The Annals of Statistics
2019, Vol. 47, No. 3, 1443–1474
https://doi.org/10.1214/18-AOS1720

TWO-SAMPLE AND ANOVA TESTS FOR HIGH DIMENSIONAL MEANS[1]

BY SONG XI CHEN, JUN LI AND PING-SHOU ZHONG

Peking University, Kent State University and University of Illinois at Chicago

This paper considers testing the equality of two high dimensional means. Two approaches are utilized to formulate L_2-type tests for better power performance when the two high dimensional mean vectors differ only in sparsely populated coordinates and the differences are faint. One is to conduct thresholding to remove the nonsignal bearing dimensions for variance reduction of the test statistics. The other is to transform the data via the precision matrix for signal enhancement. It is shown that the thresholding and data transformation lead to attractive detection boundaries for the tests. Furthermore, we demonstrate explicitly the effects of precision matrix estimation on the detection boundary for the test with thresholding and data transformation. Extension to multi-sample ANOVA tests is also investigated. Numerical studies are performed to confirm the theoretical findings and demonstrate the practical implementations.

1. Introduction. Modern statistical data in biological and financial studies are increasingly high dimensional, but with relatively small sample sizes. This is the so-called “large p, small n” paradigm, where classical multivariate procedures originally designed for fixed dimension problems may no longer be feasible. New methods which are adaptive to the “large p, small n” paradigm are needed.

An important high dimensional inferential task is to test the equality of the mean vectors between two populations. Let $\mathbf{X}_{i1}, \dots, \mathbf{X}_{in_i}$ be an IID sample drawn from a p-dimensional distribution F_i, for $i = 1$ and 2, respectively. The dimensionality p can be much larger than the sample sizes n_1 and n_2 so that $p/n_i \to \infty$. Let $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ be the mean and the covariance of F_i. The primary interest is testing

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \qquad \text{versus } H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2. \tag{1.1}$$

Hotelling’s T^2 test (Hotelling (1931)) is a classical test for the above hypotheses with fixed dimension p, and is still applicable if $p < n_1 + n_2 - 2$. However, Bai and Saranadasa (1996) showed that Hotelling’s test suffers from a power loss when

Received June 2016; revised July 2017.

[1]Supported by China’s National Key Research Special Program Grants 2016YFC020701 and 2015CB856000, China’s National Natural Science Foundation Grants 11131002, 71371016 and 71532001, and NSF Grants DMS-1309210, DMS-1309156 and DMS-1462156.

MSC2010 subject classifications. Primary 62H15; secondary 62E20.

Key words and phrases. ANOVA, data transformation, large p small n, sparse signals, thresholding, two-sample tests for means.

$p/(n_1 + n_2 - 2)$ approaches to 1 from below. When $p > n_1 + n_2 - 2$, the test is inapplicable as the sample covariance matrix is no longer invertible.

There have been proposals to modify Hotelling's T^2 statistic for high dimension. Bai and Saranadasa (1996) removed the inverse of the sample covariance from the Hotelling's formulation. Chen and Qin (2010) (CQ) considered a linear combination of U-statistics and showed that the corresponding test can operate under much relaxed conditions regarding p and sample sizes without assuming $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$. Srivastava, Katayama and Kano (2013) proposed using the diagonal matrix of the sample covariance to replace the sample covariance under the normality. Gregory et al. (2015) proposed using an average of the squared univariate two-sample t-statistics over p components as the test statistic. These four tests are basically all targeted on the L_2-norm or a weighted L_2-norm between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Cai, Liu and Xia (2014) (CLX) proposed a test based on the max-norm of marginal t-statistics. More importantly, they implemented a data transformation designed to increase the signal strength under sparsity as discovered by Hall and Jin (2010) in the one-sample innovated higher criticism test.

The L_2-norm based tests are known to be effective in detecting dense signals when the differences between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are located over a large number of components. However, the tests encounter a power loss under the sparse signal settings. Meanwhile, although Hall and Jin (2010) discovered that transforming data with a known precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ for the Gaussian data leads to enhanced signal strength and a lowered detection boundary, it is uncertain if these results can be maintained with estimated $\boldsymbol{\Omega}$ for the sub-Gaussian data.

With these as the motivation, this paper considers two modifications to CQ's test formulation. First of all, we apply a multi-level thresholding approach to removing the nonsignal bearing dimensions via a multi-layer of threshold levels to be adaptive to faint signals. The second alteration is to transform the data by an estimated precision matrix followed by the multi-level thresholding trying to enlarge the signal strength for more power gain. The idea of thresholding to remove the nonsignal bearing dimensions was advocated in Donoho and Johnstone (1994) for selecting significant wavelet coefficients and Fan (1996) for testing the mean with IID Gaussian distributed components; see also Ji and Jin (2012) for variable selection in high dimensional regression. In this paper, we show that a two-sample test based on the multiple thresholding levels (multi-level thresholding test) attains a detection boundary that coincides with the the optimal detection boundary for Gaussian data with identity covariance matrix. Furthermore, it is found that the detection boundary can be lowered by adding the data transformation in its formulation. A contribution of the current paper is that we explicitly establish the effect of precision matrix estimation on the detection boundary for the test with thresholding and data transformation.

In addition to the two-sample tests, we extend our analysis to the ANOVA test for m populations:

$$H_0^*: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_m \qquad \text{versus } H_1^*: \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \text{ for some } i \neq j, \tag{1.2}$$

where $\boldsymbol{\mu}_i$ is the mean of F_i for $i = 1, \dots, m$ and $m \geq 2$. Multi-level thresholding ANOVA tests with and without data transformation via the precision matrices are proposed. It is shown that the detection boundaries of the ANOVA tests resemble those of the two samples outlined above. As far as we are aware, the results regarding the detection boundary for ANOVA tests are the first of this kind in high dimensional testing for the means.

