Statistica Sinica **29** (2019), 983-1005
doi:https://doi.org/10.5705/ss.202017.0189

# LARGE-SCALE SIMULTANEOUS TESTING OF CROSS-COVARIANCE MATRICES WITH APPLICATIONS TO PheWAS

Tianxi Cai, T. Tony Cai, Katherine Liao and Weidong Liu

*Harvard T.H Chan School of Public Health, University of Pennsylvania, Brigham and Women's Hospital and Shanghai Jiao Tong University*

*Abstract:* Motivated by applications in phenome-wide association studies (PheWAS), we consider in this paper simultaneous testing of columns of high-dimensional cross-covariance matrices and develop a multiple testing procedure with theoretical guarantees. It is shown that the proposed testing procedure maintains a desired false discovery rate (FDR) and false discovery proportion (FDP) under mild regularity conditions. We also provide results on the magnitudes of the signals that can be detected with high power. Simulation studies demonstrate that the proposed procedure can be substantially more powerful than existing FDR controlling procedures in the presence of correlation of unknown structure. The proposed multiple testing procedure is applied to a PheWAS of two auto-immune genetic markers using a rheumatoid arthritis patient cohort constructed from the electronic medical records of Partners Healthcare System.

*Key words and phrases:* Covariance, false discovery rate, multiple responses, multiple testing, PheWAS.

## 1. Introduction

Simultaneously assessing the associations among a large number of variables is an important problem in statistics with a wide range of applications. For example, large-scale testing tools are frequently needed for 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); Nelson et al. (2004)); and phenome-wide association studies (PheWAS) which examine the relationships between a large number of disease phenotypes and some candidate genomic markers (Denny et al. (2010)).

A critical step in performing a large-scale multiple testing is to control the false discovery rate (FDR). Standard FDR control procedures, such as the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg (1995)), typically built under the independence assumption, would fail to provide desired

error controls in the presence of strong correlation. Alternative methods that allow for the general dependency, including the Benjamini-Yekutieli (BY) procedure (Benjamini and Yekutieli (2001)) and those proposed in Romano and Shaikh (2006a,b) and Guo and Rao (2008), 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, Han and Gu (2012); Fan and Han (2017), e.g.). However, multivariate outcomes are often of interest in many applications such as the PheWAS. These aforementioned methods may not be valid or powerful for such PheWAS settings due to the more complex correlation structure of the test statistic and the exact null distribution being unknown.

In this paper, we formulate the problem of assessing the association between a large number of variables and a vector of outcomes as 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 structures.

### 1.1. The problem

We consider a problem of multiple testing for columns of high-dimensional cross-covariance matrices. Let $\mathscr{D} = \{(\boldsymbol{Y}_k', \boldsymbol{X}_k')', 1 \le k \le n\}$, where $\boldsymbol{Y}_k = (Y_{k1}, \ldots, Y_{kd})'$ and $\boldsymbol{X}_k = (X_{k1}, \ldots, X_{kp})'$, be a random sample consisting of $n$ independent and identically distributed copies of $(\boldsymbol{Y}, \boldsymbol{X})$. In the PheWAS setting, $\boldsymbol{Y}$ may be a vector of genomic markers and $\boldsymbol{X}$ represents all phenotypic disease conditions of interest. For $1 \le i \le p$, define the cross-covariance vector between $\boldsymbol{Y}$ and $X_i$ by $\boldsymbol{\sigma}_i = (\sigma_1^{(i)}, \ldots, \sigma_d^{(i)})'$, where $\sigma_j^{(i)} = \mathsf{Cov}(Y_j, X_i)$. Thus, $\boldsymbol{\sigma}_i$ is the $i$th column of the cross-covariance matrix

$$\boldsymbol{\Sigma}_{YX} \equiv \mathsf{Cov}(\boldsymbol{Y}, \boldsymbol{X}) = [\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_p]$$

between $\boldsymbol{Y}$ and $\boldsymbol{X}$. We wish to simultaneously test the collection of $p$ hypotheses

$$H_{0i} : \boldsymbol{\sigma}_i = \boldsymbol{0} \quad \text{versus} \quad H_{1i} : \boldsymbol{\sigma}_i \neq \boldsymbol{0}, \quad 1 \le i \le p \tag{1.1}$$

based on the random sample $\mathscr{D}$, while controlling the overall false discovery rate (FDR) and false discovery proportion (FDP). We are interested in multiple testing of the columns, not individual entries, of the cross-covariance matrix $\boldsymbol{\Sigma}_{YX}$. Here the $i$th hypothesis examines whether the marker vector $\boldsymbol{Y}$ 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. Here, we consider linear associations between $\boldsymbol{Y}$ and

$\boldsymbol{X}$ but note that transformations can be performed on the original data. For example we typically transform the ICD9 code $X_j$ as $\log(1 + X_j)$ to increase the power of analysis since the count of ICD9 codes tend to be highly skewed.

Large-scale multiple testing of the columns of a high-dimensional cross-covariance matrix is technically difficult due to the complex entrywise dependence structures. No existing multiple testing procedures can be applied to this problem to provide accurate FDR or FDP control. Available methods and theoretical results largely focus on $z$- or $t$-tests for univariate $Y$ and commonly require the knowledge of the null distribution. For example, Fan, Han and Gu (2012) and Fan and Han (2017) 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 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 current setting necessitates the use of test statistics with only approximate null distributions. No existing theoretical results can ensure accurate control of the 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 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 BH method can converge to one. It is thus critical to justify the use of asymptotic null distribution for proper FDR control. If the number of true signals $p_1 \geq cp$ for some $c > 0$, the method in Storey, Taylor and Siegmund (2004) incorporates the estimation of the proportion of the true nulls into the BH method. In the PheWAS setting, the signal is sparse and $p_1$ is of order $o(p)$, which leads to the equivalence of the Storey, Taylor and Siegmund (2004) and BH 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, Taylor and Siegmund (2004) and BH procedures is able to control FDR or FDP in our setting, as confirmed via our simulation studies.

We first discuss the motivating application of PheWAS with multiple outcomes and then discuss the challenges in large-scale multiple testing of columns of high-dimensional cross-covariance matrices and summarize the main contribu-

tions of the paper.

### 1.2. PheWAS of a set of genomic markers

Complementary to GWAS, PheWAS investigates the association between a set of candidate genomic markers and a diverse range of phenotypes. PheWAS enables the discovery of genetic markers with pleiotropic effects, and thus may provide a broader view of the relationship between genetic variation and networks of phenotypes (Pendergrass et al. (2013); 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 patient level phenotypic data including ICD9 codes for a wide range of disease conditions. To enable efficient genetic research, research institutions such as Partners Healthcare System (PHS) also link the EMR to bio-specimen repositories that collect blood samples. Anonymized patient level phenotype data linked with genotype data can be extracted and stored for clinical research (Liao et al. (2015)). Through such a process, a cohort of 1,837 rheumatoid arthritis (RA) patients with their genomic and phenotypic information available has been established at PHS for discovery research (Liao et al. (2010); Kurreeman et al. (2011); Liao et al. (2013)).

