Biometrics 73, 1300–1310
December 2017

DOI: 10.1111/biom.12695

Simulation-Based Hypothesis Testing of High Dimensional Means Under Covariance Heterogeneity

Jinyuan Chang,[1,*] Chao Zheng,[2,] Wen-Xin Zhou,[3,***] and Wen Zhou[4,****]**

[1]School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China
[2]School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia
[3]Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544, U.S.A.
[4]Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, U.S.A.
*email: changjinyuan@swufe.edu.cn
**email: zhengc1@student.unimelb.edu.au
***email: wenxinz@princeton.edu
****email: riczw@stat.colostate.edu

Summary. In this article, we study the problem of testing the mean vectors of high dimensional data in both one-sample and two-sample cases. The proposed testing procedures employ maximum-type statistics and the parametric bootstrap techniques to compute the critical values. Different from the existing tests that heavily rely on the structural conditions on the unknown covariance matrices, the proposed tests allow general covariance structures of the data and therefore enjoy wide scope of applicability in practice. To enhance powers of the tests against sparse alternatives, we further propose two-step procedures with a preliminary feature screening step. Theoretical properties of the proposed tests are investigated. Through extensive numerical experiments on synthetic data sets and an human acute lymphoblastic leukemia gene expression data set, we illustrate the performance of the new tests and how they may provide assistance on detecting disease-associated gene-sets. The proposed methods have been implemented in an R-package HDtest and are available on CRAN.

Key words: Feature screening; High dimension; Hypothesis testing; Normal approximation; Parametric bootstrap; Sparsity.

1. Introduction

The problems of comparing a particular sample to a hypothetical population with known prior information or comparing two parallel groups, such as a control group and a treatment group, have both important applications in modern genomics and bio-medical research and become the foundation of scientific discoveries. They have been employed widely for identifying biologically interesting gene-sets for drug design, evolutionary studies, and mutation detection. Our interests in these problems are motivated by a microarray study on human acute lymphoblastic leukemia (Chiaretti et al., 2004). This study consists of 75 patients of B-lymphocyte type leukemia, who were classified into two groups: 35 patients with BCR/ABL fusion and 40 patients with cytogenetically normal NEG. It is known that genes tend to work collectively in groups to achieve certain biological tasks. Our analysis focuses on such groups of genes (gene sets) defined with the gene ontology (GO) framework, which are referred to as GO terms. Identifying disease-relevant GO terms based on their average expression levels provides information on differential gene pathways associated with the leukemia. Many GO terms contain a large number of (in the data, as many as 3145) genes with very complex gene-wise dependence structures. The large dimension of data and the complex dependency among genes make the problem of comparing population means extremely challenging.

Let $\mathbf{X}$ and $\mathbf{Y}$ be two p-dimensional random vectors with means $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})^{\mathrm{T}}$ and $\boldsymbol{\mu}_2 = (\mu_{21}, \ldots, \mu_{2p})^{\mathrm{T}}$, covariance matrices $\boldsymbol{\Sigma}_1 = (\sigma_{1,k\ell})_{1\le k,\ell\le p}$ and $\boldsymbol{\Sigma}_2 = (\sigma_{2,k\ell})_{1\le k,\ell\le p}$, respectively. It is then of general interest in testing the hypotheses

- (One-sample problem) $H_0^{(\mathrm{I})} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_0$ versus $H_1^{(\mathrm{I})} : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_0$ for a specified p-dimensional vector $\boldsymbol{\mu}_0$, which, without loss of generality, is equivalent to

$$H_0^{(\mathrm{I})} : \boldsymbol{\mu}_1 = \mathbf{0} \quad \text{versus} \quad H_1^{(\mathrm{I})} : \boldsymbol{\mu}_1 \neq \mathbf{0}; \tag{1.1}$$

- (Two-sample problem)

$$H_0^{(\mathrm{II})} : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1^{(\mathrm{II})} : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{1.2}$$

When p is fixed, traditional tests have been extensively studied for testing both (1.1) and (1.2). For example, the properties for both the one-sample and two-sample Hotelling's T^2 tests have been examined under normality assumption (Anderson, 2003). We refer to Liu and Shao (2013) for a moderate deviation result in the absence of normality.

Generally, the sum of squares-type and the maximum-type statistics are used to test the hypotheses (1.1) and (1.2) in

the high dimensional settings. The sum of squares-type statistics aim to mimic the weighted Euclidean norms, $|\mathbf{A}\boldsymbol{\mu}_1|_2^2$ or $|\mathbf{A}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)|_2^2$ for certain linear transformation $\mathbf{A}$, and the corresponding tests are powerful for detecting relatively dense signals (Bai and Saranadasa, 1996; Chen and Qin, 2010). Statistics of the maximum-type, on the other hand, are preferable for detecting relatively sparse signals (Cai et al., 2014) and have been used in a variety of applications including the medical image problem (James et al., 2001) and gene selections (Martens et al., 2005).

Most existing testing procedures for (1.1) and (1.2) rely on the derivation the pivotal limiting distribution of test statistics, from which the critical value is approximated. In the high dimensional scenarios, various structural assumptions on the unknown covariance matrices have been imposed (Zhong et al., 2013; Cai et al., 2014). However, in many applications, these assumptions can be very restrictive or difficult to be verified, and therefore limit the scope of applicability for the limiting distribution calibration approach. First, the existence of a pivotal asymptotic distribution relies heavily on the structural assumptions on the unknown covariance/correlation structures, which may not be true in practice. For example, it is very common that the expression levels are highly correlated for genes regulated by the same pathway (Wolen and Miles, 2012) or associated with the same functionality (Katsani et al., 2014), which results in a complex and non-sparse covariance structure. These empirical evidences indicate that the strong structural assumptions on the covariance matrices may sometimes be unrealistic in real-world applications. Another concern, as pointed out by Cai et al. (2014), is that the convergence rate to the extreme value distribution of maximum-type statistics is usually slow. Taking the extreme distribution of type I as an example, the convergence rate is of order $O\{\log(\log n)/\log(n)\}$. Although the convergence rate may be improved by using suitable intermediate approximations, still its validity relies on the dependence structure of the underlying distribution.

Driven by the above two concerns, we revisit the problem of testing hypotheses (1.1) and (1.2) from a different perspective. Motivated by applications in genomic analysis and image analysis, we are particularly interested in detecting discrepancies when $\boldsymbol{\mu}_1$ and $\mathbf{0}$ or $\boldsymbol{\mu}_2$ are distinguishable to a certain extent in at least one coordinate. We develop a fully data driven procedure to compute the critical values using the Monte Carlo simulations. The validity of our procedure is established without enforcing structural assumptions of any kind on the unknown covariances. The main idea is based on the approximation of empirical processes by Gaussian processes (Chernozhukov et al., 2013), and to some degree, is similar to that of Liu and Shao (2013) that utilizes the intermediate approximation. However, instead of generating independent standard multivariate normal vectors, our approach takes into account correlations among the features and therefore is automatically adapted to the underlying dependence.

The rest of the article is organized as follows. In Section 2, we describe the simulation-based testing procedures for both hypotheses (1.1) and (1.2). Theoretical properties of the tests are studied in Section 3. Numerical studies are reported in Section 4 to assess the performance of the proposed tests comparing to the peer methods. In Section 5, we applied the proposed tests to the acute lymphoblastic leukemia data for identifying disease-associated gene-sets based on the gene expression levels. The underpinning technical details, as well as additional simulation results and empirical data analysis, are relegated to the supplementary material.

2. Methodology

Throughout the article, we denote by $|\boldsymbol{\beta}|_\infty=\max_{1\le k\le p}|\beta_k|$ for a p-dimensional vector $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^{\mathrm{T}}$. For a matrix $\mathbf{A}=(a_{k\ell})_{p\times p}$, define $|\mathbf{A}|_\infty=\max_{1\le k,\ell\le p}|a_{k\ell}|$. Let $\mathbf{D}_1=\mathrm{diag}\,(\boldsymbol{\Sigma}_1)$ and $\mathbf{D}_2=\mathrm{diag}\,(\boldsymbol{\Sigma}_2)$. Denote by $\mathbf{R}_1$ and $\mathbf{R}_2$ the corresponding correlation matrices. Let $\mathcal{X}_n=\{\mathbf{X}_1,\ldots,\mathbf{X}_n\}$ and $\mathcal{Y}_m=\{\mathbf{Y}_1,\ldots,\mathbf{Y}_m\}$ be two independent samples consisting of independent and identically distributed (i.i.d.) observations drawn from the distributions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Let $N=n+m$. For each $i=1,\ldots,n$ and $j=1,\ldots,m$, write $\mathbf{X}_i=(X_{i1},\ldots,X_{ip})^{\mathrm{T}}$ and $\mathbf{Y}_j=(Y_{j1},\ldots,Y_{jp})^{\mathrm{T}}$.

2.1. *Test Procedures*

2.1.1. One-sample case. Consider the maximum-type statistics in the following forms:

$$T_{\mathrm{ns}}^{(\mathrm{I})}=\max_{1\le k\le p}\sqrt{n}|\bar{X}_k| \qquad \text{or} \qquad T_{\mathrm{s}}^{(\mathrm{I})}=\max_{1\le k\le p}\frac{\sqrt{n}|\bar{X}_k|}{\hat{\sigma}_{1k}}, \tag{2.1}$$

where $\bar{X}_k=n^{-1}\sum_{i=1}^n X_{ik}$ and $\hat{\sigma}_{1k}^2=n^{-1}\sum_{i=1}^n(X_{ik}-\bar{X}_k)^2$. Throughout, the statistic $T_{\mathrm{s}}^{(\mathrm{I})}$ is referred as the *studentized* statistic, while $T_{\mathrm{ns}}^{(\mathrm{II})}$ is referred as the *non-studentized* statistic. Intuitively, large values of $T_{\mathrm{ns}}^{(\mathrm{I})}$ or $T_{\mathrm{s}}^{(\mathrm{I})}$ provide evidences against $H_0^{(\mathrm{I})}$ in (1.1) so that the corresponding tests are of the form $\Psi_{\mathrm{ns},\alpha}^{(\mathrm{I})}=I\{T_{\mathrm{ns}}^{(\mathrm{I})}>\mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{I})}\}$ or $\Psi_{\mathrm{s},\alpha}^{(\mathrm{I})}=I\{T_{\mathrm{s}}^{(\mathrm{I})}>\mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{I})}\}$, where $\mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{I})}$ and $\mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{I})}$ are the critical values.