The rest of the paper is organized as follows. We analyze the power performance of the CQ test and the Oracle test under the sparse setting in Section 2. Thresholding tests without and with data transformation are proposed in Section 3 for detecting faint signals. Section 4 studies the multi-level thresholding tests. Extension to the ANOVA tests is provided in Section 5. Simulation results are presented in Section 6. Section 7 concludes the paper with discussions. Key technical details are reported in the Appendix, whereas additional proofs and simulation results, and an empirical study to select differentially expressed gene-sets for a human breast cancer data set are given in the Supplementary Material.

2. L_2-Norm based tests under sparsity. The statistic proposed by Chen and Qin (2010) (herein CQ test) can be written as $T_n = \sum_{k=1}^{p} T_{nk}$ where

$$\begin{aligned} T_{nk} = {} & \frac{1}{n_1(n_1-1)} \sum_{i \neq j}^{n_1} X_{1i}^{(k)} X_{1j}^{(k)} + \frac{1}{n_2(n_2-1)} \sum_{i \neq j}^{n_2} X_{2i}^{(k)} X_{2j}^{(k)} \\ & - \frac{2}{n_1 n_2} \sum_{i}^{n_1} \sum_{j}^{n_2} X_{1i}^{(k)} X_{2j}^{(k)}, \end{aligned} \tag{2.1}$$

and $X_{ij}^{(k)}$ denotes the kth component of $\mathbf{X}_{ij}$. It is readily shown that T_{nk} is unbiased to $(\mu_{1k} - \mu_{2k})^2$, a form of the signal in the kth dimension.

To facilitate simpler notation, we modify the statistic T_n by rescaling each T_{nk} by $\sigma_{1,kk}/n_1 + \sigma_{2,kk}/n_2$, the variance of $\bar{X}_1^{(k)} - \bar{X}_2^{(k)}$ where $\sigma_{i,kk}$ is the kth diagonal component of $\boldsymbol{\Sigma}_i$ $(i = 1, 2)$ and is assumed to be known. If $\sigma_{1,kk}$ and $\sigma_{2,kk}$ are unknown, we can use $\hat{\sigma}_{1,kk}/n_1 + \hat{\sigma}_{2,kk}/n_2$ where $\hat{\sigma}_{1,kk}$ and $\hat{\sigma}_{2,kk}$ are the usual sample variances at the kth dimension. This will make the CQ test invariant under the scale transformation; see Feng et al. (2015) for a related investigation. To expedite discussion, we assume $\sigma_{i,kk}$ are known and equal to one without loss of generality. This leads to a modified CQ statistic

$$\tilde{T}_n = n \sum_{k=1}^{p} T_{nk} \qquad \text{with } n = n_1 n_2/(n_1 + n_2). \tag{2.2}$$

Similar to Chen and Qin (2010), by defining

$$\rho_{kl} = \operatorname{Cov}\{\sqrt{n}(\bar{X}_1^{(k)} - \bar{X}_2^{(k)}), \sqrt{n}(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})\} = n\left(\frac{\sigma_{1,kl}}{n_1} + \frac{\sigma_{2,kl}}{n_2}\right), \tag{2.3}$$

the variance of $\tilde{T}_n$ under H_0 is $\sigma^2_{\tilde{T}_n,0} = 2p + 2\sum_{i\neq j}\rho^2_{ij}$, and under H_1 is

$$\sigma^2_{\tilde{T}_n,1} = 2p + 2\sum_{i\neq j}\rho^2_{ij} + 4n\sum_{k,l\in S_\beta}\delta_k\delta_l\rho_{kl}, \tag{2.4}$$

where $\delta_k = \mu_{1k} - \mu_{2k}$ and $S_\beta = \{k : \delta_k \neq 0\}$ is the set of nonzero δ_k locations.

Under a general multivariate model and some conditions on the covariance, it can be shown (Chen and Qin (2010)) that

$$\frac{\tilde{T}_n - \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2}{\sigma_{\tilde{T}_n,1}} \xrightarrow{d} \mathrm{N}(0, 1) \qquad \text{as } p \to \infty \text{ and } n \to \infty.$$

So the modified CQ test rejects H_0 if $\tilde{T}_n/\hat{\sigma}_{\tilde{T}_n,0} > z_\alpha$ where z_α is the upper α quantile of N(0, 1) and $\hat{\sigma}_{\tilde{T}_n,0}$ is a consistent estimator of $\sigma_{\tilde{T}_n,0}$.

To see the performance of the CQ test under the sparse setting, let $|S_\beta| = p^{1-\beta}$ where $|\cdot|$ represents the cardinality of a set and $\beta \in (0, 1)$ is the sparsity index. The power of the CQ test is

$$\beta_{\tilde{T}_n}(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|) = \Phi\left(-\frac{\sigma_{\tilde{T}_n,0}}{\sigma_{\tilde{T}_n,1}} z_\alpha + \frac{p^{1-\beta}n\bar{\delta}^2}{\sigma_{\tilde{T}_n,1}}\right), \tag{2.5}$$

where $\Phi(\cdot)$ is the cumulative distribution function of N(0, 1), and $\bar{\delta}^2 = \sum_{k\in S_\beta}\delta^2_k/p^{1-\beta}$ in (2.5) is the average standardized signal.