Such linked EMR data enables us to conduct PheWAS to rigorously study the association between a large number of disease phenotypes $\boldsymbol{X}$ and any given set of candidate genomic markers $\boldsymbol{Y}$. In contrast to GWAS, the large number of phenotypic variables in PheWAS are often substantially correlated. The complex yet unknown correlation along with the multi-dimensionality of $\boldsymbol{Y}$ contribute significantly to the difficulty of the multiple testing problem. We translate the PheWAS problem into the problem of testing high dimensional cross-covariance matrices and our proposed procedures can overcome such difficulties to draw valid conclusions from the PheWAS.

### 1.3. Our contributions

There are two main challenges for simultaneous testing of columns of high-dimensional cross-covariance matrices: (i) the construction of suitable test statistics for individual hypotheses and providing proper estimates for the null distribution of the test statistics; (ii) the construction of a good procedure to account for the multiplicity of the tests so that the overall FDR and FDP are controlled

in the presence of dependency and additional approximation error due to the use of the estimated null distribution. We summarize the main contributions of this paper as follows.

1. We propose a test statistic $T_i$ that mimics the Hotelling's $T^2$ statistic for testing an individual hypothesis $H_{0i}$. It is shown that in the presence of correlation induced by the observed data, the null distribution of this test statistic can be well approximated by a $\chi^2$ distribution within an appropriate range.

2. We develop a large-scale multiple testing procedure for $\{H_{0i}, i = 1, \ldots, p\}$ by thresholding the test statistics $\{T_i,\ i = 1, \ldots, p\}$ with theoretical guarantees for FDR and FDP control under flexible correlation structures. Both theoretical and numerical properties are investigated.

   It is proven that, under mild regularity conditions, the proposed multiple testing procedure based on the asymptotic null distribution of $T_i$ controls both the FDR and FDP. In addition, we study the power property of the procedure and 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.

3. Our methods also contribute to the practice of PheWAS using EMR data. Standard PheWAS defines each disease phenotype $X_j$ based on whether a patient has at least 1 or 2 ICD9 billing codes and then tests for the association between the binary phenotype and a single marker $Y$ (Denny et al. (2010); Liao et al. (2013)). Our proposed procedures allow clinical investigators to make use of the counts of ICD9 codes, which could be more powerful and more robust since it is difficult to choose the appropriate threshold for each disease as the ICD9 codes have varying degree of accuracy (Liao et al. (2010)).

   Our methods allow one to identify the subset of phenotypes that are associated with multiple, say $d$, related markers as a group. An alternative strategy is to perform FDR-controlled marginal testing of $dp$ hypotheses to identify the set of phenotype and marker pairs that are associated. However, this approach is generally less powerful than the proposed approach based on Hotelling $T$ statistics. Similar patterns have been previously reported when comparing entry-wise testing versus group level testing (Xia, Cai and Cai (2017), e.g.).

4. The proposed procedure is applied to a PheWAS of two important autoimmune genetic markers using the EMR RA cohort described above. The goal is to comprehensively evaluate how these genetic variations may contribute to comorbidities of RA. The results show that these risk alleles are potentially associated with RA severity, chronic fatigue syndrome, back pain and anemia.

### 1.4. Structure of the paper

The rest of the paper is organized as follows. The proposed methodology detailing the test statistics for individual hypotheses as well as the simultaneous testing procedure is introduced in Section 2. A theoretical analysis of the multiple testing procedure, presented in Section 3 and proved in the supplementary materials, 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, along with those of the application of the proposed procedures to the aforementioned RA cohort. Section 5 discusses a few related issues. Proofs for the theoretical results are given in the Supplementary Materials.

## 2. Methodology

We detail the proposed multiple testing procedure in this section. We first introduce the test statistics for testing the individual hypothesis $H_{0i} : \boldsymbol{\sigma}_i = \mathbf{0}$. Let $\hat{\boldsymbol{Z}}_{ki} = \big(\boldsymbol{Y}_k - \bar{\boldsymbol{Y}}\big)\big(X_{ki} - \bar{X}_i\big)$ where $\bar{\boldsymbol{Y}} = n^{-1}\sum_{k=1}^n \boldsymbol{Y}_k$ and $\bar{X}_i = n^{-1}\sum_{k=1}^n X_{ki}$ for $1 \le i \le p$ and $1 \le k \le n$. Then

$$\hat{\boldsymbol{\sigma}}_i = n^{-1}\sum_{k=1}^n \hat{\boldsymbol{Z}}_{ki}$$

is a consistent and asymptotically unbiased estimator of $\boldsymbol{\sigma}_i$, the covariance vector between $\boldsymbol{Y}$ and $X_i$. A test for $H_{0i}$ may be constructed based on the observed $n^{1/2}\hat{\boldsymbol{\sigma}}_i$ along with its covariance matrix that can be approximated using the sample covariance matrix of $\big\{\hat{\boldsymbol{Z}}_{ki},\, k = 1, \ldots, n\big\}$,

$$\hat{\boldsymbol{\Sigma}}_{Zi} = n^{-1}\sum_{k=1}^n \big(\hat{\boldsymbol{Z}}_{ki} - \hat{\boldsymbol{\sigma}}_i\big)\big(\hat{\boldsymbol{Z}}_{ki} - \hat{\boldsymbol{\sigma}}_i\big)'.$$

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

$$T_i = n(\hat{\boldsymbol{\sigma}}_i)'\big(\hat{\boldsymbol{\Sigma}}_{Zi}\big)^{-1}\hat{\boldsymbol{\sigma}}_i. \tag{2.1}$$

Let $\mathcal{H}_0 = \{i : \boldsymbol{\sigma}_i = \mathbf{0}\}$ be the index set of all null hypotheses, $\mathcal{H}_1 = \{i : \boldsymbol{\sigma}_i \neq 0\}$, and $p_0 = \mathrm{Card}(\mathcal{H}_0)$. As indicated in Lemma 1 in the Supplementary Materials, under the conditions of Theorem 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 $G(t) = \mathbb{P}\big(\chi^2_d \geq t\big)$ and

$$a_p = 2\log p + (d-1)\log\log p. \tag{2.2}$$

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 $\{T_i, i = 1, \ldots, p\}$. 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 $\mathscr{D}$ is

$$\mathrm{FDP}(t) = \frac{\sum_{i\in\mathcal{H}_0} I(T_i \geq t)}{\max\{\sum_{i=1}^{p} I(T_i \geq t), 1\}}. \tag{2.3}$$

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

$$\hat{t}_0 = \inf\{t : \mathrm{FDP(t)} \leq \alpha\}.$$

Since the denominator of $\mathrm{FDP}(t)$ in (2.3) 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

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

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 $p_0^{-1}\sum_{i\in\mathcal{H}_0} I(T_i \geq t)$. Here $p_0$ can be further estimated by $p$ due to the sparsity in the number of alternative hypotheses in many data applications. Based on this analysis, we propose a multiple testing procedure for simultaneously testing the hypotheses in (1.1):

**FDR control procedure.** Calculate $T_i$ in (2.1) and, for any given nominal FDR level $\alpha \in (0, 1)$, reject $H_{0i}$ whenever $T_i \geq \hat{t}$, where

$$\hat{t} = \inf\left\{0 \le t \le b_p : G(t) \le \frac{\alpha \max\left\{\sum_{1\le i\le p} I(T_i \ge t), 1\right\}}{p}\right\} \tag{2.4}$$

if the righthand side of (2.4) exists; and $\hat{t} = a_p$ otherwise.