Under the null hypothesis $H_0^{(\mathrm{I})}:\boldsymbol{\mu}_1=\mathbf{0}$, we motivate from the multivariate central limit theorem with fixed p to calculate critical values $\mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{I})}$ and $\mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{I})}$ as follows: let $\widetilde{\boldsymbol{\Sigma}}_1$ be an estimate of $\boldsymbol{\Sigma}_1$ from the sample $\mathcal{X}_n$, and set $\widetilde{\mathbf{R}}_1=\widetilde{\mathbf{D}}_1^{-1/2}\widetilde{\boldsymbol{\Sigma}}_1\widetilde{\mathbf{D}}_1^{-1/2}$ with $\widetilde{\mathbf{D}}_1=\mathrm{diag}\,(\widetilde{\boldsymbol{\Sigma}}_1)$. Given $\mathcal{X}_n$, let $\mathbf{W}_{\mathrm{ns}}^{(\mathrm{I})}\sim\mathrm{N}(\mathbf{0},\widetilde{\boldsymbol{\Sigma}}_1)$ and $\mathbf{W}_{\mathrm{s}}^{(\mathrm{I})}\sim\mathrm{N}(\mathbf{0},\widetilde{\mathbf{R}}_1)$ be two Gaussian random vectors, the critical values can be computed by $\mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{I})}=\inf\{t\in\mathbb{R}:\mathbb{P}(|\mathbf{W}_{\mathrm{ns}}^{(\mathrm{I})}|_\infty>t\,|\mathcal{X}_n)\le\alpha\}$ and $\mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{I})}=\inf\{t\in\mathbb{R}:\mathbb{P}(|\mathbf{W}_{\mathrm{s}}^{(\mathrm{I})}|_\infty>t\,|\mathcal{X}_n)\le\alpha\}$. Practically, let $\{\mathbf{W}_{\mathrm{ns},\ell}\}_{\ell=1}^M\overset{\text{i.i.d.}}{\sim}\mathrm{N}(\mathbf{0},\widetilde{\boldsymbol{\Sigma}}_1)$ and $\{\mathbf{W}_{\mathrm{s},\ell}\}_{\ell=1}^M\overset{\text{i.i.d.}}{\sim}\mathrm{N}(\mathbf{0},\widetilde{\mathbf{R}}_1)$. Then, $\mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{I})}$ and $\mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{I})}$ can be estimated by $\widehat{\mathrm{cv}}_{\mathrm{ns},\alpha}^{(\mathrm{I})}=\inf\{t\in\mathbb{R}:\widehat{F}_{\mathrm{ns},M}^{(\mathrm{I})}(t)\ge1-\alpha\}$ and $\widehat{\mathrm{cv}}_{\mathrm{s},\alpha}^{(\mathrm{I})}=\inf\{t\in\mathbb{R}:\widehat{F}_{\mathrm{s},M}^{(\mathrm{I})}(t)\ge1-\alpha\}$, where $\widehat{F}_{\mathrm{ns},M}^{(\mathrm{I})}(t)=M^{-1}\sum_{\ell=1}^M I\{|\mathbf{W}_{\mathrm{ns},\ell}|_\infty\le t\}$ and $\widehat{F}_{\mathrm{s},M}^{(\mathrm{I})}(t)=M^{-1}\sum_{\ell=1}^M I\{|\mathbf{W}_{\mathrm{s},\ell}|_\infty\le t\}$. For $\nu\in\{\mathrm{ns},\mathrm{s}\}$, the empirical version of test $\Psi_{\nu,\alpha}^{(\mathrm{I})}$ is therefore defined by

$$\widehat{\Psi}_{\nu,\alpha}^{(\mathrm{I})}(M)=I\{T_\nu^{(\mathrm{I})}>\widehat{\mathrm{cv}}_{\nu,\alpha}^{(\mathrm{I})}\}, \tag{2.2}$$

such that the null hypothesis $H_0^{(\mathrm{I})}$ is rejected whenever $\widehat{\Psi}_{\nu,\alpha}^{(\mathrm{I})}(M)=1$. The proposed testing procedures are fully data driven and easily computed. In Section 2.2, we discuss the constructions of $\widetilde{\boldsymbol{\Sigma}}_1$, from which the wide applicability of the test (2.2) will be explored.

2.1.2. Two-sample case. The above procedures can be naturally extended to deal with the two-sample problem (1.2). Analogously to (2.1), we define the non-studentized and studentized test statistics by $T_{\mathrm{ns}}^{(\mathrm{II})} = \max_{1\le k\le p}\sqrt{nm}|\bar{X}_k - \bar{Y}_k|/\sqrt{n+m}$ and $T_{\mathrm{s}}^{(\mathrm{II})} = \max_{1\le k\le p}\sqrt{nm}|\bar{X}_k - \bar{Y}_k|/(m\hat{\sigma}_{1k}^2 + n\hat{\sigma}_{2k}^2)^{1/2}$, respectively, where $\bar{X}_k = n^{-1}\sum_{i=1}^n X_{ik}$, $\bar{Y}_k = m^{-1}\sum_{j=1}^m Y_{jk}$, $\hat{\sigma}_{1k}^2 = n^{-1}\sum_{i=1}^n (X_{ik}-\bar{X}_k)^2$, and $\hat{\sigma}_{2k}^2 = m^{-1}\sum_{j=1}^m (Y_{jk}-\bar{Y}_k)^2$. For nominal significance level α, we define tests of the form $\Psi_{\mathrm{ns},\alpha}^{(\mathrm{II})} = I\{T_{\mathrm{ns}}^{(\mathrm{II})} > \mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{II})}\}$ or $\Psi_{\mathrm{s},\alpha}^{(\mathrm{II})} = I\{T_{\mathrm{s}}^{(\mathrm{II})} > \mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{II})}\}$ with appropriate critical values $\mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{II})}$ and $\mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{II})}$. Let $\widetilde{\boldsymbol{\Sigma}}_1$ and $\widetilde{\boldsymbol{\Sigma}}_2$ be estimates of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Define

$$\widetilde{\boldsymbol{\Sigma}}_{1,2} = \frac{m}{N}\widetilde{\boldsymbol{\Sigma}}_1 + \frac{n}{N}\widetilde{\boldsymbol{\Sigma}}_2, \quad \widetilde{\mathbf{D}}_{1,2} = \mathrm{diag}\left(\widetilde{\boldsymbol{\Sigma}}_{1,2}\right),$$
$$\widetilde{\mathbf{R}}_{1,2} = \widetilde{\mathbf{D}}_{1,2}^{-1/2}\widetilde{\boldsymbol{\Sigma}}_{1,2}\widetilde{\mathbf{D}}_{1,2}^{-1/2}, \tag{2.3}$$

and let $\{\mathbf{W}_{\mathrm{ns},\ell}\}_{\ell=1}^M \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_{1,2})$ and $\{\mathbf{W}_{\mathrm{s},\ell}\}_{\ell=1}^M \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \widetilde{\mathbf{R}}_{1,2})$. Then, $\mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{II})}$ and $\mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{II})}$ can be estimated by $\widehat{\mathrm{cv}}_{\mathrm{ns},\alpha}^{(\mathrm{II})} = \inf\{t\in\mathbb{R} : \widehat{F}_{\mathrm{ns},M}^{(\mathrm{II})}(t) \ge 1-\alpha\}$ and $\widehat{\mathrm{cv}}_{\mathrm{s},\alpha}^{(\mathrm{II})} = \inf\{t\in\mathbb{R} : \widehat{F}_{\mathrm{s},M}^{(\mathrm{II})}(t) \ge 1-\alpha\}$, where $\widehat{F}_{\mathrm{ns},M}^{(\mathrm{II})}(t) = M^{-1}\sum_{\ell=1}^M I\{|\mathbf{W}_{\mathrm{ns},\ell}|_\infty \le t\}$ and $\widehat{F}_{\mathrm{s},M}^{(\mathrm{II})}(t) = M^{-1}\sum_{\ell=1}^M I\{|\mathbf{W}_{\mathrm{s},\ell}|_\infty \le t\}$. Similarly to (2.2), for $\nu\in\{\mathrm{ns},\mathrm{s}\}$, we define the empirical version of $\Psi_{\nu,\alpha}^{(\mathrm{II})}$ by $\widehat{\Psi}_{\nu,\alpha}^{(\mathrm{II})}(M) = I\{T_\nu^{(\mathrm{II})} > \widehat{\mathrm{cv}}_{\nu,\alpha}^{(\mathrm{II})}\}$, such that the null hypothesis $H_0^{(\mathrm{II})}$ is rejected as long as $\widehat{\Psi}_{\nu,\alpha}^{(\mathrm{II})}(M) = 1$.

2.2. *Estimation of Covariance Matrices*

As a part of proposed tests, we need estimates of the covariance matrices. Many existing tests rely on the operator-norm consistent estimation of the covariance matrices that requires extra structural assumptions on the unknown covariances such as banding or sparsity. In contrast, the proposed tests require much less restrictions on covariance estimates, which grants its wide scope of applicability. In fact, the validity of the proposed testing procedures only entails the covariance estimators $\widetilde{\boldsymbol{\Sigma}}_1$ and $\widetilde{\boldsymbol{\Sigma}}_2$ to satisfy $|\widetilde{\boldsymbol{\Sigma}}_1 - \boldsymbol{\Sigma}_1|_\infty = o_P(1)$ and $|\widetilde{\boldsymbol{\Sigma}}_2 - \boldsymbol{\Sigma}_2|_\infty = o_P(1)$.