Since $\sigma^2_{\tilde{T}_n,1} \geq \sigma^2_{\tilde{T}_n,0}$, the first term within $\Phi(\cdot)$ in (2.5) is bounded. Then the power is largely determined by the second term

$$\mathrm{SNR}_{\tilde{T}_n} =: \frac{p^{1-\beta}n\bar{\delta}^2}{\sqrt{2p + 2\sum_{i\neq j}\rho^2_{ij} + 4n\sum_{k,l\in S_\beta}\delta_k\delta_l\rho_{kl}}}, \tag{2.6}$$

which is the signal-to-noise ratio since the numerator is the average signal strength and the denominator is the standard deviation of the test statistic under H_1. An inspection on (2.6) reveals that while the numerator of $\mathrm{SNR}_{\tilde{T}_n}$ is contributed only by those signal bearing dimensions, the standard deviation in the denominator is contributed by all T_{nk} including those with non-signals. Specifically, if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$, $\mathrm{SNR}_{\tilde{T}_n} = p^{1-\beta}n\bar{\delta}^2/\sqrt{2p + 4p^{1-\beta}n\bar{\delta}^2}$. If $\beta > 1/2$ and $\bar{\delta} = o(n^{-1/2}p^{\beta/2-1/4})$, $\mathrm{SNR}_{\tilde{T}_n} = o(1)$, which implies that the test has little power beyond the significant level. A reason for the power loss is that the variance of $\tilde{T}_n$ is inflated by those nonsignal bearing T_{nk}.

To put the above analysis in a broader perspective, we consider an Oracle who has the knowledge of the signal bearing set $S_\beta = \{k : \delta_k \neq 0\}$. The Oracle test statistic is

$$O_n = n\sum_{k\in S_\beta} T_{nk}. \tag{2.7}$$

Similar to the CQ test, the power of the Oracle test is determined by

$$\text{SNR}_{O_n} =: \frac{p^{1-\beta} n \bar{\delta}^2}{\sigma_{O_n}} \quad \text{where} \tag{2.8}$$

$$\sigma_{O_n}^2 = 2p^{1-\beta} + 2 \sum_{i \neq j \in S_\beta} \rho_{ij}^2 + 4n \sum_{k,l \in S_\beta} \delta_k \delta_l \rho_{kl}. \tag{2.9}$$

Since $\sigma_{O_n}^2 \ll \sigma_{\tilde{T}_n,1}^2$, $\text{SNR}_{O_n} \gg \text{SNR}_{\tilde{T}_n}$. Specially, if $\Sigma_1 = \Sigma_2 = \mathbf{I}_p$,

$$\text{SNR}_{O_n} = \frac{p^{1-\beta} n \bar{\delta}^2}{\sqrt{2p^{1-\beta} + 4p^{1-\beta} n \bar{\delta}^2}} = \frac{p^{(1-\beta)/2} n \bar{\delta}^2}{\sqrt{2 + 4n\bar{\delta}^2}} \tag{2.10}$$

that tends to infinity for $\beta > 1/2$ as long as $\bar{\delta} \asymp n^{-1/2} p^{\beta/4-1/4+\varepsilon}$ for any $\varepsilon > 0$, which is much smaller than $n^{-1/2} p^{\beta/2-1/4}$ required for the CQ test, indicating the test is able to detect much fainter signal.

3. Thresholding and data transformation. The power of the Oracle test is in its exclusion of the nonsignal bearing dimensions, whose locations are unknown in reality. Thresholding can be carried out to exclude those nonsignal bearing dimensions. Based on the large deviation results (Petrov (1995)), we use a thresholding level $\lambda_n(s) = 2s \log p$ for $s \in (0, 1)$ to strike a balance between removing nonsignal bearing T_{nk} and keeping those with signals. The thresholding statistic is

$$L_1(s) = \sum_{k=1}^{p} nT_{nk} I\{nT_{nk} + 1 > \lambda_n(s)\}, \tag{3.1}$$

where $I(\cdot)$ is the indicator function. A closely related statistic is

$$L_2(s) = \sum_{k=1}^{p} \{n(\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2 - 1\} I\{n(\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2 > \lambda_n(s)\}, \tag{3.2}$$

where $\bar{X}_i^{(k)}$ is the sample means of $X_{ij}^{(k)}$. Here, we use $L_n(s)$ to refer either $L_1(s)$ or $L_2(s)$ because both $L_1(s)$ and $L_2(s)$ have very similar properties. The proposed thresholding statistic can accommodate the column-wise dependence, which is defined via the α-mixing among the components of $\mathbf{X} = (X^{(1)}, \dots, X^{(p)})^T$.

We take a time series view on the dependence among the components of the high dimensional $\mathbf{X}$. For any integers $a < b$, let $\mathcal{F}_{\mathbf{X},(a,b)}$ to be the σ-algebra generated by $\{X^{(m)} : m \in (a, b)\}$ and define the α-mixing coefficient

$$\alpha_{\mathbf{X}}(k) = \sup_{m \in \mathcal{N}, A \in \mathcal{F}_{\mathbf{X},(1,m)}, B \in \mathcal{F}_{\mathbf{X},(m+k,\infty)}} |P(A \cap B) - P(A)P(B)|,$$

where $\mathcal{N}$ denotes the set of natural numbers. The following conditions are assumed in our analysis:

(C1): As $n \to \infty$, $p/n \to \infty$, $n_1/(n_1+n_2) \to \kappa$ and $\log p = o(n^{1/3})$.

(C2): Let $\mathbf{X}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{W}_{ij}$. There exists a positive constant H such that for $h \in [-H, H]^2$, $\mathrm{E}\{e^{h^T \cdot [(W_{ij}^{(k)})^2, (W_{ij}^{(l)})^2]}\} < \infty$ for $k \neq l$.