When $\hat{t}$ in (2.4) does not exist, the FDR control procedure simply thresholds the test statistics at the value $a_p$ given in (2.2). Unlike the conventional BH 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 (2.4) may not exist and the true FDP of the BH 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. Proof of these results can be found in the Supplementary Materials. Numerical performance of the procedure will be studied in Section 4. We first state some conditions on the correlation structure. Let $\boldsymbol{Z}_i = (\boldsymbol{Y} - \boldsymbol{\mu}_Y)(X_i - \mu_i)$, $\Sigma_{Zi} = \mathsf{Cov}(\boldsymbol{Z}_i)$ and $\boldsymbol{\xi}_i = \boldsymbol{\Sigma}_{Zi}^{-1/2}(\boldsymbol{Z}_i - \boldsymbol{\sigma}_i)$. As in canonical correlation analysis, define the maximum correlation coefficients $\rho_{ij}^* = \max_{\|\boldsymbol{a}\|=1,\|\boldsymbol{b}\|=1} |corr\big(\boldsymbol{a}'\boldsymbol{\xi}_i, \boldsymbol{b}'\boldsymbol{\xi}_j\big)|$, where $\|\cdot\|$ denotes the Euclidean norm. $\rho_{ij}^*$ characters the dependence between $T_i$ and $T_j$.

Define

$$\mathcal{B}(\delta) = \{(i,j) :\ i \in \mathcal{H}_0,\ j \in \mathcal{H}_0,\ \rho_{ij}^* \ge \delta, i \ne j\} \quad \text{with } \delta \in (0,1).$$

The set $\mathcal{A}(\varepsilon) = \mathcal{B}\{(\log p)^{-2-\varepsilon}\}$ includes the pairs $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ that are strongly correlated for $i, j \in \mathcal{H}_0$. 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

$$\sum_{(i,j)\in\mathcal{A}(\varepsilon)} p^{\{2\rho_{ij}^*/(1+\rho_{ij}^*)\}+\delta} = O\big(p^2(\log p)^{-2}\big). \tag{3.1}$$

**Remark 1.** When $(\boldsymbol{Y}', \boldsymbol{X}')$ has an elliptically contoured distribution, it is easy to see that $\rho_{ij}^* \le |\rho_{ij}|$ for $i, j \in \mathcal{H}_0$, where $(\rho_{ij})_{1\le i,j\le p}$ is the correlation matrix of $\boldsymbol{X}$. In this case, (3.1) is reduced to

$$\sum_{(i,j)\in\mathcal{A}_1(\varepsilon)} p^{\{2|\rho_{ij}|/(1+|\rho_{ij}|)\}+\delta} = O\Big(p^2(\log p)^{-2}\Big), \tag{3.2}$$

where $\mathcal{A}_1(\varepsilon)$ is defined as $\mathcal{A}(\varepsilon)$ with $\rho_{ij}^*$ replaced by $|\rho_{ij}|$. Condition (3.1) holds if $\mathrm{Card}\{\mathcal{B}(\delta)\} = O(p^\rho)$ for any $0 < \delta < 1$ and some $\rho < 2/(1+\delta)$, and $\mathrm{Card}\{\mathcal{A}(\varepsilon)\} = O(p^\rho)$ for some $\rho < 2$ and $\varepsilon > 0$. The correlation condition (3.1) 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 $\mathrm{Card}\{\mathcal{B}(\delta)\}$ and $\mathrm{Card}\{\mathcal{A}(\varepsilon)\}$ are reasonable.

**Remark 2.** When $i \in \mathcal{H}_0$ and $j \in \mathcal{H}_0$, we have $\boldsymbol{Y}$ uncorrelated with $(X_i, X_j)$. If we assume that $\boldsymbol{Y}$ is independent from $(X_i, X_j)$ for $i \in \mathcal{H}_0$ and $j \in \mathcal{H}_0$, then $\rho_{ij}^* = |\rho_{ij}|$. In this case, (C1) is reduced to a correlation condition on $\boldsymbol{X}$ which is quite natural.

### 3.1. FDR and FDP control

For the multiple testing procedure defined in Section 2, the FDP and FDR are given by

$$\mathrm{FDP} = \frac{\sum_{i\in\mathcal{H}_0} I\big(T_i \ge \hat{t}\big)}{\max\big\{\sum_{1\le i\le p} I\big(T_i \ge \hat{t}\big), 1\big\}} \quad \text{and} \quad \mathrm{FDR} = \mathbb{E}(\mathrm{FDP}).$$

The next two theorems show 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,

$$m_1(c) = \mathrm{Card}\left\{ i : 1 \le i \le p,\ \boldsymbol{\sigma}_i' \boldsymbol{\Sigma}_{Zi}^{-1} \boldsymbol{\sigma}_i \ge \frac{c(\log p)}{n} \right\},$$

is not too small for some $c$. Let $\lambda_{i1}$ and $\lambda_{id}$ be the largest and smallest eigenvalues of $\Sigma_{Zi}$, respectively.

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

$$\lim_{(n,p)\to\infty} \mathbb{P}(\mathrm{FDP} \le \alpha + \varepsilon) = 1 \quad \textit{and} \quad \limsup_{(n,p)\to\infty} \mathrm{FDR} \le \alpha.$$

**Theorem 2.** *Suppose the conditions in* Theorem 1 *hold. If*

$$m_1(c) \ge \log p \quad \textit{for some } c > 2, \tag{3.3}$$

*then*

$$\lim_{(n,p)\to\infty} \frac{\mathrm{FDR}}{\alpha p_0/p} \to 1 \quad \textit{and} \quad \frac{\mathrm{FDP}}{\alpha p_0/p} \to 1 \quad \textit{in probability as } (n,p)\to\infty. \tag{3.4}$$

When additional assumptions are imposed on the sparsity and strength of the signals, the condition on the correlation matrix $\boldsymbol{R}$ can be further relaxed. For example, under

$$m_1(c) \geq p^{\theta} \quad \text{for some } 0<\theta<1 \text{ and } c>2(1-\theta), \tag{3.5}$$

(C1) can be weakened as follows.

**(C1$^*$).** $$\sum_{(i,j)\in\mathcal{A}(\varepsilon)} p^{\{2(1-\theta)\rho_{ij}^*/(1+\rho_{ij}^*)\}+\delta} = O\left(\frac{p^2}{(\log p)^2}\right), \text{ for some } \varepsilon>0 \text{ and } \delta>0.$$

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

**Theorem 3.** *Suppose that $p \leq n^{\beta}$ for some $\beta>0$, $\mathbb{E}\|\boldsymbol{Y}-\boldsymbol{\mu}_Y\|^{8\beta+4+\epsilon} \leq K$, $\max_{1\leq i\leq p}\mathbb{E}|X_i-\mu_i|^{8\beta+4+\epsilon} \leq K$ 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 $c_2>0$. Under* (C1$^*$) *and* (3.5)*, we have* (3.4) *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 $\mathrm{Card}\{\mathcal{A}(\varepsilon)\} \leq p^{2-\delta}$ for some $\delta>0$. The number of pairs $(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j)$ with non-trivial correlations can be as large as $p^{2-\delta}$.