It is shown in Lemma 3 in the supplementary material that for the sample covariance and correlation matrices $\widehat{\boldsymbol{\Sigma}}_q$ and $\widehat{\mathbf{R}}_q$ with $q = 1, 2$, there holds $|\widehat{\boldsymbol{\Sigma}}_q - \boldsymbol{\Sigma}_q|_\infty + |\widehat{\mathbf{R}}_q - \mathbf{R}_q|_\infty = o_P(1)$ under mild regularity conditions for $\log(p) = o(n^{\gamma/2})$ with $0 < \gamma \le 2$. Therefore, the sample covariance and correlation matrices can be directly used in the proposed tests, while the dimension p is allowed to be as large as either $O\{\exp(n^{c_1})\}$ for some $c_1 > 0$. In comparison to the existing tests, we do not enforce any structural assumptions on the unknown covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. This reflects our motivations in Section 1. As evidenced by extensive numerical studies in Section 4, our proposed procedures are fairly robust to various covariance structures with complex forms, even the long range dependence. Although the proposed tests do not require operator-norm consistent estimates of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, still one may replace the sample covariance matrix by adaptive and rate-optimal covariance estimators to improve the empirical performance when the underlying covariance satisfies certain structural assumptions.

2.3. *Screening-Based Testing Procedures*

The proposed testing procedures are valid when the dimension p is much larger than the sample size n. However, building tests based on all dimensions may result in large critical values which may compromise the power performance. To enhance the power, we propose a two-step procedure that combines the proposed simulation-based tests and a preliminary step on *feature screening*, which screens the p measurements before conducting the test. The power of this two-step procedure is expected to improve upon the proposed tests with a large number of irrelevant features excluded.

2.3.1. One-sample case. Let $\mathcal{S}_{10} = \{1 \le k \le p : \mu_{1k} = 0\}$. The preliminary procedure is aimed at eliminating irrelevant features indexed by $\mathcal{S}_{10}$. Reformulate the original global test of a mean vector to the following p marginal tests: $H_{0k}^{(\mathrm{I})} : \mu_{1k} = 0$ versus $H_{1k}^{(\mathrm{I})} : \mu_{1k} \ne 0$, for $k = 1, \ldots, p$. For the kth marginal hypothesis, a standard test statistic is the t-statistic $\mathrm{TS}_k^{(\mathrm{I})} = \sqrt{n}|\bar{X}_k|/\hat{\sigma}_{1k}$. Motivated by the idea of marginal screening (Chang et al., 2013, 2016), we define the index set $\widehat{\mathcal{S}}_1 = \{1 \le k \le p : \mathrm{TS}_k^{(\mathrm{I})} \le \sqrt{2\log(p)} + \{2\log(p)\}^{-1/2} + \sqrt{2\log(1/\alpha)}\}$. We refer to Chang et al. (2013, 2016) for more discussions on the advantages of the studenized statistics in marginal screening problems. If $|\widehat{\mathcal{S}}_1| < p$, we put $d = p - |\widehat{\mathcal{S}}_1|$ and let $\widetilde{\boldsymbol{\mu}}_1 \in \mathbb{R}^d$ be the sub-vector of $\boldsymbol{\mu}_1 \in \mathbb{R}^p$ containing only the coordinates excluded by $\widehat{\mathcal{S}}_1$. We have therefore downsized the original problem and instead, we focus on the reduced null hypothesis $\widetilde{H}_0^{(\mathrm{I})} : \widetilde{\boldsymbol{\mu}}_1 = \mathbf{0}$ against the alternative $\widetilde{H}_1^{(\mathrm{I})} : \widetilde{\boldsymbol{\mu}}_1 \ne \mathbf{0}$. Write $\widehat{T}_{\mathrm{ns}}^{(\mathrm{I})} = \max_{k\notin\widehat{\mathcal{S}}_1}\sqrt{n}|\bar{X}_k|$ and $\widehat{T}_{\mathrm{s}}^{(\mathrm{I})} = \max_{k\notin\widehat{\mathcal{S}}_1}\sqrt{n}|\bar{X}_k|/\hat{\sigma}_{1k}$. The resulting non-studentized and studentized tests are given by $\Psi_{\mathrm{ns},\alpha}^{f,(\mathrm{I})} = I\{\widehat{T}_{\mathrm{ns}}^{(\mathrm{I})} > \mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{I})}(\widehat{\mathcal{S}}_1)\}$ and $\Psi_{\mathrm{s},\alpha}^{f,(\mathrm{I})} = I\{\widehat{T}_{\mathrm{s}}^{(\mathrm{I})} > \mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{I})}(\widehat{\mathcal{S}}_1)\}$, where $\mathrm{cv}_{\mathrm{ns},\alpha}^{(\mathrm{I})}(\widehat{\mathcal{S}}_1)$ and $\mathrm{cv}_{\mathrm{s},\alpha}^{(\mathrm{I})}(\widehat{\mathcal{S}}_1)$ denote the conditional $(1-\alpha)$-quantile of $\max_{k\notin\widehat{\mathcal{S}}_1}|W_{\mathrm{ns},k}^{(\mathrm{I})}|$ and $\max_{k\notin\widehat{\mathcal{S}}_1}|W_{\mathrm{s},k}^{(\mathrm{I})}|$ given $\mathcal{X}_n$, respectively, with $\mathbf{W}_{\mathrm{ns}}^{(\mathrm{I})} = (W_{\mathrm{ns},1}^{(\mathrm{I})}, \ldots, W_{\mathrm{ns},p}^{(\mathrm{I})})^{\mathrm{T}}$ and $\mathbf{W}_{\mathrm{s}}^{(\mathrm{I})} = (W_{\mathrm{s},1}^{(\mathrm{I})}, \ldots, W_{\mathrm{s},p}^{(\mathrm{I})})^{\mathrm{T}}$ as discussed in Section 2.1.1. Whenever $|\widehat{\mathcal{S}}_1| = p$, we set $\Psi_{\mathrm{ns},\alpha}^{f,(\mathrm{I})} = \Psi_{\mathrm{s},\alpha}^{f,(\mathrm{I})} = 0$.

Notice that $\mathbb{P}_{H_0^{(\mathrm{I})}}\{\Psi_{\nu,\alpha}^{f,(\mathrm{I})} = 1\} \le \mathbb{P}_{H_0^{(\mathrm{I})}}[\Psi_{\nu,\alpha}^{f,(\mathrm{I})} = 1, \widehat{\mathcal{S}}_1 = \{1,\ldots,p\}] + \mathbb{P}_{H_0^{(\mathrm{I})}}[\widehat{\mathcal{S}}_1 \ne \{1,\ldots,p\}]$ for $\nu\in\{\mathrm{ns},\mathrm{s}\}$. Since $\Psi_{\nu,\alpha}^{f,(\mathrm{I})} = 0$ if $|\widehat{\mathcal{S}}| = p$, then $\mathbb{P}_{H_0^{(\mathrm{I})}}\{\Psi_{\nu,\alpha}^{f,(\mathrm{I})} = 1\} \le \mathbb{P}_{H_0^{(\mathrm{I})}}[\widehat{\mathcal{S}}_1 \ne \{1,\ldots,p\}]$. As shown in part D of supplementary material, $\limsup_{n\to\infty}\mathbb{P}_{H_0^{(\mathrm{I})}}[\widehat{\mathcal{S}}_1 \ne \{1,\ldots,p\}] \le \alpha$, which indicates that the size of the two-step procedure can be controlled by the prescribed significant level α. On the other hand, also stated in part D of supplementary material, $\mathbb{P}_{H_1^{(\mathrm{I})}}\{\widehat{T}_\nu^{(\mathrm{I})} = T_\nu^{(\mathrm{I})}\} \to 1$ for $\nu\in\{\mathrm{ns},\mathrm{s}\}$ which means the testing statistics with screening and without screening are almost identical under $H_1^{(\mathrm{I})}$. Since the critical value $\mathrm{cv}_{\nu,\alpha}^{(\mathrm{I})}(\widehat{\mathcal{S}}_1)$ for two-step procedure is not larger than $\mathrm{cv}_{\nu,\alpha}^{(\mathrm{I})}$ for non-screening procedure, we know with probability approaching to one that the power for two-step procedure does not decrease in comparison to the procedure without screening. The simulation studies in Section 4 also verify this.

2.3.2. Two-sample case. Similar to the one-sample case, for each $k = 1, \ldots, p$, we define $\mathrm{TS}_k^{(\mathrm{II})} = \sqrt{nm}$

$|\bar{X}_k - \bar{Y}_k|/\left(m\hat{\sigma}_{1k}^2 + n\hat{\sigma}_{2k}^2\right)^{1/2}$ and set $\widehat{\mathcal{S}}_2 = \{1 \le k \le p : \mathrm{TS}_k^{(\mathrm{II})} \le [\sqrt{2\log(p)} + \{2\log(p)\}^{-1/2} + \sqrt{2\log(1/\alpha)}\}$. If $|\widehat{\mathcal{S}}_2| < p$, the resulting tests, denoted by $\Psi_{\mathrm{ns},\alpha}^{f,(\mathrm{II})}$ and $\Psi_{\mathrm{s},\alpha}^{f,(\mathrm{II})}$, are defined in the same way as $\Psi_{\mathrm{ns},\alpha}^{f,(\mathrm{I})}$ and $\Psi_{\mathrm{s},\alpha}^{f,(\mathrm{I})}$ for one-sample case, respectively. If $|\widehat{\mathcal{S}}_2| = p$, we set $\Psi_{\mathrm{ns},\alpha}^{f,(\mathrm{II})} = \Psi_{\mathrm{s},\alpha}^{f,(\mathrm{II})} = 0$.

3. Theoretical Properties

In this section, we study the properties of the proposed tests including the asymptotic sizes and powers. In practice, taking M in thousands using numerical devices to increase simulation efficiency is now the rule rather than the exception in the Monte Carlo framework. The difference between such large values of M and using mathematically ideal value $M = \infty$ is particularly small. We therefore focus on the oracle tests $\Psi_{\nu,\alpha}^{(\mathrm{I})}$ and $\Psi_{\nu,\alpha}^{(\mathrm{II})}$ for $\nu \in \{\mathrm{ns}, \mathrm{s}\}$, and their screening-based analogues $\Psi_{\nu,\alpha}^{f,(\mathrm{I})}$ and $\Psi_{\nu,\alpha}^{f,(\mathrm{II})}$. It is shown that the proposed tests maintain the nominal size asymptotically under very general covariance structures. Moreover, the proposed tests are shown to be consistent against sparse alternatives. Recall $\mathbf{\Sigma}_1 = (\sigma_{1,k\ell})_{1\le k,\ell\le p}$, $\mathbf{\Sigma}_2 = (\sigma_{2,k\ell})_{1\le k,\ell\le p}$, $\mathbf{D}_1 = \mathrm{diag}\,(\mathbf{\Sigma}_1)$ and $\mathbf{D}_2 = \mathrm{diag}\,(\mathbf{\Sigma}_2)$. The marginally standardized version of $\mathbf{X}$ and $\mathbf{Y}$ are $\mathbf{U} = (U_1, \ldots, U_p)^{\mathrm{T}} = \mathbf{D}_1^{-1/2}\mathbf{X}$ and $\mathbf{V} = (V_1, \ldots, V_p)^{\mathrm{T}} = \mathbf{D}_2^{-1/2}\mathbf{Y}$, respectively. We only impose the following mild moment conditions.