(C3): There exists a permutation $\tilde{X}$ of the components of $\mathbf{X}$ such that $\tilde{X}$ is α-mixing satisfying $\alpha_{\tilde{\mathbf{X}}}(k) \le C\alpha^k$ for some $\alpha \in (0, 1)$ and a positive constant C. Moreover, $\sum_{l=1}^{p} |\sigma_{i,kl}| < \infty$ for $i = 1$ or 2 such that ρ_{kl} defined in (2.3) satisfies $\sum_{l=1}^{p} |\rho_{kl}| < \infty$ for any $k \in \{1, \dots, p\}$.

Condition (C1) specifies the growth rate of p relative to n under which the large deviation results can be applied. This condition is not required in Chen and Qin (2010) because it does not involve the thresholding. (C2) assumes that $(X_{ij}^{(k)}, X_{ij}^{(l)})$ has a bivariate sub-Gaussian distribution, which is more general than the Gaussian distribution. Such conditions are commonly assumed in high dimensional analysis (Bickel and Levina (2008b), Zhong, Chen and Xu (2013); and Cai, Liu and Xia (2014)). Condition (C3) prescribes weak dependence among the column components of a permuted version $\tilde{X}$ of the original vector $\mathbf{X}$, which implies that $\tilde{\mathbf{X}}$ respects an ordering such that components closer to each other are more strongly correlated than those further apart. As the thresholded L_2-norm statistics are invariant under any permutation of the data, (C3) only requires that such permutation exists and there is no need to actually identify the permutation. The exponential decay for the α-mixing coefficients can be relaxed to polynomial decays with more involved proofs. The last two restrictions in (C3) are for the quantities of the original data rather than the permuted data due to the permutation invariance of the L_2 statistics. While the α-mixing is a common approach to accommodate the column-wise dependence, the physical dependence measure of Wu (2005) may be also used for modeling the weak dependence. A parallel development based on the physical dependence is possible. It is noted that (C3) is not required in Chen and Qin (2010) because their test formulation does not involve thresholding to remove the nonsignal bearing dimensions. The asymptotic normality of the test statistic in Chen and Qin (2010) is established by the martingale central limit theorem, which requires some moment conditions and a general multivariate linear innovation model. However, in one aspect, the paper's assumption is weaker, as we do not assume the multivariate linear innovation model. This is due to the thresholding which renders a need for such model.

In addition to the thresholding, we consider enhancing signal strength via data rotation. Motivated by Hall and Jin's (2010) study on data rotation via a banded Cholesky factor for the innovated HC test, and Cai, Liu and Xia's (2014) transformation via the CLIME estimator (Cai, Liu and Luo (2011)) in their max-norm based test, we will show that the signal enhancement can be achieved by transforming the data via an estimate of $\mathbf{\Omega} = \mathbf{\Sigma}_w^{-1} = (\omega_{ij})_{p\times p}$ where $\mathbf{\Sigma}_w = (1-\kappa)\mathbf{\Sigma}_1 + \kappa\mathbf{\Sigma}_2$.

Like Hall and Jin (2010), we consider a bandable covariance matrix class

$$V(\varepsilon_0, C, \nu) = \{\mathbf{\Sigma} : 0 < \varepsilon_0 \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le \varepsilon_0^{-1}, \nu > 0,$$
$$|\sigma_{ij}| \le C(1 + |i-j|)^{-(\nu+1)} \text{ for all } i, j : |i-j| \ge 1\},$$

which satisfies both the banding and thresholding conditions of Bickel and Levina (2008b). We assume the following regarding the covariance matrices:

(C4): Both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ belong to the matrix class $V(\varepsilon_0, C, \nu)$.

Although (C3) has assumed the weak dependence among the components of $\mathbf{X}_{ij}$, imposing (C4) ensures that the data transformed by $\mathbf{\Omega}$ are also weakly dependent. Although we assume the off-diagonal decay rates of $\mathbf{\Sigma}_i$ ($i = 1, 2$) are the same to expedite the technical analysis, all the results can be generalized to the case with different decay rates. In practice, the precision matrix $\mathbf{\Omega}$ needs to be estimated. Bickel and Levina (2008a) proposed estimating $\mathbf{\Omega}$ by banding the Cholesky factor matrices. Cai, Liu and Luo (2011) introduced the CLIME estimator based on the constrained L_1 minimization. As the CLIME estimator has the same rate of convergence as the estimator of Bickel and Levina (2008a) (Cai, Liu and Luo (2011)) when $\mathbf{\Omega}$ belongs to the bandable class, we use the latter to obtain a slightly simpler banding Cholesky estimator $\hat{\mathbf{\Omega}}_\tau$ as follows.

Define $\boldsymbol{Y}_{kl} = \sqrt{1-\kappa}\mathbf{X}_{1k} - \sqrt{\kappa}\mathbf{X}_{2l}$ for $k = 1, \dots, n_1$ and $l = 1, \dots, n_2$, where $\kappa = \lim_{n\to\infty} n_1/(n_1+n_2)$. Then $\operatorname{Var}(\boldsymbol{Y}_{kl}) = \mathbf{\Sigma}_w \equiv (1-\kappa)\mathbf{\Sigma}_1 + \kappa\mathbf{\Sigma}_2$. Let $\boldsymbol{Y}$ be an IID copy of $\boldsymbol{Y}_{kl}$ for any fixed k and l such that $\boldsymbol{Y} = (Y^{(1)}, \dots, Y^{(p)})^T$. For $j = 1, \dots, p$, define $\hat{Y}^{(j)} = \mathbf{a}_j^T \mathbf{W}^{(j)}$ where $\mathbf{a}_j = \{\operatorname{Var}(\mathbf{W}^{(j)})\}^{-1} \operatorname{Cov}(\hat{Y}^{(j)}, \mathbf{W}^{(j)})$ and $\mathbf{W}^{(j)} = (Y^{(1)}, \dots, Y^{(j-1)})^T$. Let $\varepsilon_j = Y^{(j)} - \hat{Y}^{(j)}$ and $d_j^2 = \operatorname{Var}(\varepsilon_j)$, and $\boldsymbol{A}$ be the lower triangular matrix with the jth row being $(\mathbf{a}_j^T, \mathbf{0}_{p-j+1})$ and $\boldsymbol{D} = \operatorname{diag}(d_1^2, \dots, d_p^2)$ where $\mathbf{0}_s$ means a vector of 0 with length s. Then the Cholesky decomposition is $\mathbf{\Omega} = (I - \boldsymbol{A})^T \boldsymbol{D}^{-1}(I - \boldsymbol{A})$.

