epsilon > 0 \), \( K > 0 \), \( c_1 > 0 \) and \( c_2 > 0 \). Assume that (3.10) and (3.11) hold. Then we have

\[
\hat{PO} \rightarrow 1 \quad \text{in probability}.
\]

Theorem 3.4 shows that the proposed multiple testing procedure has overwhelming power in detecting the signals with magnitudes satisfy (3.11). The constant factor \( c \) in (3.11) can not be replaced by \( o(1) \). Otherwise, it is not even possible to detect the global signals \( (\sigma_i \neq 0 \text{ for some } i) \); see Donoho and Jin (2004) in the setting of signal detection under a sparse normal mixture model.

### 3.3 Relation with the existing FDR control methods

Under the PheWAS setting with multi-dimensional \( \mathbf{Y} \), controlling the FDR or FDP is more complicated due to the non-trivial dependence among the test statistics, the need to estimate the null distribution of the test statistics or \( p \)-values and the sparsity of the signals. As mentioned in the introduction, most existing FDR control procedures require the exact \( p \)-values and some well-known procedures for the dependent cases focus on the multiple \( t \) tests with stringent conditions. These existing methods cannot be applied to the present setting. When the number of the true signals is fixed as \( p \rightarrow \infty \), Proposition 2.1 in Liu and Shao (2014) shows that the B-H procedure is unable to control the FDP. The PheWAS setting is even more challenging with the additional complication of strong and complex dependence among the test statistics. As such both the B-H and Storey et al. (2004) procedures fail to control the FDR or FDP, as demonstrated in the simulation studies.

The proposed procedure differs from the B-H procedure in the additional thresholding step, which is critical in controlling the FDR and FDP. Specifically, the B-H procedure rejects
any hypotheses with the test statistics exceeding \( \hat{t}_{BH} \), where

\[
\hat{t}_{BH} = \inf \left\{ t \geq 0 : G(t) \leq \frac{\alpha \max \{ \sum_{1 \leq i \leq p} I(T_i \geq t), 1 \}}{p} \right\}.
\]

When \( \hat{t}_{BH} \leq b_p \), our procedure coincides with the B-H procedure; If \( \hat{t}_{BH} > b_p \), then our procedure instead rejects hypotheses with test statistics exceeding \( a_p \). This threshold is necessary since the asymptotic null distribution \( G(t) = \mathbb{P}(\chi_d^2 \geq t) \) may not approximate \( p^{-1} \sum_{i \in H_0} I(T_i \geq t) \) sufficiently well for large \( t \) in that \( p^{-1} \sum_{i \in H_0} I(T_i \geq t)/G(t) \not\to 1 \) when \( t \geq b_p + \log(\log p) \). As shown in our simulation studies, the probability of \( \hat{t}_{BH} > b_p \) approaches to 1 under extreme sparsity and is non-trivial under moderate sparsity. This also sheds light on why the proposed procedure controls the FDR and FDP while the B-H procedure fails under the PheWAS setting. Our thresholding rule based procedure adaptively controls the FDR and FDP without prior knowledge of the degree of sparsity.

## 4 Numerical Results

We investigate the numerical performance of the proposed multiple testing procedure in this section through simulation studies. Numerical comparison with alternative methods is given. The proposed procedure is also applied to an RA cohort to comprehensively evaluate how two important genetic markers for auto-immune diseases may contribute to comorbidities of RA.

### 4.1 Simulation Studies

We performed extensive simulations to examine the performance of the procedure in finite sample with practical sample sizes and \( p \). We let \( d = 4 \) for the outcome \( Y \) and considered \( p = 500, 1000, 2000 \) and 4000 for \( X \). For each setting, we consider \( n = 200 \) and 400. All results are summarized based on 200 simulated datasets for each configuration.

We generate the entire data vector \( \mathbf{W} := (Y_1, \ldots, Y_4, X_1, \ldots, X_p)' \) from \( \mathbf{W} = \Sigma^{1/2} \mathbf{\varepsilon} \), where four settings of \( \Sigma \) as described below were chosen to reflect different correlation structures,
and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{p+4})'$ are independent and identically distributed. The distribution of $\varepsilon_i$ is taken to be either (i) the standard normal $N(0, 1)$; or (ii) the exponential with mean 1. As our test statistics are invariant to the variances, we take the diagonal entries of $\Sigma$ to be 1. Let $\Sigma_1 = (\sigma_{ij1}) \in \mathbb{R}^{(p_1+4) \times (p_1+4)}$ and $\Sigma_2 = (\sigma_{ij2}) \in \mathbb{R}^{(p-p_1) \times (p-p_1)}$. In the following, we describe the four choices of $\Sigma$. In Models 1-3, we let $\Sigma = \text{diag}(\Sigma_1, \Sigma_2)$ and $\sigma_{ij1} = (2 \log p/n)^{I(i\neq j)/2}$.

**Model 1.** $\sigma_{ij2} = 0.8|j-i|$;

**Model 2.** $\Sigma_2 = A + (|\lambda_{\min}(A)| + 0.05)I$ with $A = (0.5\varepsilon_{ij})$, where $\varepsilon_{ij}$ are independently generated from Bernoulli(0.1). The matrix is standardized to have unit diagonals.