- **(M1)** $\max_{1\le k\le p} \max[\{\mathbb{E}(|U_k|^r)\}^{1/r}, \{\mathbb{E}(|V_k|^r)\}^{1/r}] \le K_0$ for some $r \ge 4$ and $K_0 > 0$
- **(M2)** $\max_{1\le k\le p} \max[\mathbb{E}\{\exp(K_1|U_k|^\gamma)\}, \mathbb{E}\{\exp(K_1|V_k|^\gamma)\}] \le K_2$ for some $K_1 > 0$, $K_2 > 1$ and $0 < \gamma \le 2$.

Condition (M1) indicates that the tail probability $(|U_k| > t)$ decays to zero in a faster rate than t^{-r} as $t \to \infty$. Condition (M2) requires exponentially light tails, that is, $(|U_k| > t) \le \exp(-\widetilde{K}_1 t^\gamma)$ for some $\widetilde{K}_1 > 0$ and all sufficiently large t, and implies that all moments of U_k are finite. Throughout this section, we assume that $\sigma_{1,11}, \ldots, \sigma_{1,pp}, \sigma_{2,11}, \ldots, \sigma_{2,pp}$ are uniformly bounded away from 0 and ∞, $n, p \ge 2$, $n \asymp m$, and $n \le m$.

THEOREM 1. *Let $\widetilde{\mathbf{\Sigma}}_1 = \widehat{\mathbf{\Sigma}}_1$, the sample covariance matrix, and $\nu \in \{\mathrm{ns}, \mathrm{s}\}$. As $n, p \to \infty$, $\mathbb{P}_{H_0^{(\mathrm{I})}}\{\Psi_{\nu,\alpha}^{(\mathrm{I})} = 1\} \to \alpha$ holds with either* (i) (M1) *holds and $p = O(n^{r/2-1-\delta})$ for some $\delta > 0$; or* (ii) (M2) *holds for some $\gamma \ge 1/2$ and $\log(p) = o(n^{1/7})$.*

Theorem 1 establishes the validity of the proposed one-sample tests in the sense that the testing procedures in Section 2.1.1 maintain nominal significance level asymptotically. In addition, as evidenced by the numerical experiments in Section 4, the test based on non-studentized statistics outperforms its studentized analogue in terms of maintaining the nominal significance level when the sample size is small. This, however, is not surprising since the inverse operation, say $\widehat{\mathbf{D}}_1^{-1/2}$, usually leads to an augmentation of the estimation error in $\widehat{\mathbf{D}}_1$ and therefore is more sensitive to the sample size. In the following theorem, we summarize the asymptotic power of the proposed one-sample tests under suitable conditions on the lower bound of the signal-to-noise ratios.

THEOREM 2. *Let $\widetilde{\mathbf{\Sigma}}_1 = \widehat{\mathbf{\Sigma}}_1$ be the sample covariance matrix. Assume that either condition* (M1) *holds and $p = O(n^{r/2-1-\delta})$ for some $\delta > 0$, or condition* (M2) *holds and $\log(p) = o(n^{\gamma/2})$. For given $0 < \alpha < 1$, write $\lambda(p,\alpha) = \sqrt{2\log(p)} + \sqrt{2\log(1/\alpha)}$, and let $\{\varepsilon_n\}_{n\ge1}$ be an arbitrary sequence of positive numbers satisfying $\varepsilon_n \to 0$ and $\varepsilon_n\sqrt{\log(p)} \to \infty$ as $n \to \infty$. As $n, p \to \infty$, we have* (i) $\mathbb{P}_{H_1^{(\mathrm{I})}}\{\Psi_{\mathrm{ns},\alpha}^{(\mathrm{I})} = 1\} \to 1$ *if* $\max_{1\le k\le p}|\mu_{1k}|/\max_{1\le k\le p}\sigma_{1k} \ge (1+\varepsilon_n)n^{-1/2}\lambda(p,\alpha)$, *and* (ii) $\mathbb{P}_{H_1^{(\mathrm{I})}}\{\Psi_{\mathrm{s},\alpha}^{(\mathrm{I})} = 1\} \to 1$ *if* $\max_{1\le k\le p}|\mu_{1k}|/\sigma_{1k} \ge (1+\varepsilon_n)n^{-1/2}\lambda(p,\alpha)$.

Theorem 2 reveals that the test based on studentized statistics is consistent in a larger testable region in comparison to the test based on non-studentized statistics. As a complement to Theorem 1, the asymptotic size of the proposed two-sample tests without screening is reported below.

THEOREM 3. *Let $(\widetilde{\mathbf{\Sigma}}_1, \widetilde{\mathbf{\Sigma}}_2) = (\widehat{\mathbf{\Sigma}}_1, \widehat{\mathbf{\Sigma}}_2)$ and $\nu \in \{\mathrm{ns}, \mathrm{s}\}$. Assume that either condition* (i) *or condition* (ii) *in Theorem* 1 *holds. Then as $n, p \to \infty$, $\mathbb{P}_{H_0^{(\mathrm{II})}}\{\Psi_{\nu,\alpha}^{(\mathrm{II})} = 1\} \to \alpha$.*

Theorem 3 implies that, under proper moment conditions, the proposed two-sample non-screening tests maintain nominal size α asymptotically, while allowing for either a polynomial or an exponential rate of growth of the dimension p with respect to the sample size n. In Theorem 4 below, the asymptotic power of the two-sample non-screening tests is analyzed.

THEOREM 4. *Let $(\widetilde{\mathbf{\Sigma}}_1, \widetilde{\mathbf{\Sigma}}_2) = (\widehat{\mathbf{\Sigma}}_1, \widehat{\mathbf{\Sigma}}_2)$. Assume that either condition* (M1) *holds and $p = O(n^{r/2-1-\delta})$ for some $\delta > 0$, or condition* (M2) *holds and $\log(p) = o(n^{\gamma/2})$. For given $0 < \alpha < 1$, let $\lambda(p,\alpha)$ and $\{\varepsilon_n\}_{n\ge1}$ be as in Theorem* 2. *As $n, p \to \infty$, we have* (i) $\mathbb{P}_{H_1^{(\mathrm{II})}}\{\Psi_{\mathrm{ns},\alpha}^{(\mathrm{II})} = 1\} \to 1$ *if* $\max_{1\le k\le p}|\mu_{1k} - \mu_{2k}|/\max_{1\le k\le p}(\sigma_{1k}^2/n + \sigma_{2k}^2/m)^{1/2} \ge (1+\varepsilon_n)\lambda(p,\alpha)$, *and* (ii) $\mathbb{P}_{H_1^{(\mathrm{II})}}\{\Psi_{\mathrm{s},\alpha}^{(\mathrm{II})} = 1\} \to 1$ *if* $\max_{1\le k\le p}|\mu_{1k} - \mu_{2k}|/(\sigma_{1k}^2/n + \sigma_{2k}^2/m)^{1/2} \ge (1+\varepsilon_n)\lambda(p,\alpha)$.

The following theorem establishes asymptotic properties of the proposed two-step testing procedures. Part (i) in Theorem 5 below shows that the type I error of the proposed screening-based two-step procedures can be controlled by the prescribed significance level asymptotically. Similar to the comparison between the studentized and non-studentized tests in Theorem 2, parts (ii) and (iii) in Theorem 5 below also imply that the screening-based two-step studentized test is consistent in a larger region than its non-studentized counterpart.

THEOREM 5. *Let $\widetilde{\mathbf{\Sigma}}_1 = \widehat{\mathbf{\Sigma}}_1$. Assume that either condition* (M1) *holds and $p = O(n^{r/2-1-\delta})$ for some $\delta > 0$, or condition* (M2) *holds for some $\gamma \ge \frac{1}{2}$ and $\log(p) = o(n^{1/7})$. We have* (i) $\limsup_{n\to\infty}\mathbb{P}_{H_0^{(\mathrm{I})}}\{\Psi_{\nu,\alpha}^{f,(\mathrm{I})} = 1\} \le \alpha$ *for* $\nu \in \{\mathrm{ns}, \mathrm{s}\}$, (ii) $\mathbb{P}_{H_1^{(\mathrm{I})}}\{\Psi_{\mathrm{ns},\alpha}^{f,(\mathrm{I})} = 1\} \to 1$ *if the condition for part* (i) *in Theorem*

2 *holds,* (iii) $\mathbb{P}_{H_1^{(\mathrm{I})}}\{\Psi^{f,(\mathrm{I})}_{\mathrm{s},\alpha}=1\}\to 1$ *if the condition for part* (ii) *in Theorem* 2 *holds.*

Similarly, the following theorem establishes the limiting null property and the asymptotic power for the proposed two-step procedures with pre-screening in the two-sample settings.