Let $\boldsymbol{Y}_{n,kl} = \sqrt{n_2/(n_1+n_2)}\mathbf{X}_{1k} - \sqrt{n_1/(n_1+n_2)}\mathbf{X}_{2l} := (Y_{n,kl}^{(1)}, \dots, Y_{n,kl}^{(p)})^T$. Given a τ, regress $Y_{n,kl}^{(j)}$ on $\mathbf{Y}_{n,kl,-\tau}^{(j)} = (Y_{n,kl}^{(j-\tau)}, \dots, Y_{n,kl}^{(j-1)})^T$ to obtain the least square estimate of $\mathbf{a}_{j,\tau} = (a_{j-\tau}, \dots, a_{j-1})^T$:

$$\hat{\mathbf{a}}_{j,\tau} = \left(\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} \mathbf{Y}_{n,kl,-\tau}^{(j)} \mathbf{Y}_{n,kl,-\tau}^{(j)^T}\right)^{-1} \sum_{k=1}^{n_1}\sum_{l=1}^{n_2} \mathbf{Y}_{n,kl,-\tau}^{(j)} \mathbf{Y}_{n,kl}^{(j)}.$$

Put $\hat{\mathbf{a}}_j^T = (\mathbf{0}_{\tau-1}^T, \hat{\mathbf{a}}_{j,\tau}^T, \mathbf{0}_{p-j+1}^T)$ be the jth row of a lower triangular matrix $\hat{A}_\tau$ and $\hat{D}_\tau = \operatorname{diag}(d_{1,\tau}^2, \dots, d_{p,\tau}^2)$ where $d_{j,\tau}^2 = \sum_{k=1}^{n_1}\sum_{l=1}^{n_2}(Y_{n,kl}^{(j)} - \hat{\mathbf{a}}_{j,\tau}^T \mathbf{Y}_{n,kl,-\tau}^{(j)})^2/(n_1 n_2)$. Thus, a banded estimator of $\mathbf{\Omega}$ is

$$\hat{\mathbf{\Omega}}_\tau = (I - \hat{A}_\tau)^T \hat{D}_\tau^{-1}(I - \hat{A}_\tau). \tag{3.3}$$

The consistency of $\hat{\mathbf{\Omega}}_\tau$ to $\mathbf{\Omega}$ is established in Bickel and Levina (2008a). The method of Qiu and Chen (2015) may be used for selecting the suitable banding width τ, as well as a method outlined in Section 6.

It is noted that the bandable structure assumed in (C4) is not needed in the max-norm based test of Cai, Liu and Xia (2014). However, they require other conditions which are not taken by the current paper. More importantly, the two tests have different detection boundaries by comparing those established in Theorem 4 of this paper and that given in Section 1.4.1 of Donoho and Jin (2004). In particular, when the sparsity is moderate such that $1/2 < \beta < 3/4$, the proposed test can attain the Gaussian detection boundary while the max-norm test cannot. These aspects are reflected in the simulation results and reported in Figures 1–3 of Section 6.

The transformed thresholding test statistic based on $\{\hat{\mathbf{Z}}_{1i} =: \hat{\mathbf{\Omega}}_\tau \mathbf{X}_{1i} : 1 \le i \le n_1\}$ and $\{\hat{\mathbf{Z}}_{2i} =: \hat{\mathbf{\Omega}}_\tau \mathbf{X}_{2i} : 1 \le i \le n_2\}$ is

$$\hat{J}_n(s,\tau) = \sum_{k=1}^{p} \left\{ \frac{n(\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)})^2}{\hat{\omega}_{kk}} - 1 \right\} I \left\{ \frac{n(\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)})^2}{\hat{\omega}_{kk}} > \lambda_n(s) \right\}, \tag{3.4}$$

where $\bar{\hat{Z}}_i^{(k)}$ is the sample mean of $\hat{Z}_{ij}^{(k)}$.

It is worth discussing the sparsity of the transformed signals as it directly relates to the benefits of data transformation. As shown in Lemma 6 of the Supplementary Material, under (C4) and (C5), $\mathbf{\Omega}(\tau)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is sparse if $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is, where $\mathbf{\Omega}(\tau) = \{\omega_{ij} \mathrm{I}(|i-j| \le \tau)\}_{p \times p}$ is a banded version of $\mathbf{\Omega}$. The Supplementary Material also contains a numerical confirmation on the sparsity of $\mathbf{\Omega}(\tau)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. In practical problems such as time series data analysis, (C4) and (C5) can hold simultaneously if random noises are weakly dependent when they are farther apart in time, and signals are distributed randomly without appearing in clusters. However, in many applications such as genomic studies, variables are not naturally ordered even though the correlations among them are sparse. In such a case, there are algorithms, such as the one in Wagaman and Levina (2009), which permute the original random vectors so that the covariance of permuted data vector is more compliant to (C4). We note that the hypothesis (1.1) will not be affected by the permutation. When signals appear in clusters whose locations are randomly distributed as (C5), the benefits of data transformation may also be studied similar to Hall and Jin (2010).