**Model 3.** $\Sigma_2 = \text{diag}(D_1, \ldots, D_m, I)$, where $m = [(p-p_1)/10]$, $D_k \in \mathbb{R}^{10 \times 10}$, $1 \leq k \leq m$. $I$ is the identity matrix. All of the off-diagonal entries of $D_k$ are taken to be 0.5.

**Model 4.** Let $\sigma_{ij1} = (4 \log p/n)^{I(i\neq j)/2}$ and $\Sigma_2$ be in Model 3. Let $A_1 = \text{diag}(\Sigma_1, \Sigma_2)$ and $A_2 = (a_{ij2}) \in \mathbb{R}^{(p+4) \times (p+4)}$ with $a_{ij2} = a_{ji2} = 0.5^{i-p_1}j$ for $5 \leq i \leq 4 + p_1$ and $5 + p_1 \leq j \leq p$. Other $a_{ij2}$ are set to zero. Let $A_3 = A_1 + A_2$ and the covariance matrix $\Sigma = (A_3 + (|\lambda_{\min}(A_3)| + 0.05)I)/[1 + (|\lambda_{\min}(A_3)| + 0.05)]$.

Under these models, only $X_1, X_2, \ldots, X_{p_1}$ are correlated with $Y$.

We first examine the probability that $\hat{t}$ in (2.5) exists under different settings, which reflects the degree to which our procedure differs from the B-H procedure. Figure 1 summarizes the frequency that $\hat{t}$ in (2.5) exists for Models 1-4 when $\alpha = 0.1$, $n = 200$ and $p = 1000$. The probability that $\hat{t}$ does not exist approaches to 1 when the signals are extremely sparse, is non-trivial under moderate sparsity and gradually decreases to near 0 as $p_1$ further increases. These results demonstrate that the proposed procedure differs substantially from the B-H procedure under sparse settings.

We next investigate the performance of the proposed procedure in controlling for the FDR in finite samples. Table 1 presents the empirical FDR for $p_1 = 46$, $\alpha = 0.1, 0.2$ and 0.3 with various choices of $p, n$ under Models 1-4. The results show that the proposed procedures are able to maintain the desired FDR levels across all settings.
Figure 1: The frequency that \( \hat{t} \) exists. \( \alpha = 0.1, n = 200 \) and \( p = 1000 \). The x-axis denotes the values of \( p_1 \) and the y-axis denotes the values of frequency.
<table>
<thead>
<tr>
<th>Model 1</th>
<th>N(0,1)</th>
<th>Exp(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 200</td>
<td>n = 400</td>
</tr>
<tr>
<td>100α</td>
<td>10 20 30</td>
<td>10 20 30</td>
</tr>
<tr>
<td>p = 500</td>
<td>11.4 21.0 30.6</td>
<td>10.9 20.3 29.4</td>
</tr>
<tr>
<td>p = 1000</td>
<td>12.9 23.0 32.6</td>
<td>11.8 21.8 31.2</td>
</tr>
<tr>
<td>p = 2000</td>
<td>12.1 22.2 32.6</td>
<td>12.0 22.1 31.6</td>
</tr>
<tr>
<td>p = 4000</td>
<td>12.7 23.1 33.9</td>
<td>11.8 21.5 31.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 2</th>
<th>N(0,1)</th>
<th>Exp(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 200</td>
<td>n = 400</td>
</tr>
<tr>
<td>p = 500</td>
<td>12.0 21.7 31.2</td>
<td>11.5 20.6 29.7</td>
</tr>
<tr>
<td>p = 1000</td>
<td>12.5 22.6 32.9</td>
<td>12.0 21.6 31.3</td>
</tr>
<tr>
<td>p = 2000</td>
<td>12.9 23.5 33.9</td>
<td>11.9 22.1 31.9</td>
</tr>
<tr>
<td>p = 4000</td>
<td>12.9 23.7 34.3</td>
<td>12.1 22.0 32.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 3</th>
<th>N(0,1)</th>
<th>Exp(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 200</td>
<td>n = 400</td>
</tr>
<tr>
<td>p = 500</td>
<td>11.3 21.3 30.9</td>
<td>11.2 19.9 29.2</td>
</tr>
<tr>
<td>p = 1000</td>
<td>12.5 22.5 32.6</td>
<td>12.0 21.7 31.1</td>
</tr>
<tr>
<td>p = 2000</td>
<td>12.1 23.4 32.6</td>
<td>12.1 22.0 31.9</td>
</tr>
<tr>
<td>p = 4000</td>
<td>12.6 23.4 33.8</td>
<td>12.3 22.3 32.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 4</th>
<th>N(0,1)</th>
<th>Exp(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 200</td>
<td>n = 400</td>
</tr>
<tr>
<td>p = 500</td>
<td>11.7 21.5 31.1</td>
<td>11.5 20.5 29.5</td>
</tr>
<tr>
<td>p = 1000</td>
<td>12.4 23.0 32.7</td>
<td>11.9 21.8 31.7</td>
</tr>
<tr>
<td>p = 2000</td>
<td>12.4 22.6 33.0</td>
<td>12.1 22.4 32.3</td>
</tr>
<tr>
<td>p = 4000</td>
<td>12.5 23.2 33.8</td>
<td>12.1 22.1 32.3</td>
</tr>
</tbody>
</table>