When $d=1$, Owen (2005) considered multiple tests assessing the association between $\boldsymbol{Y}$ and $X_i$ for $i=1,\ldots,p$ 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 $X_i$, $1\leq i\leq p$. 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 $\boldsymbol{X}$ is related to the number of signals. As $\theta$ in (3.5) increases, the total number of signals increases and the condition on the correlation becomes weaker.

### 3.2. Power properties

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

$$\widehat{PO} = \frac{\sum_{i\in\mathcal{H}_1} I(T_i \geq \hat{t})}{\mathrm{Card}(\mathcal{H}_1)}.$$

Suppose (3.5) holds and the magnitude of all signals in $\mathcal{H}_1$ satisfies

$$\boldsymbol{\sigma}_i' \boldsymbol{\Sigma}_{Zi}^{-1} \boldsymbol{\sigma}_i \geq \frac{c \log p}{n} \quad \text{for } i \in \mathcal{H}_1. \tag{3.6}$$

**Theorem 4.** *Suppose that* $p \leq n^\beta$ *for some* $\beta > 0$, $\mathbb{E}\|\boldsymbol{Y} - \boldsymbol{\mu}_Y\|^{8\beta+4+\epsilon} \leq K$, $\max_{1\leq i\leq p} \mathbb{E}|X_i - \mu_i|^{8\beta+4+\epsilon} \leq K$ *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* $c_2 > 0$. *If* (3.5) *and* (3.6) *hold, then*

$$\widehat{PO} \to 1 \quad \textit{in probability.}$$

Theorem 4 shows that the proposed multiple testing procedure has overwhelming power in detecting the signals whose magnitudes satisfy (3.6). The constant factor $c$ in (3.6) cannot be replaced by $o(1)$. Otherwise, it is not even possible to detect the global signals ($\boldsymbol{\sigma}_i \neq 0$ 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 $\boldsymbol{Y}$, 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. Most existing FDR control procedures require the exact $p$-values. When the number of the true signals is fixed as the dimension $p \to \infty$, Proposition 2.1 in Liu and Shao (2014) shows that the BH 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.

Compared with the BH procedure, the proposed method differs in the additional thresholding step, which is critical in controlling the FDR and FDP. To see the differences in the properties of the two procedures, we investigate the theoretical performance of the BH procedure. Let the $p$-values $\mathrm{p}_i = G(T_i)$. The BH procedure rejects $H_{0i}$ if $\mathrm{p}_i \leq \mathrm{p}_{\left(\hat{k}\right)}$, where the $p$-values $\{\mathrm{p}_i, 1 \leq i \leq p\}$ are sorted as $\mathrm{p}_{(1)} \leq \cdots \leq \mathrm{p}_{(p)}$ and $\hat{k}$ satisfies

$$\hat{k} = \max\left\{k : \ \mathrm{p}_{(k)} \leq \frac{\alpha k}{p}\right\}.$$

The BH procedure is equivalent to rejecting $H_{0i}$ if $T_i \geq \hat{t}_{BH}$, where $\hat{t}$ is defined by

$$\hat{t}_{BH} = \inf\left\{t \geq 0 : G(t) \leq \frac{\alpha \max\left\{\sum_{1\leq i\leq p} I(T_i \geq t), 1\right\}}{p}\right\}. \tag{3.7}$$

The FDR and FDP of the BH procedure are given by

$$\mathrm{FDP}_{BH} = \frac{\sum_{i\in\mathcal{H}_0} I\{T_i \geq \hat{t}_{BH}\}}{\max\left(1, \sum_{1\leq i\leq p} I\{T_i \geq \hat{t}_{BH}\}\right)} \quad \text{and} \quad \mathrm{FDR}_{BH} = \mathbb{E}(\mathrm{FDP}_{BH}).$$

We show that the BH method cannot control the FDP if the number of true alternatives $|\mathcal{H}_1|$ is fixed as $p \to \infty$. To simplify the proof, we illustrate the point in the case that $\mathbf{\Sigma} = \mathsf{Cov}(\boldsymbol{X})$ is diagonal.

**Proposition 1.** *Suppose the conditions in* Theorem 3.1 *hold and* $\mathbf{\Sigma}$ *is diagonal. If* $|\mathcal{H}_1|$ *is fixed as* $p \to \infty$*, then for any* $0 < \xi < 1$*, we have*

$$\liminf_{(n,p)\to\infty} \mathbb{P}(\mathrm{FDP}_{BH} \geq \xi) \geq \eta \tag{3.8}$$

*for some* $\eta > 0$ *which may depend on* $\xi$ *and* $|\mathcal{H}_1|$.

This proposition indicates that in the extremely sparse case, the BH method is not suitable for the control of FDP. In contrast, Theorem 3.1 shows that our procedure can still control the FDP when $|\mathcal{H}_1|$ is fixed.

When $|\mathcal{H}_1|$ goes to infinity at certain rates, we will have $\hat{t} \leq b_p$ with probability tending to one. In this case, $\hat{t}$ in (2.4) exists and our procedure is equivalent to the BH method. So the BH method can control the FDP/FDR asymptotically.

**Proposition 2.** *Suppose the conditions in* Theorem 3.1 *hold. If* (3.3) *holds, we have*

$$\lim_{(n,p)\to\infty} \frac{\mathrm{FDR}_{BH}}{\alpha p_0/p} = 1 \quad \textit{and} \quad \frac{\mathrm{FDP}_{BH}}{\alpha p_0/p} \to 1 \quad \textit{in probability as } (n,p) \to \infty.$$

When $\hat{t}_{BH} \leq b_p$, our procedure coincides with the BH procedure; if $\hat{t}_{BH} > b_p$, then our procedure rejects hypotheses with test statistics exceeding $a_p$. This threshold is necessary since the asymptotic null distribution $G(t) = \mathbb{P}\big(\chi^2_d \geq t\big)$ may not approximate $p_0^{-1}\sum_{i\in\mathcal{H}_0} I(T_i \geq t)$ sufficiently well for large $t$ in that $p_0^{-1}\sum_{i\in\mathcal{H}_0} I(T_i \geq t)/G(t) \nrightarrow 1$ when $t \geq b_p + \log(\log p)$. As shown in our simulation studies, the probability of $\hat{t}_{BH} > b_p$ approaches 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 BH procedure fails under the PheWAS setting. As shown in Theorem 3.1, our thresholding rule based procedure adaptively controls the FDR and FDP without prior knowledge of the degree of sparsity.