THEOREM 1. *Assume Conditions* (C1)–(C4), $p = n^{1/\theta}$ *for* $0 < \theta < 1$ *and* $\tau \asymp (n^{-1} \log p)^{-1/\{2(\nu+1)\}}$, *then for any* $s \in (1 - \nu\theta/(\nu+1), 1)$,

$$\sigma_{J_n(s,\tau)}^{-1} \{ \hat{J}_n(s,\tau) - \mu_{J_n(s,\tau)} \} \xrightarrow{d} N(0,1),$$

where $\mu_{J_n(s,\tau)}$ *and* $\sigma_{J_n(s,\tau)}$ *are defined by* (A.3) *and* (A.4), *respectively.*

That requiring $\tau \asymp (n^{-1} \log p)^{-1/\{2(\nu+1)\}}$ is to allow consistent estimation of $\mathbf{\Omega}$. In addition, that imposing $p = n^{1/\theta}$ is to control the accumulated error due to the increase of dimension. Note that if θ is arbitrarily close to 0, p will grow exponentially fast with n. Moreover, a sufficient rate of convergence of $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\tau}}$ to $\boldsymbol{\Omega}$ requires the thresholding level s being larger than $1-\nu\theta/(\nu+1)$ which depends on the sparsity parameter ν for the bandable class in (C4). In practice, θ can be "estimated" by $\log(n)/\log(p)$. A liberal choice for the lower bound may be $1-\log(n)/\log(p)$ provided $\nu \gg 1$. To gain knowledge of ν, we may model the decay of σ_{ij} parametrically, for instance $|\sigma_{ij}|=\gamma^{|j-i|}$ for a $\gamma \in (0,1)$ or $|\sigma_{ij}|=(1+|j-i|)^{-\xi}$ for a $\xi>0$, which represent models with the exponential and the polynomial decay, respectively. Both models were considered in Hall and Jin (2010) and Qiu and Chen (2015). We can also use the generalized method of moment estimator advocated in He and Chen (2016) to estimate the parameters γ and ξ under both models, which can be translated to estimates of ν.

For the thresholding statistic $L_n(s)$ without the data transformation, less conditions than those in Theorem 1 are required in establishing its asymptotic normality, as shown in the following proposition. In particular, that $p=n^{1/\theta}$ and the lower bound on s are not needed.

PROPOSITION 1. *Assume Conditions* (C1)–(C3). *For any* $s \in (0,1)$,

$$\sigma_{L_n(s)}^{-1}\{L_n(s)-\mu_{L_n(s)}\} \xrightarrow{d} N(0,1), \tag{3.5}$$

where $\mu_{L_n(s)}$ *and* $\sigma_{L_n(s)}$ *are given by* (A.1) *and* (A.2), *respectively.*

The transformed thresholding test rejects H_0 at the level α if

$$\hat{J}_n(s,\tau) > z_\alpha \hat{\sigma}_{J_n(s,\tau),0} + \hat{\mu}_{J_n(s,\tau),0}, \tag{3.6}$$

where $\hat{\mu}_{J_n(s,\tau),0}$ and $\hat{\sigma}^2_{J_n(s,\tau),0}$ are, respectively, consistent estimators of

$$\mu_{J_n(s,\tau),0} = \left\{\frac{2}{\sqrt{2\pi}}(2s\log p)^{\frac{1}{2}}p^{1-s}\right\}\{1+o(1)\}$$

and

$$\sigma^2_{J_n(s,\tau),0} = \left\{\frac{2}{\sqrt{2\pi}}\{(2s\log p)^{\frac{3}{2}}+(2s\log p)^{\frac{1}{2}}\}p^{1-s}\right\}\{1+o(1)\},$$

satisfying

$$\mu_{J_n(s,\tau),0}-\hat{\mu}_{J_n(s,\tau),0} = o\{\sigma_{J_n(s,\tau),0}\} \quad \text{and} \quad \hat{\sigma}_{J_n(s,\tau),0}/\sigma_{J_n(s,\tau),0} \xrightarrow{p} 1. \tag{3.7}$$

Moreover, the asymptotic power of the transformed thresholding test is

$$\beta_{\hat{J}_n(s,\tau)}(\|\mu_1-\mu_2\|) = \Phi\left(-\frac{z_\alpha\sigma_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau)}}+\frac{\mu_{J_n(s,\tau)}-\mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau)}}\right),$$

which is mainly determined by

$$\mathrm{SNR}_{\hat{J}_n(s,\tau)} =: \frac{\mu_{J_n(s,\tau)}-\mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau)}}. \tag{3.8}$$

A test based on $L_n(s)$ without the data transformation can be proposed in an analogy to (3.6). It can be shown based on (3.5) that the asymptotic power of the thresholding only test is

$$\beta_{L_n(s)}(\|\mu_1 - \mu_2\|) = \Phi\left(-\frac{z_\alpha \sigma_{L_n(s),0}}{\sigma_{L_n(s)}} + \frac{\mu_{L_n(s)} - \mu_{L_n(s),0}}{\sigma_{L_n(s)}}\right),$$

which is determined by its signal-to-noise ratio

$$\mathrm{SNR}_{L_n(s)} =: \frac{\mu_{L_n(s)} - \mu_{L_n(s),0}}{\sigma_{L_n(s)}}, \tag{3.9}$$

where $\mu_{L_n(s),0}$ and $\sigma_{L_n(s),0}$ are the values of $\mu_{L_n(s)}$ and $\sigma_{L_n(s)}$ under H_0 and can be obtained by ignoring all the summation terms in (A.1) and (A.2).

We want to compare $\mathrm{SNR}_{\hat{J}_n(s,\tau)}$ with SNR_{L_n} with the following condition regarding the distribution of the signals:

(C5): The sparse elements of S_β with $\beta \in (1/2, 1)$ are randomly distributed among $\{1, 2, \ldots, p\}$.