Table 1: Empirical FDR at various desired levels of α = 0.1, 0.2 and 0.3 under Model 1, 2 and 3 with error distributions for ε_i being N(0,1) and exponential with mean 1 (Exp(1)).
We now turn to the comparison of powers between our procedure and the B-Y procedure which is robust to the dependence structure. We use the proposed test statistics $T_i$ and the $\chi^2(4)$ distribution to calculate the $p$-values in B-Y method. We only report the powers for Models 3 and 4, $p_1 = 46$, $p = 1000$ and 2000 and $\alpha = 0.1$ because the powers for Models 1-2 are quite similar to those for Model 3. We let $\sigma_{ij1} = \sqrt{(1 + \delta) \log p/n}$ in Model 3 and $\sigma_{ij1} = \sqrt{(3 + \delta) \log p/n}$ in Model 4 for $i \neq j$, where $\delta$ grows from 0 to 1. Figure 2 summarizes the empirical power of the procedures over different values of $\delta$ for the sample size $n = 200$. Across all settings, the power of the proposed procedure is substantially higher than that of the B-Y method.

Figure 2: Comparing the power of our proposed procedure (++) to that of the B-Y method (⋄⋄) for Models 3 and 4 when $n = 200$. 
A simulation study is also carried out to compare the performance of the proposed method to that of the B-H and Storey et al. (2004) (Storey) procedures in terms of the FDP and FDR control when $p_1$ is not large. To this end, we consider the following model:

**Model 5.** Let $\Sigma_1 = (\sigma_{ij1}) \in \mathbb{R}^{20 \times 20}$ and $\Sigma_2 = (\sigma_{ij2}) \in \mathbb{R}^{(p-16) \times (p-16)}$. We let $\sigma_{ii1} = \sigma_{ii2} = 1$ for $1 \leq i \leq p + 4$. and $\sigma_{ij1} = \sqrt{2\log p/n}$ for $i \neq j$. $\Sigma_2 = \text{diag}(D_1, \ldots, D_m, I)$, where $m = \lfloor (p - 16)/10 \rfloor$, $D_k \in \mathbb{R}^{10 \times 10}$, $1 \leq k \leq m$. $I$ is the identity matrix. All of the off-diagonal entries of $D_k$ are taken to be 0.5.

In Model 5, $Y$ is associated with $X_1, \ldots, X_{p_1}$ with $p_1 = 16$. We set the nominal FDR and FDP level at $\alpha = 0.1$. Table 2 shows the empirical FDR of the proposed, B-H and Storey procedures for various choices of $p$, $n$ and error distributions. The results suggest that the proposed method is able to achieve the target FDR level while both the B-H and Storey procedures have substantially inflated FDR under the sparse setting. Figure 3 plots the empirical survival distribution function of the observed FDP. The empirical probability that FDP $\geq 0.2$ of B-H method can be as large as 30% and the empirical probability of FDP $\geq 0.25$ exceeds 10%. Our method provides a much tighter control of the FDP with FDP $\geq 0.2$ occurring less than 4% of the time and FDP $\geq 0.25$ less than 1%.

These numerical results are consistent with the theoretical analysis discussed earlier. Both the B-H and the Storey procedures fail to control for the FDR or FDP due to the complex correlation structure and the estimated null distribution under the sparse setting. On the other hand, the B-Y procedure is overly conservative with substantially lower power compared to our method.

### 4.2 Application to PheWAS of Autoimmune Risk Alleles with EMR

We applied our proposed procedure to a cohort of 1837 RA subjects identified from the EMR of Brigham and Womens Hospital and Massachusetts General Hospital (Liao et al., 2010). Blood samples were collected for these RA cases using the BWH Specimen Bank from
Table 2: Empirical FDR ($\times 100$) of the proposed new procedure (New), B-H and Storey et al (2014) procedures at desired level of $\alpha = 0.1$ under Model 5 with error distributions for $\epsilon_i$ being $N(0,1)$ and exponential with mean 1 ($\text{Exp}(1)$).

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
<th>$n = 200$</th>
<th>$n = 400$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New</td>
<td>B-H</td>
<td>Storey</td>
<td>New</td>
</tr>
<tr>
<td>500</td>
<td>9.89</td>
<td>15.3</td>
<td>16.1</td>
<td>10.0</td>
</tr>
<tr>
<td>1000</td>
<td>10.3</td>
<td>16.8</td>
<td>17.5</td>
<td>9.15</td>
</tr>
<tr>
<td>2000</td>
<td>10.2</td>
<td>16.3</td>
<td>16.8</td>
<td>9.13</td>
</tr>
<tr>
<td>4000</td>
<td>8.95</td>
<td>14.9</td>
<td>15.4</td>
<td>8.76</td>
</tr>
</tbody>
</table>

Figure 3: Empirical survival distributions of FDP, $\hat{S}(\delta)$, for the proposed procedure (++), the B-H procedure (○○) and the Storey et al (2014) procedure (○○) under Model 5 with $\alpha = 0.1$, $n = 200$ and $N(0,1)$. 