## 4. Numerical Results

We investigated the numerical performance of the proposed multiple testing procedure through simulation studies. Numerical comparison with alternative

methods is given. The proposed procedure was applied to an RA cohort to comprehensively evaluate how two important genetic markers for auto-immune diseases can 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 $\boldsymbol{Y}$, considered $p = 500, 1{,}000$ and $2{,}000$ for $\boldsymbol{X}$, and the sample size $n = 100$ and 150. Under each configuration, the results are summarized based on 500 simulated datasets for FDR estimates and 100 simulated datasets for power calculations. For each dataset, we performed testing based on our methods and several existing methods In addition to to the BH, BY methods, we compared to the Storey, Taylor and Siegmund (2004) (JS) procedure, the Romano and Shaikh (2006b) step-up procedure controlling FDP at 50% with probability at least $1-\alpha$ (RSuFDP), the Romano and Shaikh (2006a) step-down procedure controlling FDP at 50% with probability at least $1-\alpha$ (RSdFDP) and controlling FDR at $\alpha$ (RSdFDR), as well as the Guo and Rao (2008) procedure (GR).

We generated the entire data vector $\boldsymbol{W} := (Y_1, \ldots, Y_4, X_1, \ldots, X_p)'$ from $\boldsymbol{W} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\varepsilon}$, with two settings of $\boldsymbol{\Sigma}$, as described below, were chosen to reflect different correlation structures, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{p+4})'$ are independent and identically distributed. The distribution of $\varepsilon_i$ is taken to be (i) the standard normal $N(0,1)$; (ii) the exponential with mean 1; and (iii) a mixture of $N(-1, 0.5^2)$ with probability 0.1 and $N(0, 0.5^2)$ with probability 0.9. As our test statistics are invariant to the variances, we take the diagonal entries of $\boldsymbol{\Sigma}$ to be 1. Let $\boldsymbol{\Sigma}_1 = (\sigma_{ij1}) \in \mathbb{R}^{(p_1+4)\times(p_1+4)}$ and $\boldsymbol{\Sigma}_2 = (\sigma_{ij2}) \in \mathbb{R}^{(p-p_1)\times(p-p_1)}$. In all models, we let $\boldsymbol{\Sigma} = \operatorname{diag}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$ and $\sigma_{ij1} = \{(2+\delta)\log p/n\}^{I(i\neq j)/2}$.

**Model 1:** $\sigma_{ij2} = \wp^{|j-i|}$. Here $\wp = 0.5$ in Model (1A) and $\wp = 0.8$ in Model (1B).

**Model 2:** $\boldsymbol{\Sigma}_2 = \operatorname{diag}(D_1, \ldots, D_m, I)$, where $m = \{(p-p_1)/10\}$, $D_k \in \mathbb{R}^{10\times 10}$, $1 \le k \le m$. I is the identity matrix. All off-diagonal entries of $D_k$ are taken to be $\wp$. We let $\wp = 0.8$ in Model (2A) and $\wp = 0.5$ in Model (2B).

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

We examined the probability that $\hat{t}$ in (2.4) exists under different settings, which reflects the degree to which our procedure differs from the BH procedure. Figure 1 summarizes the frequency that $\hat{t}$ in (2.4) exists when $\delta = -0.5$, $\alpha = 0.05$

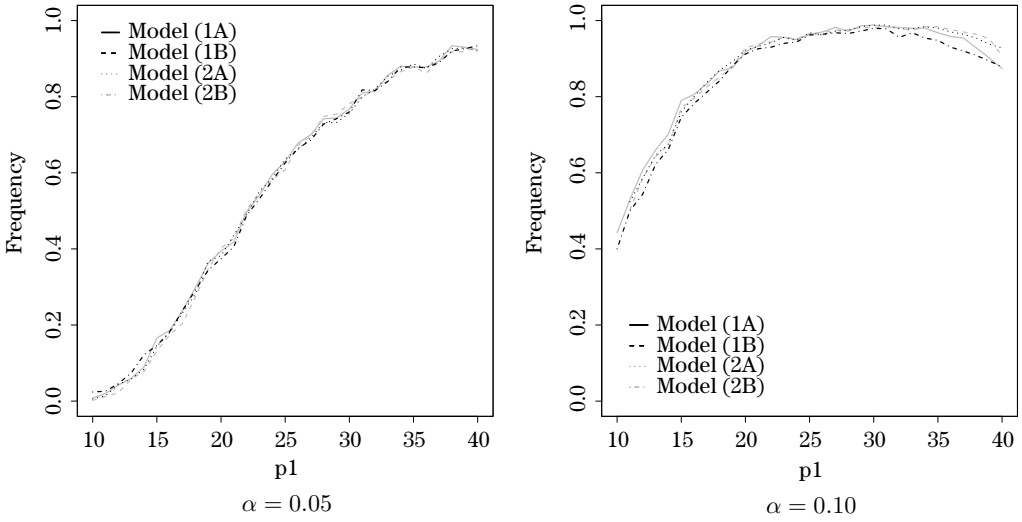


Figure 1. The frequency that $\hat{t}$ exists when $n = 100$, $p = 1{,}000$ with normal error.

and 0.10, $n = 100$ and $p = 1{,}000$. The probability that $\hat{t}$ does not exist approaches 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 BH procedure under sparse settings.

We next investigated the performance of various procedure in controlling for the FDR in finite samples. Figures 2 and 3 present the empirical FDR for $\alpha = 0.05$ and 0.1when $\delta = -0.5$ with various choices of $p$, $p_1$ and $n$. The results show that the proposed procedures maintain the desired FDR levels well when $p_1 = 30$. For $p_1 = 10$, our method also generally maintains the FDR level but with slight inflation for a few scenarios when $n = 100$. On the other hand, the BH procedure tends to result in inflated FDR under a variety of settings and the inflation can be quite substantial for some scenarios. The JS procedure, requiring an estimate of the null proportion, tends to give substantially inflated FDR, due to the inaccurate estimate of the null proportion. All other procedures, including the BY and FDP controlling procedures are able to control the FDR but tend to be conservative with empirical FDR substantially lower than the target levels. For example, when $p = 2{,}000$, $p_1 = 10$, $n = 100$, and $\alpha = 0.1$, the empirical FDR was 0.110, 0.111 and 0.068 for the proposed procedure, and 0.144, 0.151 and 0.097 for the BH procedure, under Model (1A) with mixture, normal and exponential error distributions, respectively. To better preserve the FDR, our procedure also