THEOREM 6. *Let* $(\widetilde{\boldsymbol{\Sigma}}_1,\widetilde{\boldsymbol{\Sigma}}_2)=(\widehat{\boldsymbol{\Sigma}}_1,\widehat{\boldsymbol{\Sigma}}_2)$. *Assume that either condition* (M1) *holds and* $p=O(n^{r/2-1-\delta})$ *for some* $\delta>0$, *or condition* (M2) *holds for some* $\gamma\geq\frac{1}{2}$ *and* $\log(p)=o(n^{1/7})$. *We have* (i) $\limsup_{n\to\infty}\mathbb{P}_{H_0^{(\mathrm{II})}}\{\Psi^{f,(\mathrm{II})}_{\nu,\alpha}=1\}\leq\alpha$ *for* $\nu\in\{\mathrm{ns},\mathrm{s}\}$, (ii) $\mathbb{P}_{H_1^{(\mathrm{II})}}\{\Psi^{f,(\mathrm{II})}_{\mathrm{ns},\alpha}=1\}\to 1$ *if the condition for part* (i) *in Theorem* 4 *holds, and* (iii) $\mathbb{P}_{H_1^{(\mathrm{II})}}\{\Psi^{f,(\mathrm{II})}_{\mathrm{s},\alpha}=1\}\to 1$ *if the condition for part* (ii) *in Theorem* 4 *holds.*

4. Simulation Studies

In this section, we report the simulation results from several experiments to evaluate the performance of the proposed tests, including the non-studentized test without screening $\Psi_{\mathrm{ns},\alpha}$, the studentized test without screening $\Psi_{\mathrm{s},\alpha}$, the non-studentized test with screening $\Psi^f_{\mathrm{ns},\alpha}$ and the studentized test with screening $\Psi^f_{\mathrm{s},\alpha}$, for both one- and two-sample problems. For ease of exposition, we suppress the superscripts (I) and (II). To demonstrate the proposed tests, we also implemented peer testing procedures for comparison. For the one-sample problem, we compared the proposed tests with the test by Zhong et al. (2013) (denoted by ZCX hereafter) and the Higher Criticism (HC) procedure by Donoho and Jin (2004) . We used the method proposed by Li and Siegmund (2015) to obtain more accurate approximation of the critical values in HC procedure. For the two-sample problem, we experimented the tests by Chen and Qin (2010) (denoted by CQ hereafter) and Cai et al. (2014) (denoted by CLX hereafter) as well as the HC procedure.

In the simulation studies, we considered a wide range of covariance structures, including both the sparse and dense settings to investigate the numerical performance of the proposed tests. We generate data with sample sizes $n=40$ or 80 in one-sample case and $(n,m)=(40,40)$ or $(80,80)$ in two-sample case. The dimension p took values in 120, 360, or 1080. The empirical size and power were defined as the proportion of the rejection among 1500 replications. We used the sample covariance matrices to generate $M=1500$ Monte Carlo samples to compute the critical values for our proposed tests. We only report the results for six models in this section and more models are considered in the supplementary material.

4.1. *One-Sample Case*

We took $\boldsymbol{\mu}_1=\mathbf{0}$ under the null hypothesis, whereas, under the alternative, we took $\boldsymbol{\mu}_1=(\mu_{11},\ldots,\mu_{1p})^{\mathrm{T}}$ to have $\lfloor\kappa p^r\rfloor$ non-zero entries uniformly and randomly drawn from $\{1,\ldots,p\}$, where κ was an integer and $\lfloor x\rfloor$ denotes the integer part of x. We took $r=0,0.4,0.5,0.7$, and 0.85, where $\kappa=8$ if $r=0$ and $\kappa=1$ otherwise. The choices of $r=0$ and $r=0.7$ or 0.85 correspond to the sparse and non-sparse settings, respectively. The magnitudes of non-zero entries $\mu_{1\ell}$ were set to be $\{2\beta\sigma_{1,\ell\ell}\log(p)/n\}^{1/2}$, where $\sigma_{1,\ell\ell}$ denotes the ℓth diagonal entry of $\boldsymbol{\Sigma}_1$. We took $\beta=0.01,0.2,0.4,0.6$, and use $\beta=0.01$ to mimic the scenario of weak signals.

The following two models were used to generate random samples $\mathbf{X}_i=\mathbf{Z}_i+\boldsymbol{\mu}_1$ for $i=1,\ldots,n$, where $\{\mathbf{Z}_i\}_{i=1}^n\overset{\text{i.i.d}}{\sim}\mathrm{N}(\mathbf{0},\boldsymbol{\Sigma}_1)$ with $\boldsymbol{\Sigma}_1=(\sigma_{1,k\ell})_{1\leq k,\ell\leq p}$.

- Model $1^{(\mathrm{I})}$: $\sigma_{1,k\ell}=0.4^{|k-\ell|}$ for $1\leq k,\ell\leq p$.
- Model $2^{(\mathrm{I})}$: Let $\{\theta_k\}_{k=1}^p\overset{\text{i.i.d.}}{\sim}\mathrm{Unif}(1,2)$. We took $\sigma_{1,kk}=\theta_k$ and $\sigma_{1,k\ell}=\rho_\alpha(|k-\ell|)$ for $k\neq\ell$, where $\rho_\alpha(e)=\frac{1}{2}\{(e+1)^{2H}+(e-1)^{2H}-2e^{2H}\}$ with $H=0.9$.

Model $1^{(\mathrm{I})}$ has sparse covariance structure while Model $2^{(\mathrm{I})}$ takes long range dependence into account which exhibits a non-sparse structure. In addition, we considered the following model with non-Gaussian data to study the robustness of the proposed tests against Gaussian assumptions. The covariance structure in the following Model $3^{(\mathrm{I})}$ is non-sparse.

- Model $3^{(\mathrm{I})}$: Let $\{\mathbf{X}_i\}_{i=1}^n\overset{\text{i.i.d.}}{\sim}t_\omega(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$, where $t_\omega(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ is the non-central multivariate t-distribution with non-central parameter $\boldsymbol{\mu}_1$, degrees of freedom $\omega=5$, and $\sigma_{1,k\ell}=0.995^{|k-\ell|}$.

Simulation results for the tests $\Psi_{\mathrm{ns},\alpha}$, $\Psi_{\mathrm{s},\alpha}$, $\Psi^f_{\mathrm{ns},\alpha}$, and $\Psi^f_{\mathrm{s},\alpha}$ and the ZCX and HC tests are summarized in Table 1 and Figure 1. Table 1 displays the empirical sizes of all the tests. It can be seen that in all the models, the empirical sizes of the non-studentized tests $\Psi_{\mathrm{ns},\alpha}$ and $\Psi^f_{\mathrm{ns},\alpha}$ are reasonably close to the nominal level 0.05 for both $n=40$ and $n=80$. The proposed studentized tests $\Psi_{\mathrm{s},\alpha}$ and $\Psi^f_{\mathrm{s},\alpha}$ have slightly inflated size when n is relatively small but improve with larger sample sizes. The ZCX test maintains the nominal size for Model $1^{(\mathrm{I})}$ but fails in the presence of long range dependence or non-sparse covariance structures. The HC procedure also fails in maintaining the nominal significance when the sample size n is small or the dependency is strong and complex.

To compare the empirical powers, we took $n=80$ and $p=1080$. For Model $1^{(\mathrm{I})}$, we compared the proposed tests with the ZCX test (column (a) in Figure 1), whereas, for the other two models, we only focused on comparing the four proposed tests as they maintain the nominal size reasonably well and other tests fail in size control. Column (a) in Figure 1 shows that $\Psi_{\mathrm{s},\alpha}$, $\Psi^f_{\mathrm{s},\alpha}$, and $\Psi^f_{\mathrm{ns},\alpha}$ provide non-trivial powers against alternatives with sparse signals ($r=0$) even under the weak signal settings ($\beta=0.01$); in contrast, the ZCX test improves its power as the signal getting dense, which is expected for sum of squares-type statistics. As the signal strength increases, all tests under consideration gain powers. The proposed tests with screening, $\Psi^f_{\mathrm{ns},\alpha}$ and $\Psi^f_{\mathrm{s},\alpha}$, outperform the ZXC test under sparse alternatives ($r=0,0.4$), and their powers are close to that of the ZCX test for dense signals ($r\geq 0.7$). From columns (b) and (c) in Figure 1, we observe that the screening procedure substantially improves the power performance of the tests for all settings, which reflects the heuristic discussions and motivations in Section 2.3.1. The non-studentized test with screening $\Psi^f_{\mathrm{ns},\alpha}$ performs comparably to, or better than, the studentized test without screening $\Psi_{\mathrm{s},\alpha}$ under sparse alternatives ($r\leq 0.5$). This suggests that $\Psi^f_{\mathrm{ns},\alpha}$ is more preferable in practice given its capability in maintaining the nominal significance for small sample size.

Table 1

Empirical sizes of the proposed tests (non-studentized without screening $\Psi_{\mathrm{ns},\alpha}$, studentized without screening $\Psi_{\mathrm{s},\alpha}$, non-studentized with screening $\Psi^f_{\mathrm{ns},\alpha}$, and studenzied with screening $\Psi^f_{\mathrm{s},\alpha}$) for the one-sample problem (1.1)*, along with those of the tests by Zhong et al. (2013) (ZCX), and Donoho and Jin (2004) (HC) at 5% nominal significance. Models with Gaussian data and sparse or long range dependence (non-sparse) covariance matrices, and the autoregressive model with t-distributed innovations are considered when $n = 40, 80$ and $p = 120, 360, 1080$.*

	Model $1^{(\mathrm{I})}$			Model $2^{(\mathrm{I})}$			Model $3^{(\mathrm{I})}$		
tests/p	120	360	1080	120	360	1080	120	360	1080
					$n = 40$				
$\Psi_{\mathrm{ns},\alpha}$	0.037	0.027	0.021	0.025	0.028	0.023	0.054	0.044	0.033
$\Psi_{\mathrm{s},\alpha}$	0.133	0.126	0.168	0.093	0.113	0.202	0.065	0.080	0.096
$\Psi^f_{\mathrm{ns},\alpha}$	0.044	0.045	0.043	0.039	0.027	0.039	0.054	0.046	0.033
$\Psi^f_{\mathrm{s},\alpha}$	0.150	0.154	0.194	0.095	0.170	0.218	0.060	0.058	0.093
ZCX	0.064	0.078	0.089	1	1	1	0.382	0.487	0.673
HC	0.123	0.225	0.316	0.129	0.249	0.320	0.274	0.377	0.468
					$n = 80$				
$\Psi_{\mathrm{ns},\alpha}$	0.037	0.036	0.029	0.040	0.032	0.042	0.049	0.047	0.040
$\Psi_{\mathrm{s},\alpha}$	0.060	0.082	0.092	0.082	0.083	0.094	0.058	0.058	0.067
$\Psi^f_{\mathrm{ns},\alpha}$	0.048	0.045	0.043	0.051	0.045	0.040	0.049	0.048	0.044
$\Psi^f_{\mathrm{s},\alpha}$	0.086	0.097	0.094	0.095	0.091	0.110	0.060	0.058	0.069
ZCX	0.080	0.072	0.071	1	1	1	0.404	0.506	0.702
HC	0.063	0.119	0.142	0.079	0.145	0.175	0.267	0.363	0.471

4.2. *Two-Sample Case*

We took $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$ under the null hypothesis, whereas, under the alternative, we let $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})^{\mathrm{T}}$ to have $\lfloor \kappa p^r \rfloor$ non-zero entries uniformly and randomly drawn from $\{1, \ldots, p\}$, where κ is an integer. As before, we considered $r = 0, 0.4, 0.5, 0.7$, and 0.85, where $\kappa = 8$ if $r = 0$ and $\kappa = 1$ otherwise. The magnitudes of non-zero entries $\mu_{1\ell}$ were set to be $\{2\beta\sigma_{\ell\ell}\log(p)(1/n + 1/m)\}^{1/2}$, where $\sigma_{\ell\ell}$ is the ℓth diagonal entry of the pooled covariance matrix $\boldsymbol{\Sigma}_{1,2}$ as in (2.3). We took $\beta = 0.01, 0.2, 0.4, 0.6$.