20
2009-2010 and a range of single nucleotide polymorphisms (SNPs) associated with various auto-immune diseases were genotyped (Kurreeman et al., 2011). We limited our study to 1237 individuals of European ancestry because this was the majority (80%) of our cohort. For this analysis, we focused on the two major genetic risk factors for RA, the human leukocyte antigen shared epitope (HLA-SE) region (tag SNP rs6910071) and a loci in the PTPN22 gene (rs2476601). We are interested in conducting a PheWAS using these two SNPs to determine whether carriage of these risk alleles are associated with subphenotypes in RA. To perform the PheWAS, the ICD9 codes were grouped into clinically relevant diseases, termed as PheWAS code as suggested in Denny et al. (2010). For example, ICD9 codes 401-405, representing different types of hypertension and complications are grouped into a hypertension PheWAS code. For each of the PheWAS code, the value of the variable represents total count of ICD9 codes a patient received across all their hospital encounters. This analysis included 352 PheWAS codes that have a prevalence of $\geq 2\%$, where the prevalence is calculated as the fraction of patients with non-zero counts. Since these counts are highly skewed, a log($x + 1$) transformation was applied to each of the code count to obtain the $p = 352$ dimensional vector $X$ for the association analysis. The 2-dimensional vector $Y$ consists of the allele counts of the HLA-SE and PTPN22 SNPs, taking values 0, 1 or 2.

At an FDR level of 0.05, our proposed testing procedures identified 4 phenotypes as significantly associated with the HLA-SE and PTPN22: RA ($p$-value = 3.8E-7), chronic fatigue syndrome ($p$-value = 1.1E-4), back pain ($p$-value = 1.2E-4) and anemias ($p$-value = 7.0E-4). It is interesting to note that all 4 phenotypes are related to severity of the disease. Patients with more severe RA tends to have a higher number of RA ICD9 codes, suggesting more visits related to RA. Anemia of chronic disease is also frequently associated in RA patients with high disease activity and higher levels of inflammation (Masson, 2011). Our findings are consistent with previous studies demonstrating an association between carriage of the HLA-SE and RA disease severity (Weyand et al., 1992; Jaraquemada et al., 1986) and hence is likely to be associated with the phenotypes associated with RA severity. Chronic fatigue syndrome
(CFS), a multifactorial condition, is highly related to autoimmune diseases. High levels of inflammation result in symptoms of profound fatigue and may explain the association with chronic fatigue syndrome (CFS). In addition, recent gene expression studies have confirmed that several immune function related genes are candidate markers of CFS (Fang et al., 2006). Since both the HLA-SE and PTPN22 genes are important in immune function and both predispose to other autoimmune diseases (Criswell et al., 2005; Davidson and Diamond, 2001), it is interesting to see their association with CFS.

5 Discussion

We introduced in this paper a multiple testing procedure for simultaneously testing columns of high-dimensional cross-covariance matrices. As mentioned in Section 2, there is an important difference between our FDR control procedure given in (2.5) and the well-known B-H procedure. When the test statistics $T_i$ are used and the $p$-values are calculated from $\mathbb{P}(\chi^2_d \geq t)$, the B-H procedure is equivalent to rejecting $H_0$ whenever $T_i \geq \hat{t}_{BH}$. For our procedure, as indicated in (2.5), if $t_{BH} > b_p$ we do not use $\hat{t}_{BH}$ but instead threshold at $a_p = 2 \log p + (d - 1) \log(\log p)$ to control the FDP and FDR. In other words, for $t > b_p$, we do not use $pG(t)$ to estimate $\hat{R}_0(t) \equiv \sum_{i \in H_0} I\{T_i \geq t\}$ while the B-H method does. It is important to note that, when $pG(t)$ is bounded or converges to infinity slowly, it is not a good estimator for $\hat{R}_0(t)$ and can even be inconsistent. For example, if we treat $T_i$ as i.i.d. random variables with the cumulative distribution function $1 - G(t)$, then $\hat{R}_0(t)$ is a binomial random variable with success probability $G(t)$. So if $pG(t)$ is bounded, then $\hat{R}_0(t)$ approximately follows a Poisson distribution with rate $p_0G(t)$ and $pG(t)$ is no longer a consistent estimator for $\hat{R}_0(t)$. Thus thresholding test statistics at $\hat{t}_{BH}$ would lead to unstable behavior of the FDP and ultimately fail to control the FDR when the signals are very sparse.

We next argue that, if $\hat{t}$ in (2.5) does not exist, then thresholding the test statistics at $a_p = 2 \log p + (d - 1) \log(\log p)$ is a reasonable way to control FDP. To explain this point
clearly, we would assume the number of true alternative \( p_1 = 10. \) Let \( \hat{m}_0 \) be the number of wrong rejections by any multiple tests procedure. Then \( FDP \geq \hat{m}_0/(10 + \hat{m}_0). \) So, if we want to control \( FDP \leq 0.05, \) for example, then it is necessary to make sure \( \hat{m}_0 = 0. \) In this case, control \( FDP \) is equivalent to control FWER. For general fixed \( p_1, \) control \( FDP \) at level \( 1/(p_1 + 1) \) is essentially equivalent to control FWER. When \( p_1 \) is fixed as \( p \to \infty, \) \( \hat{t} \) in (2.5) does not exist with probability tending to one and our procedure would simply threshold the test statistics at \( a_p \) to control \( FDP. \) In fact, when \( p_1 \) is fixed, it can be shown that the B-H method is unable to control \( FDP \) at any level \( 0 < \alpha < 1. \) Specifically, consider \( p \) tests with independent \( p \)-values \( P_1, \ldots, P_p \). If \( \min_{i \in H_1} P_i = o_p(p^{-1}) \), then for any \( 0 < \alpha < 1, \) there exists \( c_0 > 0 \) such that \( \lim \inf_{p \to \infty} P(FDP \geq \alpha) \geq c_0. \)

6 Proofs

We prove the main theorems in this section.