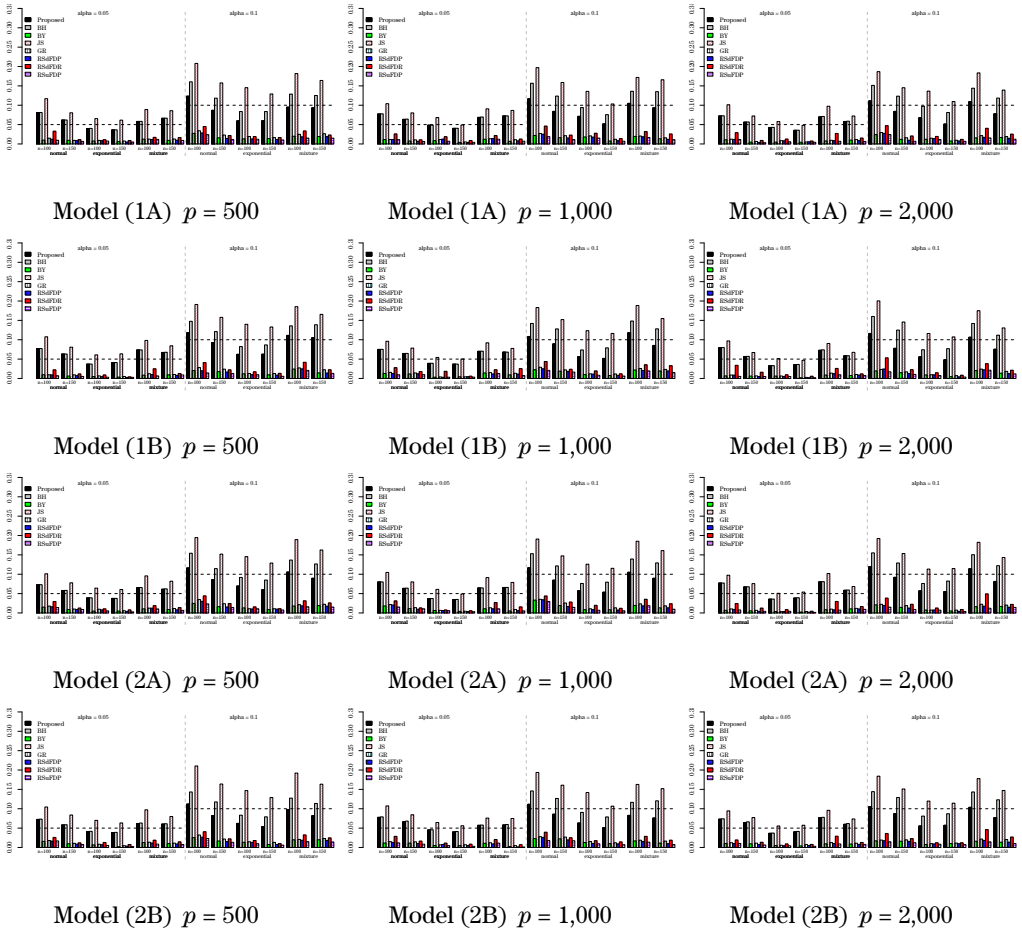


Figure 2. Empirical FDR of various procedures, including the proposed, BH, BY, Storey, Taylor and Siegmund (2004) (JS), Guo and Rao (2008) (GR), the Romano and Shaikh (2006b) step-up procedure controlling FDP at 50% with probability at least $1-\alpha$ (RSuFDP), the Romano and Shaikh (2006a) step-down procedure controlling FDP at 50% with probability at least $1-\alpha$ (RSdFDP) as well as the Romano and Shaikh (2006a) step-down FDR controlling procedure (RSdFDR), under the very sparse setting with $p_1 = 10$.

demonstrates superior performance in maintaining the FDP when compared to the BH and the JS procedure, as illustrated in Figure 4.

We now turn to the comparison of powers between our procedure and existing procedures that are robust to the dependence structure. We varied the signal strength by letting $\delta$ in $\sigma_{ij1}$ vary from $-0.5$ to 1. Figure 5 summarizes the empirical power of various procedures over different values of $\delta$ when $n = 100$,

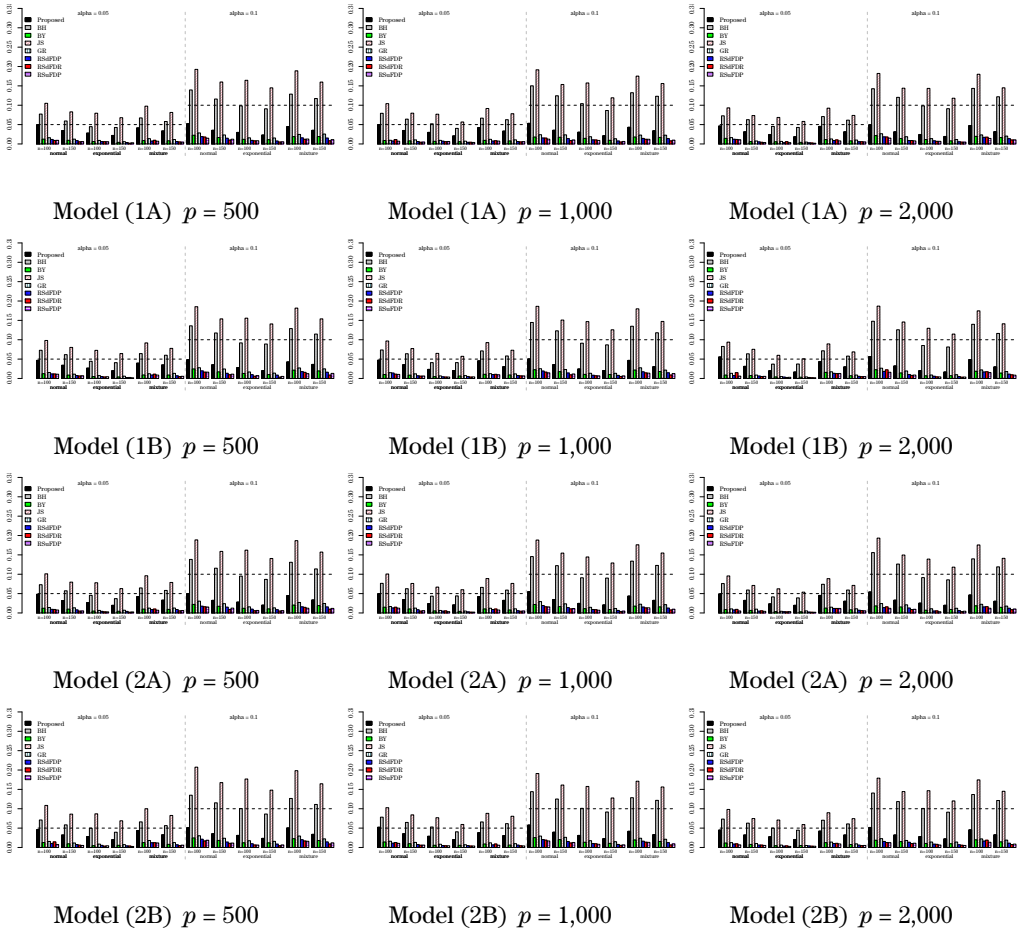


Figure 3. Empirical FDR of various procedures, including the proposed, BH, BY, Storey, Taylor and Siegmund (2004) (JS), Guo and Rao (2008) (GR), the Romano and Shaikh (2006b) step-up procedure controlling FDP at 50% with probability at least $1 - \alpha$ (RSuFDP), the Romano and Shaikh (2006a) step-down procedure controlling FDP at 50% with probability at least $1-\alpha$ (RSdFDP) as well as the Romano and Shaikh (2006a) step-down FDR controlling procedure (RSdFDR), under the moderately sparse setting with $p_1 = 30$.

$p = 2{,}000$ and $\alpha = 0.1$. Across all settings, the power of our procedure is substantially higher than that of competing methods, with the advantage even more significant under the very sparse setting of $p_1 = 10$.