The following two models were used to generate random samples $\mathbf{X}_i = \mathbf{Z}_{1,i} + \boldsymbol{\mu}_1, \mathbf{Y}_j = \mathbf{Z}_{2,j} + \boldsymbol{\mu}_2$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where $\{\mathbf{Z}_{1,i}\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ and $\{\mathbf{Z}_{2,j}\}_{j=1}^m \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_2)$ with $\boldsymbol{\Sigma}_1 = (\sigma_{1,k\ell})_{1\le k,\ell\le p}$ and $\boldsymbol{\Sigma}_2 = (\sigma_{2,k\ell})_{1\le k,\ell\le p}$, respectively.

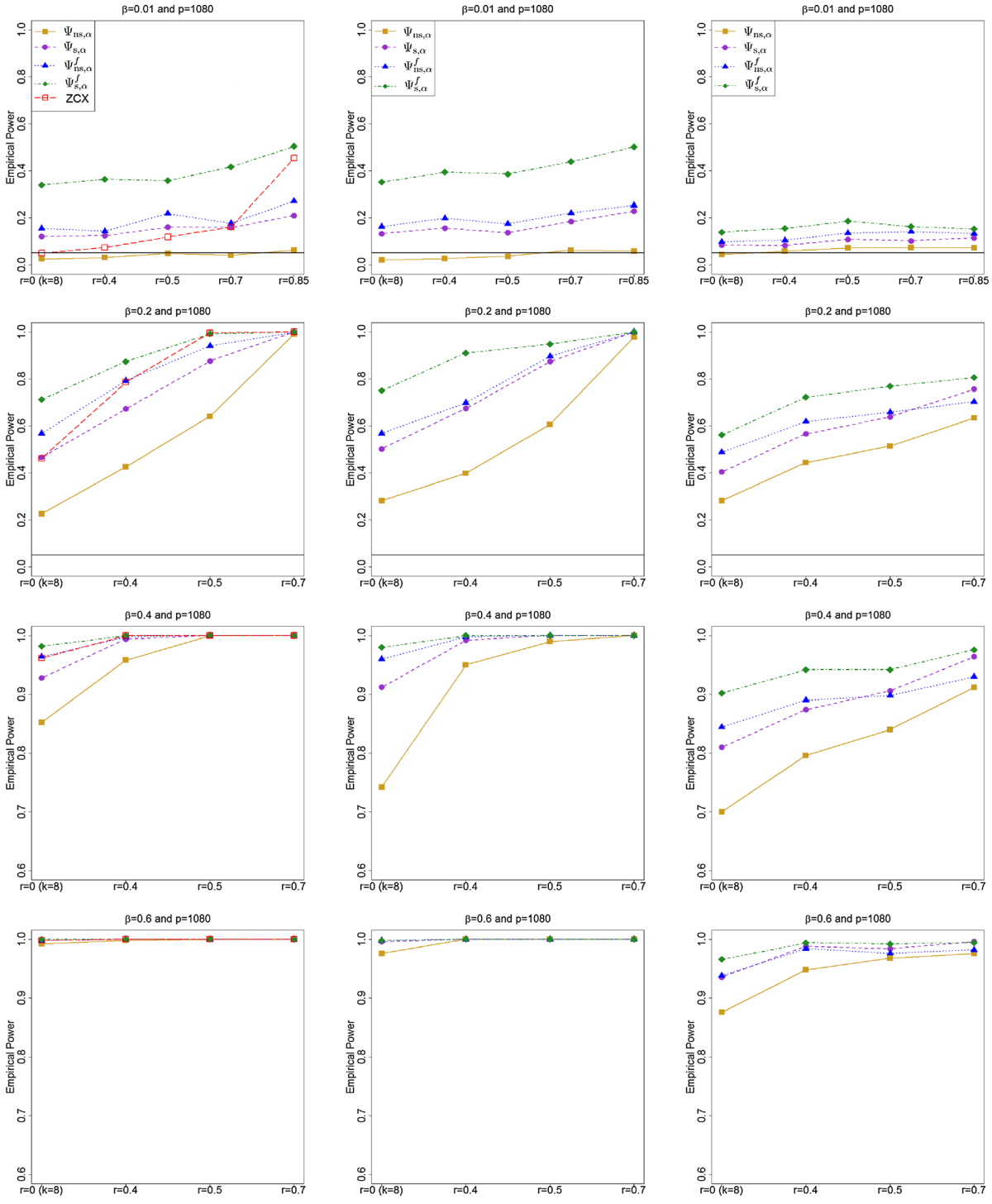
- Model $1^{(\mathrm{II})}$: For $k = 1, \ldots, p$ and $q = 1, 2$, $\sigma_{q,kk} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(2, 3)$, $\sigma_{q,k\ell} = 0.7$ for $10(t-1) + 1 \le k \ne \ell \le 10t$, where $t = 1, \ldots, \lfloor p/10 \rfloor$, and $\sigma_{q,k\ell} = 0$ otherwise.
- Model $2^{(\mathrm{II})}$: Let $\mathbf{F} = (f_{k\ell})_{1\le k,\ell\le p}$ with $f_{kk} = 1$, $f_{k,k+1} = f_{k+1,k} = 0.5$, $\mathbf{U}_q \sim \mathcal{U}(\mathcal{V}_{p,k_0})$, the uniform distribution on the Stiefel manifold for $q = 1, 2$, and $\boldsymbol{\Theta} = \mathrm{diag}\{\theta_{11}, \ldots, \theta_{pp}\}$ with $\theta_{kk} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(1, 6)$. Set $k_0 = 10$ and put $\boldsymbol{\Sigma}_q = \boldsymbol{\Theta}^{1/2}(\mathbf{F} + \mathbf{U}_q\mathbf{U}_q^{\mathrm{T}})\boldsymbol{\Theta}^{1/2}$ for $q = 1, 2$.

Model $1^{(\mathrm{II})}$ and Model $2^{(\mathrm{II})}$ are with sparse and non-sparse covariance structures, respectively. In addition, we considered the following model with non-Gaussian data.

- Model $3^{(\mathrm{II})}$: Let $\{\mathbf{X}_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} t_{\omega_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\{\mathbf{Y}_j\}_{j=1}^m \overset{\text{i.i.d.}}{\sim} t_{\omega_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, where $\omega_1 = 5, \omega_2 = 7$, $\sigma_{1,k\ell} = 0.995^{|k-\ell|}$ and $\sigma_{2,k\ell} = 0.7^{|k-\ell|}$.

The numerical results on the proposed tests $\Psi_{\mathrm{ns},\alpha}$, $\Psi_{\mathrm{s},\alpha}$, $\Psi^f_{\mathrm{ns},\alpha}$, and $\Psi^f_{\mathrm{s},\alpha}$ and the HC, CQ, and CLX tests are summarized in Table 2 and Figure 2. Table 2 displays the empirical sizes. It can be seen that in all the models, the empirical sizes for $\Psi_{\mathrm{ns},\alpha}$ and $\Psi^f_{\mathrm{ns},\alpha}$ are reasonably close to the nominal level 0.05 for both $(n, m) = (40, 40)$ and $(80, 80)$. The studentized tests, $\Psi_{\mathrm{s},\alpha}$ and $\Psi^f_{\mathrm{s},\alpha}$, have slightly inflated significance when the sample size is relatively small but improve when the sample size increases. Additionally, the CLX test fails to maintain the nominal size for Model $3^{(\mathrm{II})}$ due to the strong dependency in the covariance structures. Analogous to the observation in Section 4.1, it is difficult for the HC procedure to maintain the nominal significance when the sample size is small or the dependency is strong and complex. The CQ test maintains the nominal significance reasonably well in all the models.

To evaluate the power, we compared the proposed tests with the CQ and CLX tests for $(n, m) = (80, 80)$ and $p = 1080$. It can be seen that the tests with screening, $\Psi^f_{\mathrm{ns},\alpha}$ and $\Psi^f_{\mathrm{s},\alpha}$, outperform both the CQ and CLX tests against alternatives with sparse signals ($r = 0$) for different signal strength β. On the other hand, all the tests perform similarly when the signals become less sparse and strong. The CQ test gains more powers when signals become less sparse, as expected for sum of squares-type statistics. Its power approaches to those of the proposed tests with screening $\Psi^f_{\mathrm{ns},\alpha}$ and $\Psi^f_{\mathrm{s},\alpha}$ when the signals become less sparse and stronger ($r \ge 0.5, \beta \ge 0.4$) in the models except Model $3^{(\mathrm{II})}$. In Model



(a) Model $1^{(\mathrm{I})}$ (b) Model $2^{(\mathrm{I})}$ (c) Model $3^{(\mathrm{I})}$

Figure 1. Empirical powers of the proposed tests (non-studentized without screening $\Psi_{\mathrm{ns},\alpha}$, studentized without screening $\Psi_{\mathrm{s},\alpha}$, non-studentized with screening $\Psi^f_{\mathrm{ns},\alpha}$, and also studenzied with screening $\Psi^f_{\mathrm{s},\alpha}$) against alternatives with different levels of the signal strength (β) and sparsity ($1-r$) for the one-sample problem (1.1) when $n = 80$ and $p = 1080$ at 5% nominal significance for the Gaussian data and sparse covariance matrices in Model $1^{(\mathrm{I})}$ (column (a)), the Gaussian data and long range dependence covariance matrices in Model $2^{(\mathrm{I})}$ (column (b)), and the autoregressive process model, Model $3^{(\mathrm{I})}$, with t-distributed innovations (column (c)). Column (a) also displays the powers of the test by Zhong et al. (2013) (ZCX).