6.1 Proof of Theorems 3.1 and 3.2.

Without loss of generality, we assume \( \sigma_{ii} = 1 \) for \( 1 \leq i \leq p. \) The proof of Theorems 3.1 and 3.2 mainly relies on the distribution of the test statistic \( T_i = n(\hat{\sigma}_i)'\Sigma^{-1}_{Z_i}\hat{\sigma}_i \) and their tail probabilities. To approximate the distribution of \( T_i, \) consider

\[
T_i^o = n(\hat{\sigma}_i)'\Sigma^{-1}_{Z_i}\hat{\sigma}_i = \left\| \sqrt{n}\Sigma^{-1/2}_{Z_i}\hat{\sigma}_i \right\|^2 = \left\| \sum_{k=1}^{n} \xi_{ki} \right\|^2,
\]

where \( \hat{\sigma}_i = n^{-1}\sum_{k=1}^{n} (Z_{ki} - \sigma_i), \xi_{ki} = \Sigma^{-1/2}_{Z_i}(Z_{ki} - \sigma_i), Z_{ki} = (Y_k - \mu_Y)(X_{ki} - \mu_i) \) and \( \| \cdot \| \) denotes the Euclidean norm. Define a truncated version of \( \xi_{ki}, \)

\[
\hat{\xi}_{ki} = \xi_{ki}I\{\|\xi_{ki}\| \leq \sqrt{n}/(\log p)^4\} = E[\xi_{ki}I\{\|\xi_{ki}\| \leq \sqrt{n}/(\log p)^4\}].
\]

Then, uniformly in \( 1 \leq i \leq p, \)

\[
P\left\{ \left\| n^{-1/2}\sum_{k=1}^{n} (\xi_{ki} - \hat{\xi}_{ki}) \right\| \geq (\log p)^{-2} \right\} \leq nP\left\{ \|\xi_{1i}\| \geq \sqrt{n}/(\log p)^{4}\right\} = O(p^{-1-\epsilon_1})
\]
for some $\epsilon_1 > 0$. By Theorem 1 in Zaitsev (1987), we have for any $x \in \mathbb{R}^d$,
\[
\mathbb{P}\left(\left\|n^{-1/2} \sum_{k=1}^{n} \hat{\xi}_{ki} + x\right\| \geq t\right) \leq \mathbb{P}\left\{\left\|\hat{W} + x\right\| \geq t - (\log p)^{-2}\right\} + c_{1d}e^{-c_{2d}(\log p)^2}
\]
and
\[
\mathbb{P}\left(\left\|n^{-1/2} \sum_{k=1}^{n} \hat{\xi}_{ki} + x\right\| \geq t\right) \geq \mathbb{P}\left\{\left\|\hat{W} + x\right\| \geq t + (\log p)^{-2}\right\} - c_{1d}e^{-c_{2d}(\log p)^2},
\]
uniformly in $t \in \mathbb{R}$ and $1 \leq i \leq p$, where $\hat{W}$ is a $d$-dimensional normal random vector with mean zero and covariance matrix $\text{Cov}(\hat{\xi}_{ki})$, $c_{1d}$ and $c_{2d}$ are some constants depending only on $d$. We have $\|\text{Cov}(\hat{\xi}_{ki}) - I\| \leq Cn^{-2\beta}$. Then it is easy to show that
\[
\mathbb{P}\left\{\left\|\hat{W} + x\right\| \geq t - (\log p)^{-2}\right\} \leq \mathbb{P}(\|W + x\| \geq t - 2(\log p)^{-2}) + c_{3d}e^{-c_{4d}n^{2\beta}/(\log p)^4}
\]
and
\[
\mathbb{P}\left\{\left\|\hat{W} + x\right\| \geq t + (\log p)^{-2}\right\} \geq \mathbb{P}(\|W + x\| \geq t + 2(\log p)^{-2}) - c_{3d}e^{-c_{4d}n^{2\beta}/(\log p)^4},
\]
where $W$ is the standard normal random vector. Hence, for some $\epsilon_1 > 0$,
\[
\mathbb{P}(\|n^{-1/2} \sum_{k=1}^{n} \xi_{ki} + x\| \geq t) \leq \mathbb{P}(\|W + x\| \geq t - 2(\log p)^{-2}) + O(p^{-1-\epsilon_1}),
\]
\[
\mathbb{P}(\|n^{-1/2} \sum_{k=1}^{n} \xi_{ki} + x\| \geq t) \geq \mathbb{P}(\|W + x\| \geq t + 2(\log p)^{-2}) - O(p^{-1-\epsilon_1}),
\]
where $O(1)$ is uniformly in $t \in \mathbb{R}$ and $1 \leq i \leq p$. This yields that, for any fixed $\delta > 0$,
\[
\mathbb{P}\left\{\max_{1 \leq i \leq p} \left\|n^{-1/2} \sum_{k=1}^{n} \xi_{ki}\right\|^2 \leq (2 + \delta) \log p\right\} \rightarrow 1.
\]