These numerical results are consistent with the theoretical analysis discussed earlier. Both the BH and the Storey procedures fail to control for the FDR or FDP due to the complex correlation structure and the estimated null distribution

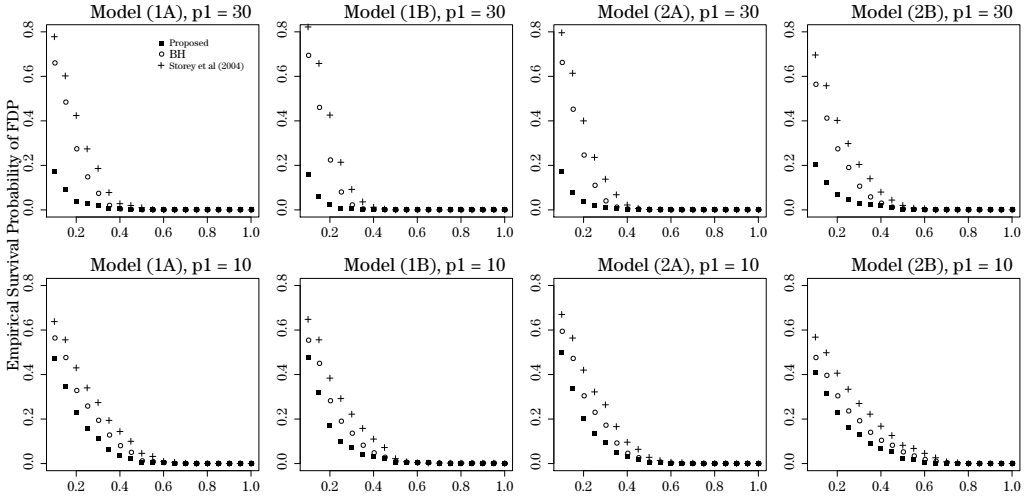


Figure 4. Empirical survival distribution of the FDP under different models with $p_1 = 30$ or 10 when $n = 100$, $p = 1{,}000$, $\alpha = 0.1$ and the error is normal.

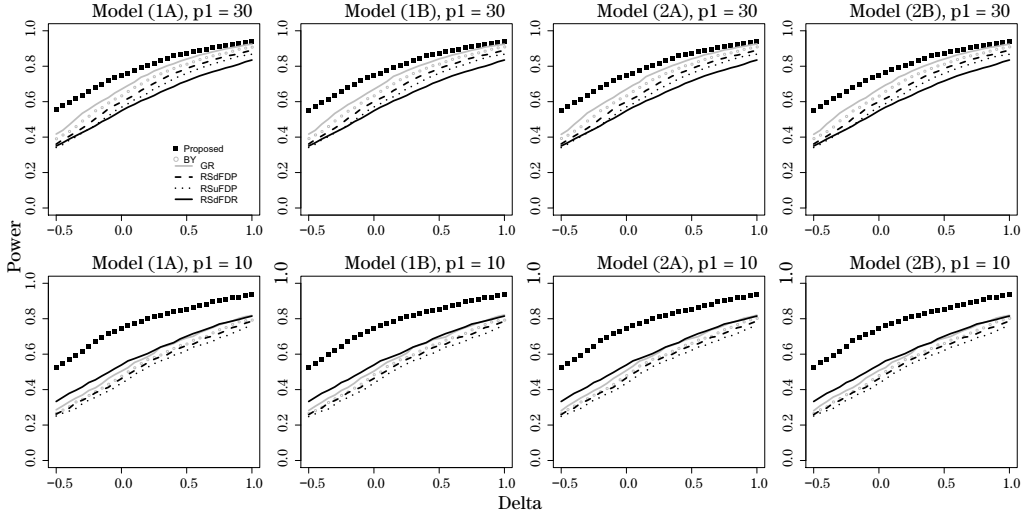


Figure 5. Comparing power of different procedures under different models with $p_1 = 30$ or 10 when $n = 100$, $p = 2{,}000$, $\alpha = 0.1$ and the error is normal.

under the sparse setting. On the other hand, the BY procedure along with other competing FDR and FDP controlling procedures are overly conservative with substantially lower power compared to our method.

### 4.2. Application to PheWAS of autoimmune risk alleles with EMR

We applied our procedure to an EMR cohort of 1837 RA subjects (Liao et al. (2010)). Blood samples were collected for these RA cases using the BWH Specimen Bank from 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 1,237 individuals of European ancestry, 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 were interested in conducting a PheWAS using these two SNPs to determine whether carriage of these risk alleles are associated with RA comorbidities. 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 each of the PheWAS codes, the value of the variable represents total count of ICD9 codes a patient received across all 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 $\boldsymbol{X}$ for the association analysis. The 2-dimensional vector $\boldsymbol{Y}$ 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 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 here that all 4 phenotypes are related to severity of the disease. Patients with more severe RA tend 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 con-

firmed 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. When we applied the BY method to this example, only RA was deemed as significant at FDR level of 0.05. This again demonstrates that our proposed method is more powerful than the BY method.

## 5. Discussion

We introduced in this paper a multiple testing procedure for simultaneously testing columns of high-dimensional cross-covariance matrices. There is an important difference between our FDR control procedure given in (2.4) and the well-known BH procedure. When the test statistics $T_i$ are used and the $p$-values are calculated from $\mathbb{P}(\chi^2_d \geq t)$, the BH procedure is equivalent to rejecting $H_{0i}$ whenever $T_i \geq \hat{t}_{BH}$. For our procedure, 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. Thus, for $t > b_p$, we do not use $pG(t)$ to estimate $\hat{R}_0(t) \equiv \sum_{i\in\mathcal{H}_0} I\{T_i \geq t\}$ while the BH method does. 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 (2.4) 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, we 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 controlling FWER. For general fixed $p_1$, controlling FDP at level $1/(p_1+1)$ is essentially equivalent to controlling FWER. When $p_1$ is fixed as $p \to \infty$, $\hat{t}$ in (2.4) 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, Liu and Shao (2014) showed that the BH method is unable to control FDP

at any level $0 < \alpha < 1$. Specifically, consider $p$ tests with independent $p$-values $\mathcal{P}_1, \ldots, \mathcal{P}_p$. If $\min_{i \in \mathcal{H}_1} \mathcal{P}_i = o_P(p^{-1})$, then for any $0 < \alpha < 1$, there exists $c_0 > 0$ such that $\liminf_{p \to \infty} \mathbb{P}(\text{FDP} \geq \alpha) \geq c_0$.

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 $X_i$. This 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.

Our procedure tests for correlatedness between $\boldsymbol{Y}$ and $X_i$'s, which is an easier task than testing for independence, especially in the setting of high dimension and low sample size. In the Gaussian case, uncorrelatedness is equivalent to independence. When data are not Gaussian, we can take transformations of the data prior to testing such that the transformed data are approximately normal. The covariance testing is also valid for detecting dependency under other models. For example, if

$$h(X_i) = \boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{Y} + \epsilon_i, \epsilon_i \perp \boldsymbol{Y}, h(\cdot) \text{ strictly increasing}$$

then it is not difficult to show that

$$\boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{\sigma} = cov(\boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{Y}, X_i) = cov(h(X_i), X_i) = E[\{h(X_i) - h(\mu_i)\}(X_i - \mu_i)] > 0$$

under mild conditions on the distribution of $X_i$. Thus, our test based on $\boldsymbol{\sigma}_i$ can in fact detect non-linear relationships although proper transformation may increase power. Future research is warranted to test for more complex non-linear associations in the high dimensional setting.