Table 2

Empirical sizes of the proposed tests (non-studentized without screening $\Psi_{\mathrm{ns},\alpha}$, studentized without screening $\Psi_{\mathrm{s},\alpha}$, non-studentized with screening $\Psi^f_{\mathrm{ns},\alpha}$, and studenzied with screening $\Psi^f_{\mathrm{s},\alpha}$) for the two-sample problem (1.2)*, along with those of the tests by Donoho and Jin (2004) (HC), Chen and Qin (2010) (CQ), and Cai et al. (2014) (CLX) at 5% nominal significance. Models with Gaussian data and sparse or non-sparse covariance matrices, and with non-Gaussian data are considered when $n = m = 40$ or 80 and $p = 120, 360, 1080$.*

	Model $1^{(\mathrm{II})}$			Model $2^{(\mathrm{II})}$			Model $3^{(\mathrm{II})}$		
tests/p	120	360	1080	120	360	1080	120	360	1080
				$(n, m) = (40, 40)$					
$\Psi_{\mathrm{ns},\alpha}$	0.039	0.041	0.041	0.042	0.044	0.039	0.052	0.036	0.042
$\Psi_{\mathrm{s},\alpha}$	0.094	0.112	0.125	0.092	0.097	0.116	0.086	0.090	0.092
$\Psi^f_{\mathrm{ns},\alpha}$	0.055	0.048	0.057	0.049	0.055	0.054	0.055	0.039	0.052
$\Psi^f_{\mathrm{s},\alpha}$	0.092	0.120	0.152	0.098	0.131	0.053	0.090	0.094	0.094
HC	0.086	0.156	0.157	0.078	0.144	0.148	0.172	0.237	0.283
CQ	0.044	0.049	0.034	0.046	0.049	0.051	0.064	0.066	0.054
CLX	0.101	0.103	0.138	0.081	0.087	0.098	0.204	0.181	0.137
				$(n, m) = (80, 80)$					
$\Psi_{\mathrm{ns},\alpha}$	0.054	0.039	0.046	0.053	0.040	0.040	0.046	0.045	0.047
$\Psi_{\mathrm{s},\alpha}$	0.074	0.062	0.086	0.058	0.064	0.090	0.059	0.065	0.074
$\Psi^f_{\mathrm{ns},\alpha}$	0.065	0.052	0.060	0.063	0.050	0.058	0.047	0.048	0.056
$\Psi^f_{\mathrm{s},\alpha}$	0.088	0.076	0.098	0.070	0.080	0.093	0.062	0.069	0.086
HC	0.068	0.086	0.099	0.053	0.085	0.085	0.165	0.239	0.263
CQ	0.046	0.039	0.048	0.048	0.038	0.048	0.044	0.054	0.056
CLX	0.107	0.090	0.104	0.057	0.057	0.089	0.289	0.352	0.297

$3^{(\mathrm{II})}$, all the proposed tests outperform the CQ test substantially as the sum of squares-type test statistics may lose power for heavy tailed sampling distributions. The CLX test performs similarly to the $\Psi_{\mathrm{ns},\alpha}$ and $\Psi_{\mathrm{s},\alpha}$, but is outperformed by the proposed tests with screening for all settings. The simulation results agree with the heuristic discussion and the theoretical justification that the screening step substantially improves the power of proposed tests. Similar to the observations in Section 4.1, $\Psi^f_{\mathrm{ns},\alpha}$ is preferable in practice whenever the sample size is relatively small.

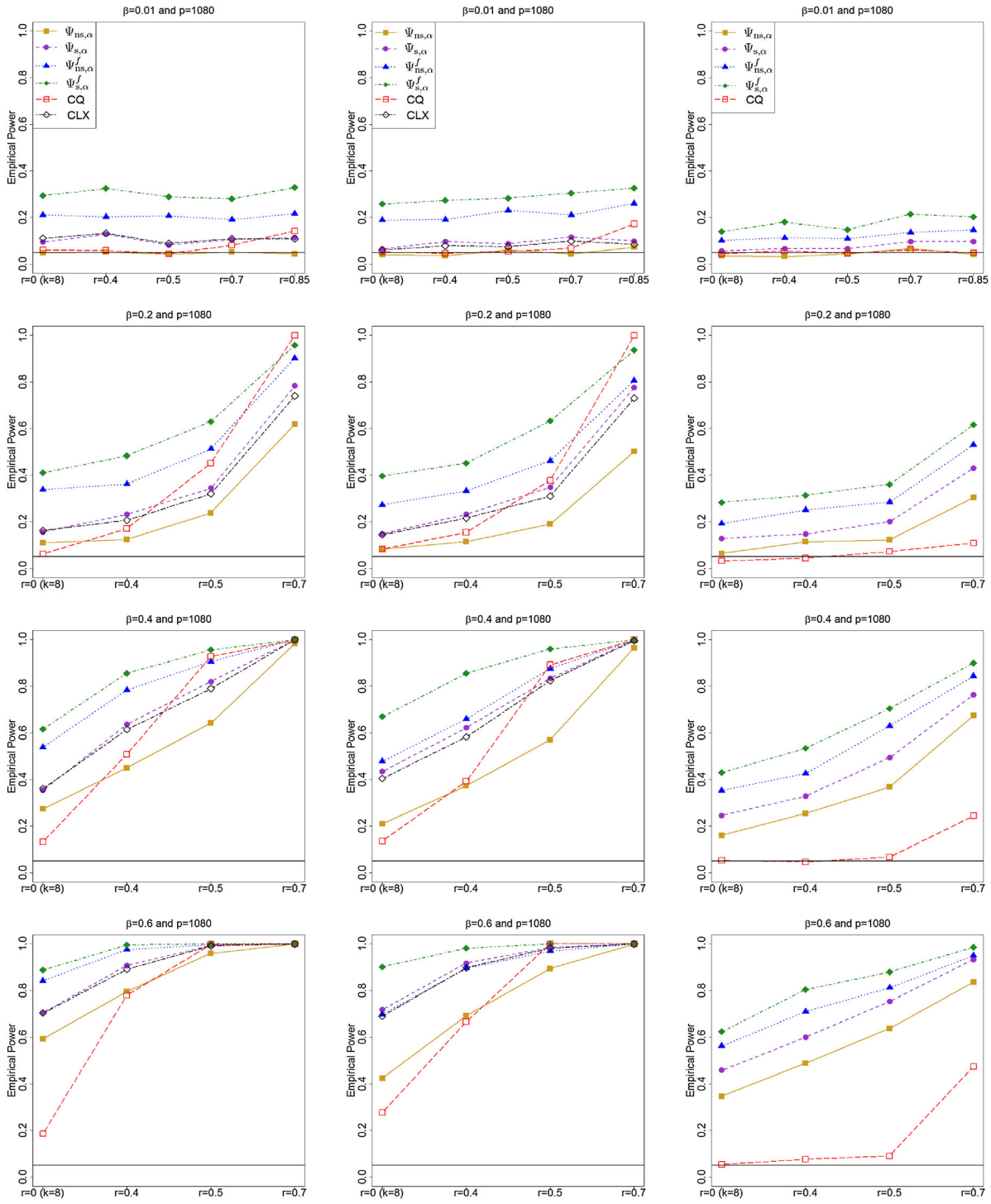
In summary, the numerical results show that the proposed tests, particularly the studentized tests and the non-studentized test with screening, $\Psi_{\mathrm{s},\alpha}$, $\Psi^f_{\mathrm{s},\alpha}$, and $\Psi^f_{\mathrm{ns},\alpha}$, outperform the existing methods when the covariance structure is non-sparse and complex. The proposed tests are robust against both unknown covariance structures and Gaussianity. The $\Psi^f_{\mathrm{ns},\alpha}$ maintains the nominal significance for small sample sizes and has good powers against sparse alternatives, which is recommended for practical applications with relatively small sample size. The $\Psi^f_{\mathrm{s},\alpha}$ is more powerful and thus is preferable in applications with relatively large samples, such as biomedical research with a large cohort.

More extensive simulations were carried out for dimensions $p = 120$ and 360, from which the comparisons are consistent with the cases that are reported here. The empirical powers of all the tests also increase in p. All the additional simulation results are placed in the online supplementary materials. Furthermore, extra simulations were reported in the supplementary materials to demonstrate that the proposed procedures may benefit from using regularized covariance estimations when the covariance matrices do admit special structures.

5. Empirical Study

Analysis and interpretation based on gene-sets or GO terms derive more power than focusing on individual gene in extracting biological insights (Subramanian et al., 2005). It has drawn increasing attentions to identify GO terms associated with biological states of interest (Subramanian et al., 2005; Efron and Tibshirani, 2007; Recknor et al., 2008). A particular GO term belongs to one of the three categories of gene ontologies of interest: biological processes (BP), cellular components (CC), and molecular functions (MF).