Since $\tilde{\sigma}_i = \sigma_i + \tilde{\sigma}_i - (\bar{Y} - \mu_Y)(\bar{X}_i - \mu_X_i)$, we may write
\[
T_i^{1/2} = \sqrt{n} \left\|\Sigma_{Zi}^{-1/2} \sigma_i - \Sigma_{Zi}^{-1/2} (\bar{Y} - \mu_Y)(\bar{X}_i - \mu_X_i) + (\Sigma_{Zi}^{-1/2} - \Sigma_{Zi}^{-1/2}) \tilde{\sigma}_i + \Sigma_{Zi}^{-1/2} \tilde{\sigma}_i\right\|. \tag{6.13}
\]
By the proof of Lemma 2 in Cai and Liu (2011), we have for some $C > 0$,
\[
\mathbb{P}\left(\max_{1 \leq i \leq p} |\bar{X}_i - \mu_i| \geq C \sqrt{\frac{\log p}{n}}\right) \rightarrow 0, \quad \mathbb{P}\left(\|\bar{Y} - \mu_Y\| \geq C \sqrt{\frac{\log p}{n}}\right) \rightarrow 0, \tag{6.14}
\]
\[
\mathbb{P}\left(\max_{1 \leq i \leq p} |\tilde{\sigma}_i - \sigma_i| \geq C \sqrt{\frac{\log p}{n}}\right) \rightarrow 0, \quad \text{and} \quad \mathbb{P}\left(\max_{1 \leq i \leq p} \|\hat{\Sigma}_{Zi} - \Sigma_{Zi}\| \geq C \sqrt{\frac{\log p}{n}}\right) \rightarrow 0. \tag{6.15}
\]
By (6.13), (6.14) and (6.15), we obtain that
\[
\mathbb{P}\left(\max_{1 \leq i \leq p} \left|\frac{1}{T_i^{1/2}} - \left\|\sqrt{n} \Sigma_{Zi}^{-1/2} \tilde{\sigma}_i + \sqrt{n} \Sigma_{Zi}^{-1/2} \sigma_i\right\|\right| \geq C \sqrt{\frac{(\log p)^2}{n}}\right) \rightarrow 0. \tag{6.16}
\]
This, together with the above arguments, implies that the following lemma.
Lemma 6.1. We have
\[
\max_{i \in H_0} \left| \frac{P(T_i \geq t)}{G(t)} - 1 \right| \rightarrow 0
\]
uniformly in \( t \in [0, a_p] \).

Next, define
\[
H_1(c) = \{ i : \sigma_i' \Sigma^{-1}_{Z_i} \sigma_i \geq c \log \frac{p}{n} \} \quad \text{and} \quad \overline{H}_1(c) = \{ i : \sigma_i' \Sigma^{-1}_{Z_i} \sigma_i < c \log \frac{p}{n} \}.
\]
For \( i \in H_1(10) \), by (6.12), (6.15) and (6.16), \( P(T_i \geq 2 \log p) \rightarrow 1 \) uniformly in \( i \). On the other hand,
\[
P\left( \max_{i \in \overline{H}_1(10)} \left| T_i^{1/2} - \sqrt{n} \Sigma^{-1/2}_{Z_i} \sigma_i + \sqrt{n} \Sigma^{-1/2}_{Z_i} \sigma_i \right| \geq C \sqrt{\frac{(\log p)^2}{n}} \right) \rightarrow 0. \tag{6.17}
\]
For \( i \in \overline{H}_1(10) \cap H_1(c) \) for some \( c > 2 \), uniformly in \( i \) we have
\[
P \left\{ \| W + \sqrt{n} \Sigma^{-1/2}_{Z_i} \sigma_i \| \geq \sqrt{2 \log p + 2(\log p)^2} \right\} \rightarrow 1.
\]

It follows from (6.12), (6.16) and (6.17) that \( P(T_i \geq 2 \log p) \rightarrow 1 \) uniformly in \( i \in H_1(c) \) for any \( c > 2 \). Thus, whenever \( H_1(c) \neq \emptyset \), we have
\[
\frac{\sum_{i \in H_1(c)} I\{T_i \geq b_p\}}{\text{Card}\{H_1(c)\}} \rightarrow 1, \quad \text{in probability.} \tag{6.18}
\]
If (3.8) holds, then we have \( \text{Card}\{H_1(c)\} \geq (1 - \varepsilon) \log p \) for any \( \varepsilon > 0 \). In this case, \( P(\hat{t} \leq b_p) \rightarrow 1 \).

Now with these distributional properties of \( T_i \), we return to the proof of Theorems 3.1 and 3.2. When \( \hat{t} \) in (2.5) exists, by the continuity of \( G(t) \) and the monotonicity of the indicator function,
\[
G(\hat{t}) = \frac{\alpha \max\{\sum_{1 \leq i \leq p} I(T_i \geq \hat{t}), 1\}}{p}
\]
and hence
\[
\text{FDP} = \frac{\alpha \sum_{i \in H_0} I(T_i \geq \hat{t})}{pG(\hat{t})}.
\]
If $\hat{t}$ in (2.5) does not exist, then $\{FDP \geq \varepsilon\} \subseteq \{\max_{i \in H_0} T_i \geq a_p\}$. Note that, by (6.12) and (6.16),

$$P(\max_{i \in H_0} T_i \geq a_p) \leq 2pG(a_p - 3(\log p)^{-1}) + O(p^{-\epsilon_1}) = O((\log p)^{-1/2}).$$

To prove Theorems 3.1 and 3.2, it suffices to show that

$$\sup_{0 \leq t \leq b_p} \left| \frac{\sum_{i \in H_0} I(T_i \geq t)}{p_0 G(t)} - 1 \right| \to 0 \text{ in probability.}$$

Let $b'_p = b_p + (\log p)^{-2}$. By (6.16), it is enough to prove that

$$\sup_{0 \leq t \leq b'_p} \left| \frac{\sum_{i \in H_0} I\{T_i \geq t\}}{p_0 G(t)} - 1 \right| \to 0 \text{ in probability.}$$

By the proof of Lemma 6.3 in Liu (2013), we only need to show that the following lemma.