## Supplementary Materials

The proofs of our theoretical results are shown in the Supplementary Materials.

## Acknowledgment

This research was supported in part by NIH grants U54 H6007963, R01 CA127334, and K08AR060257, NSF Grants DMS-1208982 and DMS-1403708, NSFC Grant No.11201298, No.11 322107 and No.11431006, Program for Profes-

sor 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, 973 Program (2015CB856004) and a grant from Australian Research Council.

## References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B Statistical Methodology* **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165–1188.

Bush, W. S. and Moore, J. H. (2012). Genome-wide association studies. *PLoS Computational Biology* **8**, e1002822.

Criswell, L. A., Pfeiffer, K. A., Lum, R. F., Gonzales, B., Novitzke, J., Kern, M., Moser, K. L., Begovich, A. B., Carlton, V. E., Li, W., Lee, A. T., Ortmann, W., Behrens, T. W. and Gregersen, P. K. (2005). Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) collection: the PTPN22 620w allele associates with multiple autoimmune phenotypes. *The American Journal of Human Genetics* **76**, 561–571.

Davidson, A. and Diamond, B. (2001). Autoimmune diseases. *The New England Journal of Medicine* **345**, 340–50.

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M. and Crawford, D. C. (2010). Phewas: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**, 1205–1210.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics* **32**, 962–994.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistics Association* **99**, 96–104.

Efron, B. (2007). Size, power, and false discovery rates. *The Annals of Statistics* **35**, 1351–1377.

Fan, J. and Han, X. (2017). Estimation of false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society B Statistical Methodology* **79**, 1143–1164.

Fan, J., Han, X. and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association* **107**, 1019–1035.

Fang, H., Xie, Q., Boneva, R., Fostel, J., Perkins, R. and Tong, W. (2006). Gene expression profile exploration of a large dataset on chronic fatigue syndrome. *Pharmacogenomics* **7**, 429–440.

Guo, W. and Rao, M. B. (2008). On control of the false discovery rate under no assumption of dependency. *Journal of Statistical Planning and Inference* **138**, 3176–3188.

Hall, M. A., Verma, A., Brown-Gentry, K. D., Goodloe, R., Boston, J., Wilson, S., McClellan, B., Sutcliffe, C., Dilks, H. H., Gillani, N. B., Jin, H., Mayo, P., Allen, M., Schnetz-Boutaud, N., Crawford, D. C., Ritchie, M. D. and Pendergrass, S. A. (2014). Detection of pleiotropy through a Phenome-wide association study (PheWAS) of epidemiologic data as part of the environmental architecture for genes linked to environment (eagle) study. *PLoS Genetics*

**10**, e1004678.

Jaraquemada, D., Ollier, W., Awad, J., Young, A. and Festenstein, H. (1986). HLA and rheumatoid arthritis: susceptibility or severity? *Disease markers* **4**, 43–53.

Katagiri, F. and Glazebrook, J. (2009). Overview of mRNA expression profiling using DNA microarrays. *Current Protocols in Molecular Biology*, 22–4.

Kurreeman, F., Liao, K., Chibnik, L., Hickey, B., Stahl, E., Gainer, V., Li, G., Bry, L., Mahan, S., Ardlie, K., Thomson, B., Szolovits, P., Churchill, S., Murphy, S. N., Cai, T., Raychaudhuri, S., Kohane, I., Karlson, E. and Plenge, R. M. (2011). Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *The American Journal of Human Genetics* **88**, 57–69.

Liao, K. P., Cai, T., Gainer, V., Goryachev, S., Zeng-treitler, Q., Raychaudhuri, S., Szolovits, P., Churchill, S., Murphy, S., Kohane, I., Karlson, E. W. and Plenge, R. M. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care & Research* **62**, 1120–1127.

Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W.,Ananthakrishnan, A. N., Gainer, V. S., Shaw, S. Y., Xia, Z., Szolovits, P., Churchill, S. and Kohane, I. (2015). Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *British Medical Journal* **350**, h1885.

Liao, K. P., Kurreeman, F., Li, G., Duclos, G., Murphy, S., Guzman, R., Cai, T., Gupta, N., Gainer, V., Schur, P. and et al. (2013). Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non–rheumatoid arthritis controls. *Arthritis & Rheumatism* **65**, 571–581.

Liu, W. and Shao, Q.-M. (2014). Phase transition and regularized bootstrap in large scale $t$-tests with false discovery rate control. *The Annals of Statistics* **42**, 2003–25.

Masson, C. (2011). Rheumatoid anemia. *Joint Bone Spine* **78**, 131–137.

Nelson, P. T., Baldwin, D. A., Scearce, L. M., Oberholtzer, J. C., Tobias, J. W. and Mourelatos, Z. (2004). Microarray-based, high-throughput gene expression profiling of microRNAS. *Nature Methods* **1**, 155–161.

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society B Statistical Methodology* **67**, 411–426.

Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., Ambite, J. L., Avery, C. L., Buyske, S., Bůžková, P., Deelman, E., Fesinmeyer, M. D., Haiman, C. A., Heiss, G., Hindorff, L. A., Hsu, C-N, Jackson, R. D., Kooperberg, C., Le Marchand, L., Lin, Y., Matise, T. C., Monroe, K. R., Moreland, L., Park, S. L., Reiner, A., Wallace, R., Wilkens, L. R., Crawford, D. C. and Ritchie, M. D. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the population architecture using genomics and epidemiology (PAGE) network. *PLoS Genetics* **9**, e1003087.

Romano, J. P. and Shaikh, A. M. (2006a). On stepdown control of the false discovery proportion. In Optimality, 33–50. Institute of Mathematical Statistics.

Romano, J. P. and Shaikh, A. M. (2006b). Stepup procedures for control of generalizations of the familywise error rate. *The Annals of Statistics*, 1850–1873.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. and Smoller, J. W. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics* **14**, 483–495.

Storey, J., Taylor, J. and Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach.

*Journal of the Royal Statistical Society, Series B Statistical Methodology* **66**, 187–205.

Weyand, C. M., Hicok, K. C., Conn, D. L. and Goronzy, J. J. (1992). The influence of HLA-DRB1 genes on disease severity in rheumatoid arthritis. *Annals of Internal Medicine* **117**, 801–806.

Xia, Y., Cai, T. and Cai, T. T. (2017). Multiple testing of submatrices of a precision matrix with applications to identification of between pathway interactions. *Journal of the American Statistical Association*, `https://doi.org/10.1080/01621459.2016.1251930.`

Department of Biostatistics, Harvard School of Public Health, 655 Huntington Ave, Boston MA 02115 U.S.A.

E-mail: tcai@hsph.harvard.edu

Department of Statistics, The Wharton School, University of Pennsylvania, 3730 Walnut St, Philadelphia, PA 19104, USA.

E-mail: tcai@wharton.upenn.edu

Brigham and Women's Hospital, Department of Medicine, Rheumatology, Immunology, 75 Francis Street, Boston, MA 02115, USA.

E-mail: kliao@partner.org

Department of Mathematics, Institute of Natural Sciences and MOE-LSC, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai, China 200240.

E-mail: weidongl@sjtu.edu.cn

(Received April 2017; accepted November 2017)