Statistically, identifying interesting gene-sets out of G candidate gene-sets $\mathcal{S}_1, \ldots, \mathcal{S}_G$ based on independent samples from two biological states ($q = 1, 2$) is equivalent to test hypotheses $H_{0s} : \boldsymbol{\mu}_{1,s} = \boldsymbol{\mu}_{2,s}$ versus $H_{1s} : \boldsymbol{\mu}_{1,s} \neq \boldsymbol{\mu}_{2,s}$ for $s = 1, \ldots, G$, where $\boldsymbol{\mu}_{q,s}$ models the mean expression levels of p_s genes in the gene-set $\mathcal{S}_s$ under biological state q. It is common that gene-sets overlap with each other as one particular gene may belong to several functional groups, and the size of a gene-set p_s usually range from a small to a very large number. The selection of gene-sets therefore encounters both multiplicity and high dimensionality. Similar to Chen and Qin (2010), we applied the proposed tests to each gene-set. With p-values obtained for all G gene-sets, we further employed the multiple testing methods such as the Benjamini–Yekutieli (BY) procedure (Benjamini and Yekutieli, 2001) for controlling the false



(a) Model $1^{(II)}$ (b) Model $2^{(II)}$ (c) Model $3^{(II)}$

Figure 2. Empirical powers of the proposed tests (non-studentized without screening $\Psi_{ns,\alpha}$, studentized without screening $\Psi_{s,\alpha}$, non-studentized with screening $\Psi^f_{ns,\alpha}$, and also studenzied with screening $\Psi^f_{s,\alpha}$) against alternatives with different levels of the signal strength (β) and sparsity ($1-r$) for the two-sample problem (1.2) when $n = 80$ and $p = 1080$ at 5% nominal significance for the Gaussian data and sparse covariance matrices in Model $1^{(II)}$ (column (a)), the Gaussian data and non-sparse covariance matrices in Model $2^{(II)}$ (column (b)), and the non-Gaussian data in Model $3^{(II)}$ (column (c)). The powers of the tests by Chen and Qin (2010) (CQ) and Cai et al. (2014) (CLX) are also displayed.

Table 3

Numbers of identified BCR/ABL associated gene-sets for each GO category using different tests in conjunction with the BY procedure by Benjamini and Yekutieli (2001) for controlling FDR at 0.015. *Columns labeled by the name of tests records the number of identified gene-sets by the corresponding testing procedures, where* $\Psi_{\mathrm{ns},\alpha}$ *and* $\Psi^f_{\mathrm{ns},\alpha}$ *are the proposed non-studentized tests without and with screening, and CQ stands for the test by Chen and Qin (2010).*

GO Category	$\Psi_{\mathrm{ns},\alpha}$	$\Psi^f_{\mathrm{ns},\alpha}$ and CQ			Total	$\max_s p_s$	$\min_s p_s$	$\lfloor \bar{p}_s \rfloor$
		$\Psi^f_{\mathrm{ns},\alpha}$ only	Both	CQ only				
BP	601	0	956	560	1853	3050	20	150
CC	52	0	99	17	262	3145	19	280
MF	95	0	150	77	284	3040	19	157

discovery rate (FDR) under dependency to identify significant gene-sets.

We applied the above procedure to a human acute lymphoblastic leukemia (ALL) data set which is available at .The data contains gene expression levels from microarray experiments for patients suffering from ALL of either T-lymphocyte type or B-lymphocyte type leukemia. This data set was originally analyzed by Chiaretti et al. (2004) to provide insight into the genetic mechanism on ALL development and it was also analyzed by Dudoit et al. (2011) and Chen and Qin (2010) using different methodologies. To illustrate the proposed tests, we focus on the 75 patients of B-lymphocyte type leukemia, who were classified into two groups: 35 patients with BCR/ABL fusion and 40 patients with cytogenetically normal NEG, i.e., $n = 35$ and $m = 40$. We employed the approach in Gentleman et al. (2005) to conduct preliminary data processing. To focus on high dimensional scenarios, we also excluded gene-sets with $p_s \leq 19$. It remained $G = 1853, 262$, and 284 unique GO terms in the BP, CC, and MF categories, respectively. And the largest gene-set contained $p_s = 3050, 3145$, and 3040 genes in the BP, CC and MF categories, respectively. Given the complexity of the data processing and collection procedures, batch effects may exist and result in unreliable results. Therefore, we further employ the surrogate variable analysis (SVA) method proposed by Leek and Storey (2007) to remove the potential batch effects and other unwanted variations in the data. In summary, two surrogate variables were found by SVA and removed from the original ALL expression data. Identifications of gene-sets associated to the BCR/ABL fusion display biological insights on the development of B-lymphocyte type leukemia and provide lists of functional groups for potential clinical treatments. We aim to identify gene-sets with significantly different expression levels between the BCR/ABL and NEG groups for each of the three categories.

The sample size of the ALL data is relatively small comparing to the maximum p_s, we therefore employed the proposed two-sample non-studentized tests $\Psi_{\mathrm{ns},\alpha}$ and $\Psi^f_{\mathrm{ns},\alpha}$ in the analysis as suggested by simulation studies in Section 4. Based on empirical p-values, we further employed the BY procedure for controlling the FDR at 0.015 and identify significant gene-sets. For the proposed tests, we let $M = 50,000$ and used the sample covariance matrices to generate samples. Simulation studies in Section 4 have shown that the test by Cai et al. (2014) may inflate type I error rate for small sample size, we therefore only consider the test by Chen and Qin (2010) (CQ) as a reference. For each category, the numbers of gene-sets being identified are summarized in Table 3. All the gene-sets identified by the proposed two-step test $\Psi^f_{\mathrm{ns},\alpha}$ are also identified by CQ methods. This suggests that CQ test may over-detect some disease-associated gene-sets. Moreover, $\Psi^f_{\mathrm{ns},\alpha}$ found more disease associated gene-sets than $\Psi_{\mathrm{ns},\alpha}$, which reflects the power improvement of the proposed two-step testing procedure as discussed before.

By carefully investigating the gene-sets identified by both the proposed tests $\Psi_{\mathrm{ns},\alpha}$ and $\Psi^f_{\mathrm{ns},\alpha}$, we found that gene-sets GO:0005758 (mitochondrial intermembrane space) and GO:0004860 (protein kinase inhibitor activity) were identified as diseases-associated in the CC and MF categories. The functions of these two interesting gene-sets were recently studied and recognized associated with the development of ALL (Cui et al., 2009; Brinkmann and Kashkar, 2014). Particularly, the protein kinase inhibition has been considered to be essential for the mechanism of T-lymphocyte type ALL (Cui et al., 2009) and our finding suggests its connection with B-lymphocyte type ALL as well. The association of these gene-sets with the ALL may deserve further biological validations using the polymerase chain reaction.

6. Supplementary Materials

Web Appendices, which include proofs of the main theorems and additional numerical results referenced in Section 3 and 4 are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors thank the Co-Editor, the AE, and two anonymous referees for constructive comments and suggestions which have improved the presentation of the article. Jinyuan Chang was supported in part by the Fundamental Research Funds for the Central Universities of China (Grant No. JBK150501), NSFC (Grant No. 11501462), and the Center of Statistical Research and the Joint Lab of Data Science and Business Intelligence at Southwestern University of Finance and Economics. Wen Zhou was supported in part by NSF Grant IIS-1545994. Wen-Xin Zhou is the corresponding author.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. New York: Wiley-Interscience.

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, **6**, 311–329.

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

Brinkmann, K. and Kashkar, H. (2014). Targeting the mitochondrial apoptotic pathway: A preferred approach in hematologic malignancies? *Cell Death and Disease*, **5**, e1098.

Cai, T. T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society, Series B*, **76**, 349–372.

Chang, J., Tang, C. Y., and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics*, **41**, 2123–2148.

Chang, J., Tang, C. Y., and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *The Annals of Statistics*, **44**, 515–539.

Chen, S. X. and Qin, Y. (2010). A two sample test for high dimensional data with applications to gene-set testing. *The Annals of Statistics*, **38**, 808–835.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, **41**, 2786–2819.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., et al. (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**, 2771–2778.

Cui, J., Wang, Q., Wang, J., Lv, M., Zhu, N., Li, Y., et al. (2009). Basal c-Jun NH2-terminal protein kinase activity is essential for survival and proliferation of T-cell acute lymphoblastic leukemia cells. *Molecular Cancer Therapeutics*, **8**, 3214–3222.

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

Dudoit, S., Keles, S., and van der Laan, M. J. (2008). Multiple tests of associations with biological annotation metadata. *Institute of Mathematical Statistics. Collections*, **2**, 153–218.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics*, **1**, 107–129.

Gentleman, R., Irizarry, R. A., Carey, V. J., Dudoit, S., and Huber, W. (2005). *Bioinformtics and Computational Biology Solutions Using R and Bioconductor*. New York: Springer-Verlag.

James, D., Clymer, B. D., and Schmalbrock, P. (2001). Texture detection of simulated microcalcification susceptibility effects in magnetic resonance imaging of breasts. *Journal of Magnetic Resonance Imaging*, **13**, 876–881.

Katsani, K. R., Irimia, M., Karapiperis, C., Scouras, Z. G., Blencowe, B. J., Promponas, V. J., et al. (2014). Functional genomics evidence unearths new moonlighting roles of outer ring coat nucleoporins. *Scientific Reports*, **4**, 4655.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by "surrogate variable analysis." *PLoS Genetics*, **3**, e161.

Li, J. and Siegumnd, D. (2015). Higher criticism: p-values and criticism. *The Annals of Statistics*, **43**, 1323–1350.

Liu, W. and Shao, Q.-M. (2013). A Cramér moderate deviation theorem for Hotelling's T^2-statistic with applications to global tests. *The Annals of Statistics*, **41**, 296–322.

Martens, J. W., Nimmrich, I., Koenig, T., Look, M. P., Harbeck, N., Model, F., et al. (2005). Association of DNA methylation of phosphoserine aminotransferase with response to endocrine therapy in patients with recurrent breast cancer. *Cancer Research*, **65**, 4101–4117.

Recknor, J., Nettleton, D., and Reecy, J. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*, **24**, 192–201.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Science*, **102**, 15545–15550.

Thomas, M. A., Joshi, P. P., and Klaperb, R. D. (2011). Geneclass analysis of expression patterns induced by psychoactive pharmaceutical exposure in fathead minnow (Pimephales promelas) indicates induction of neuronal systems. *Comparative Biochemistry and Physiology C*, **155**, 109–120.

Wolen, A. R. and Miles, M. F. (2012). Identifying gene networks underlying the neurobiology of ethanol and alcoholism. *Alcohol Research: Current Reviews*, **34**, 306–317.

Zhong, P.-S., Chen, S. X., and Xu, M. (2013). Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *The Annals of Statistics*, **41**, 2820–2851.

Received February 2016. *Revised February* 2017.
Accepted February 2017.