THEOREM 2. *Under the conditions of Theorem* 1 *and* (C5), *SNR*$_{\hat{J}_n(s,\tau)} \geq$ *SNR*$_{L_n(s)}$ *with probability approaching to* 1.

Theorem 2 implies that the transformed thresholding test possesses a better power than that of the thresholding only test based on $L_n(s)$, which spells out the benefit of conducting the data transformation.

4. Multi-level thresholding. The thresholding tests with or without the data transformation in the last section depend on the thresholding level s. As shown in the proof of Proposition 1, if all the signals are strong such that $n\delta_k^2 > 2\log p$, a single level thresholding with $s = 1^-$ allows the test based on $L_n(s)$ to have the power of the Oracle test up to a slowly varying multi-log p function L_p. This echoes a result of Fan (1996) for Gaussian data with no dependence among the column components of the data. However, for weak signals, the thresholding has to be administrated at a smaller level, say $2s\log p$ for $s \in (0, 1)$. In this case, the single-level thresholding becomes inflexible. In order to adapt to the underlying signal strength, the higher criticism (HC) test (Donoho and Jin (2004)) that utilizes many levels of thresholding offers a solution.

We propose a multi-level thresholding statistic for the transformed data

$$M_{\hat{J}_n} = \max_{s \in \Lambda_n} \frac{\hat{J}_n(s,\tau) - \hat{\mu}_{J_n(s,\tau),0}}{\hat{\sigma}_{J_n(s,\tau),0}}, \tag{4.1}$$

where $\Lambda_n = \{s_k : s_k = n(\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)})^2/(2\hat{\omega}_{kk}\log p)$ for $k = 1, \ldots, p\} \cap (1 - \nu\theta/(\nu+1), 1 - \eta^\star)$ is the set of the thresholds for an arbitrarily small positive $\eta^\star$. It can be shown that the value of $M_{\hat{J}_n}$ is unchanged if we replace Λ_n

by $(1-\nu\theta/(\nu+1), 1-\eta^\star)$ where, similar to Theorem 1, the lower bound depends on θ and ν to ensure that the estimation error of $\hat{\mathbf{\Omega}}_\tau$ is negligible.

THEOREM 3. *Assume Conditions* (C1)–(C4) *and* (3.7), $p = n^{1/\theta}$ *for* $0 < \theta < 1$ *and* $\tau \asymp (n^{-1}\log p)^{-1/\{2(\nu+1)\}}$. *Then under* H_0,

$$P\left\{a(\log p)M_{\hat{J}_n} - b\left(\log p, \frac{\nu\theta}{\nu+1} - \eta^\star\right) \le x\right\} \to \exp(-e^{-x}),$$

where the two functions $a(y) = (2\log y)^{1/2}$ *and* $b(y, \nu\theta/(\nu+1) - \eta^\star) = 2\log y + 2^{-1}\log\log y - 2^{-1}\log[4\pi/\{1-\nu\theta/(\nu+1)+\eta^\star\}^2]$.

Theorem 3 implies an asymptotically α level test that rejects H_0 if

$$M_{\hat{J}_n} \ge \{q_\alpha + b(\log p, \nu\theta/(\nu+1) - \eta^\star)\}/a(\log p), \tag{4.2}$$

where q_α is the upper α quantile of the Gumbel distribution $\exp(-e^{-x})$.

We can also construct the multi-level thresholding statistic based on $L_n(s)$ (without data transformation) as

$$M_{L_n} = \max_{s\in\mathcal{S}_n} \frac{L_n(s) - \hat{\mu}_{L_n(s),0}}{\hat{\sigma}_{L_n(s),0}}, \tag{4.3}$$

where $\mathcal{S}_n = \{s_k \in (0, 1-\eta) : s_k = n(\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2/(2\log p), \text{for} k = 1, \dots, p\}$, $\hat{\mu}_{L_n(s),0}$ and $\hat{\sigma}_{L_n(s),0}$ are estimators of $\mu_{L_n(s),0}$ and $\sigma_{L_n(s),0}$ satisfying

$$\mu_{L_n(s),0} - \hat{\mu}_{L_n(s),0} = o\{\sigma_{L_n(s),0}\} \quad \text{and} \quad \hat{\sigma}_{L_n(s),0}/\sigma_{L_n(s),0} \xrightarrow{p} 1. \tag{4.4}$$

PROPOSITION 2. *Assume* (C1)–(C3) *and condition* (4.4). *Then under* H_0,

$$P\{a(\log p)M_{L_n} - b(\log p, \eta) \le x\} \to \exp(-e^{-x}),$$

where the two functions $a(\cdot)$ *and* $b(\cdot)$ *are defined in Theorem* 3.

The proposition implies an α level two-sample multi-level thresholding test without transformation that rejects H_0 if

$$M_{L_n} \ge \{q_\alpha + b(\log p, \eta)\}/a(\log p). \tag{4.5}$$

It is expected that both thresholding tests would encounter size distortion due to a slow convergence to the extreme value distribution and the second-order effects of the data dependence. To alleviate the problem, a parametric bootstrap approximation to the null distribution of $M_{\hat{J}_n}$ is considered. We first obtain $\hat{\mathbf{\Omega}}_\tau$ through the Cholesky decomposition based on the original samples. Two bootstrap resamples $\{\mathbf{Z}^*{}_{1i}\}_{i=1}^{n_1}$ and $\{\mathbf{Z}^*{}_{2i}\}_{i=1}^{n_2}$ are generated independently from $\mathrm{N}(0, \hat{\mathbf{\Omega}}_\tau)$. It is noted that there is no need to generate bootstrap versions of the original samples. Based on the two bootstrap resamples, a bootstrap version of $M_{\hat{J}_n}$ can be obtained based on (4.1), which we denote as $M^*_{\hat{J}_n}(s)$. After repeating this procedure B times, we obtain B bootstrap copies of $M_{\hat{J}_n}$: $M^{*,(1)}_{\hat{J}_n}(s), \ldots, M^{*,(B)}_{\hat{J}_n}(s)$, which are used to obtain an estimate to empirical null distribution of the transformed multi-level thresholding statistic and the upper α quantile.