**Lemma 6.2.** We have, for any $\varepsilon > 0$,

$$\sup_{0 \leq t \leq b'_p} P\left( \left| \frac{\sum_{i \in H_0} I\{T_i \geq t\} - \mathbb{P}(T_i \geq t)}{p_0 G(t)} \right| \geq \varepsilon \right) = o(1) \quad (6.19)$$

and

$$\int_0^{b'_p} P\left( \left| \frac{\sum_{i \in H_0} I\{T_i \geq t\} - \mathbb{P}(T_i \geq t)}{p_0 G(t)} \right| \geq \varepsilon \right) dt = o(v_p), \quad (6.20)$$

where $v_p = 1/\log \log p$.

To prove Lemma 6.2, define

$$B_1 = \{(i,j) : i \in H_0, j \in H_0, (i,j) \in A(\varepsilon), i \neq j\},$$

and $$B_2 = \{(i,j) : i \in H_0, j \in H_0, (i,j) \notin A(\varepsilon), i \neq j\}.$$ Then

$$\mathbb{E}\left( \sum_{i \in H_0} I\{T_i \geq t\} - \mathbb{P}(T_i \geq t) \right)^2 = \sum_{(i,j) \in B_1} \left[ \mathbb{P}(T_i \geq t, T_j \geq t) - \mathbb{P}(T_i \geq t)\mathbb{P}(T_j \geq t) \right]$$

$$+ \sum_{(i,j) \in B_2} \left[ \mathbb{P}(T_i \geq t, T_j \geq t) - \mathbb{P}(T_i \geq t)\mathbb{P}(T_j \geq t) \right]$$

$$+ \sum_{i \in H_0} \left[ \mathbb{P}(T_i \geq t) - (\mathbb{P}(T_i \geq t))^2 \right]. \quad (6.21)$$
For \((i, j) \in \mathcal{B}_2\), we have by Lemma 6.3 below,

\[
P(T_i^o \geq t, T_j^o \geq t) = (1 + A_n)P(T_i^o \geq t)P(T_j^o \geq t)
\] (6.22)

uniformly for \(0 \leq t \leq b'_p\), where \(|A_n| \leq C(\log p)^{-1-\gamma}\). For \((i, j) \in \mathcal{B}_1\), we have by Lemma 6.3, for any \(\delta > 0\),

\[
P(T_i^o \geq t, T_j^o \geq t) \leq C(t + 1)^{-1} \exp(-t/(1 + |\rho_{ij}| + \delta))
\] (6.23)

uniformly in \(0 \leq t \leq b'_p\). Submitting (6.22) and (6.23) into (6.21), we obtain

\[
E\left(\sum_{i \in H_0} \left[I\{T_i^o \geq t\} - P(T_i^o \geq t)\right]\right)^2 \leq C(\sum_{(i, j) \in A(\varepsilon)} e^{-t/|\rho_{ij}| + \delta} (1 + t)^{-1} + A_n p^2 G^2(t) + p G(t))
\]

uniformly in \(0 \leq t \leq b'_p\). Note that, by (C1) and letting \(\delta\) be sufficiently small,

\[
\sum_{(i, j) \in A(\varepsilon)} \int_0^{b'_p} \exp\left(\frac{|\rho_{ij}| + \delta}{1 + |\rho_{ij}| + \delta} t\right) dt = o(p^2 v_p).
\]

This, together with \(\int_0^{b'_p} 1/G(t) dt = O(p(\log p)^{-1/2})\), proves (6.20). (6.19) can be proved similarly. This concludes the proof of Theorem 3.2.

**Lemma 6.3.** (i). We have for any \(\delta > 0\),

\[
P(T_i^o \geq t, T_j^o \geq t) \leq C(t + 1)^{-1} \exp(-t/(1 + |\rho_{ij}| + \delta))
\]

uniformly in \(0 \leq t \leq b'_p\) and \((i, j) \in \mathcal{B}_1\). (ii). We have

\[
P(T_i^o \geq t, T_j^o \geq t) = (1 + A_n)P(T_i^o \geq t)P(T_j^o \geq t)
\]

uniformly in \(0 \leq t \leq b'_p\) and \((i, j) \in \mathcal{B}_2\), where \(|A_n| \leq C(\log p)^{-1-\gamma}\) for some \(\gamma > 0\).

To prove Lemma 6.3, we need the following lemma which comes from Lemma 6.2 in Liu (2013). Let \(\eta_k = (\eta_{k1}, \eta_{k2})'\) are independent and identically distributed 2-dimensional random vectors with mean zero.
Lemma 6.4. Suppose that $p \leq cn^r$ and $\mathbb{E}\|\eta_i\|^{2br+2+\epsilon} < \infty$ for some fixed $c > 0$, $r > 0$, $b > 0$ and $\epsilon > 0$. Assume that $\text{Var}(\eta_{11}) = \text{Var}(\eta_{12}) = 1$ and $|\text{Cov}(\eta_{11}, \eta_{12})| \leq \delta$ for some $0 \leq \delta < 1$.

Then we have

$$\mathbb{P}\left( \left| \sum_{k=1}^{n} \eta_{k1} \right| \geq t \sqrt{n}, \left| \sum_{k=1}^{n} \eta_{k2} \right| \geq t \sqrt{n} \right) \leq C(t+1)^{-2} \exp\left(-t^2/(1+|\text{Cov}(\eta_{11}, \eta_{12})|)\right)$$

uniformly for $0 \leq t \leq \sqrt{b \log p}$, where $C$ only depends on $c, b, r, \epsilon, \delta$.

Proof of Lemma 6.3. We first prove (i). Let

$$T_i^o(\alpha) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \alpha \xi_{ki}.$$  

For any $\|\alpha\| = 1$ and $\|\beta\| = 1$, we have, for $i \in \mathcal{H}_0$ and $j \in \mathcal{H}_0$,

$$|\text{Cov}(T_i^o(\alpha), T_j^o(\beta))| \leq \rho_{ij}^*.$$  

Let $\alpha_1, \ldots, \alpha_q$ satisfying $\|\alpha_j\| = 1$. For any $\|\alpha\| = 1$, there exists $\alpha_j$ such that $\|\alpha - \alpha_j\| \leq c_q$, where $c_q \to 0$ as $q \to \infty$ uniformly in $\alpha$ and $1 \leq j \leq q$.

Then

$$\left| (T_i^o)^{1/2} - \max_{1 \leq j \leq q} |T_i^o(\alpha_j)| \right| \leq c_q (T_i^o)^{1/2}.$$  

So we have

$$(T_i^o)^{1/2} \leq (1 - c_q)^{-1} \max_{1 \leq j \leq q} |T_i^o(\alpha_j)|.$$  

It follows from Lemma 6.4 that

$$\mathbb{P}(T_i^o \geq t, T_j^o \geq t) \leq \sum_{k=1}^{q} \sum_{l=1}^{q} \mathbb{P}\left\{ |T_i^o(\alpha_k)| \geq \sqrt{t}(1 - a_p), |T_i^o(\alpha_l)| \geq \sqrt{t}(1 - c_p) \right\} \leq C(t+1)^{-1} e^{-t/(1+\rho_{ij}^*+\delta)}$$  

for any $\delta > 0$ by letting $q$ sufficiently large. This proves (i).

To prove (ii), we first note that, using the similar arguments for (6.12) and Theorem 1 in Zaïtsev (1987),

$$\mathbb{P}(T_i^o \geq t, T_j^o \geq t) \leq \mathbb{P}(\|\hat{W}_1\|^2 \geq t', \|\hat{W}_2\|^2 \geq t') + c_{5d} \exp(-c_{6d}n^{2d}/(\log p)^4),$$

28
\[ P(T_i^o \geq t, T_j^o \geq t) \geq P(\|\hat{W}_1\|^2 \geq t'', \|\hat{W}_2\|^2 \geq t'') - c_5d \exp(-c_6d^{2/3}/(\log p)^4), \]
where \( t' = (\sqrt{t} - (\log p)^{-2})^2 \) and \( t'' = (\sqrt{t} + (\log p)^{-2})^2 \), and \((\hat{W}_1', \hat{W}_2')'\) is the normal random vector with mean zero and covariance matrix \( \text{Cov}(\hat{\xi}_{kij}) \), where \( \hat{\xi}_{kij} = (\hat{\xi}_{ki}', \hat{\xi}_{kj}')' \). We have
\[ \|\text{Cov}(\hat{\xi}_{kij}) - I\| \leq C/(\log p)^{2-\varepsilon}. \]

By the density of multivariate normal random vector,
\[ P(\|\hat{W}_1\|^2 \geq t', \|\hat{W}_2\|^2 \geq t') = (1 + A_n)[G(t)]^2. \]
Similar equation holds when \( t' \) is replaced by \( t'' \). This proves (ii).

6.2 Proof of Theorems 3.3 and 3.4.

By the proof of (6.18), for \( t \sim 2(1 - \theta) \log p \),
\[ \frac{\sum_{i \in \mathcal{H}_1(c)} I(T_i \geq t)}{m_1(c)} \rightarrow 1 \quad (6.24) \]
in probability. Then, for \( t \sim 2(1 - \theta) \log p \),
\[ \frac{\sum_{i=1}^{p} I(T_i \geq t)}{p} \geq (1 + o(1))p^{-1+\theta} \]
with probability tending to one. So \( P\{0 \leq \hat{t} \leq G^{-1}(\alpha p^{-1+\theta}/2)\} \rightarrow 1 \). Hence, \( \hat{FO} \rightarrow 1 \) in probability. Theorem 3.3 follows immediately by let \( b_p = G^{-1}(\alpha p^{-1+\theta}/2) \) in the proof of Theorem 3.2.

References


rheumatoid arthritis: susceptibility or severity?” *Disease markers*, 4, 43–53.