We are to establish the detection boundary of the transformed multi-level thresholding test. Specially, we will consider the effect of estimating the precision matrix on the detection boundary, which has not been investigated in the literature. To define the detection boundary of the test, let

$$\underline{\omega} = \varliminf_{p\to\infty}\Big(\min_{1\le k\le p}\omega_{kk}\Big) \quad\text{and}\quad \bar{\omega} = \varlimsup_{p\to\infty}\Big(\max_{1\le k\le p}\omega_{kk}\Big).$$

Lemma 7 in the Supplementary Material (Chen, Li and Zhong (2018)) shows that $\underline{\omega}$ and $\bar{\omega} \ge 1$. The following two functions quantify the detection boundaries of the tests:

$$\varrho(\beta) = \begin{cases} \beta - 1/2, & 1/2 \le \beta \le 3/4, \\ (1-\sqrt{1-\beta})^2, & 3/4 < \beta < 1, \end{cases}$$

and for $0 < \nu\theta/(\nu+1) < 1$,

$$\varrho_{\nu,\theta}(\beta) = \begin{cases} \left(\sqrt{1-\dfrac{\nu\theta}{\nu+1}} - \sqrt{1-\beta-\dfrac{\nu\theta}{2\nu+2}}\right)^2, \\ \qquad\qquad \dfrac{1}{2} \le \beta \le \dfrac{3}{4} - \dfrac{\nu\theta}{4(\nu+1)}, \\ \beta - \dfrac{1}{2}, \qquad \dfrac{3}{4} - \dfrac{\nu\theta}{4(\nu+1)} \le \beta \le \dfrac{3}{4}, \\ (1-\sqrt{1-\beta})^2, \qquad \dfrac{3}{4} < \beta < 1. \end{cases} \tag{4.6}$$

Ingster (1997) showed that $r = \varrho(\beta)$ is the optimal detection boundary for uncorrelated Gaussian data in the sense that if (r,β) lays above the phase diagram $r = \varrho(\beta)$, there are tests whose probabilities of type I and type II errors converge to zero simultaneously as $n \to \infty$; and if (r,β) is below the phase diagram, no such test exists. Donoho and Jin (2004) showed that the HC test attains $r = \varrho(\beta)$ as the detection boundary when $\mathbf{X}_i$ are IID $\mathrm{N}(\mu, I_p)$. Zhong, Chen and Xu (2013) showed that the L_1 and L_2-versions of the HC tests also attain $r = \varrho(\beta)$ as the detection boundary for non-Gaussian data with column-wise dependence, and have more attractive power for (r,β) above the detection boundary.

Although the phase diagram $r = \varrho_{\nu,\theta}(\beta)$ has a similar functional form as the detect boundary established by Delaigle, Hall and Jin (2011) based on the marginal t-statistics, it explicitly demonstrates the effect of estimating $\mathbf{\Omega}$ on the detection boundary in the current setting. Specifically, for moderate sparsity such that $1/2 \le \beta < 3/4 - \nu\theta/(4\nu+4)$, it can be shown that $\varrho_{\nu,\theta}(\beta) > \varrho(\beta)$ implying $\varrho_{\nu,\theta}(\beta)$ has a higher detection boundary caused by having to estimate $\mathbf{\Omega}$. However, for high sparsity such that $\beta \ge 3/4 - \nu\theta/(4\nu+4)$, the two diagrams are identical.

THEOREM 4. *Assume Conditions* (C1)–(C5) *and* (3.7).

(a) *When* $\mathbf{\Omega}$ *is known, if* $r < \bar{\omega}^{-1}\varrho(\beta)$, *the sum of type I and II errors of the transformed multi-level thresholding test converges to* 1 *as* $\alpha \to 0$ *and* $n \to \infty$; *if* $r > \underline{\omega}^{-1}\varrho(\beta)$, *the sum of type I and II errors of the transformed multi-level thresholding test converges to zero when* $\alpha = \bar{\Phi}\{(\log p)^{\varepsilon}\} \to 0$ *for an arbitrarily small* $\varepsilon > 0$ *as* $n \to \infty$.

(b) *When* $\mathbf{\Omega}$ *is unknown and* $p = n^{1/\theta}$ *for* $0 < \theta < 1$, *if* $r < \bar{\omega}^{-1}\varrho_{\nu,\theta}(\beta)$, *the sum of type I and II errors of the transformed multi-level thresholding test converges* 1 *as* $\alpha \to 0$ *and* $n \to \infty$; *if* $r > \underline{\omega}^{-1}\varrho_{\nu,\theta}(\beta)$, *the sum of type I and II errors of the transformed multi-level thresholding test converges to zero when* $\alpha = \bar{\Phi}\{(\log p)^{\varepsilon}\} \to 0$ *for an arbitrarily small* $\varepsilon > 0$ *as* $n \to \infty$.

Part (a) of Theorem 4 shows that when $\mathbf{\Omega}$ is known, the transformed multi-level thresholding test has a lower detection boundary than $r = \varrho(\beta)$. The latter, as shown in the Supplementary Material, is the detection boundary for the multi-level thresholding only test. This means that the data transformation is able to detect weaker signals (at a given level of sparsity β) than the thresholding only test, realizing the benefit of having higher signal-to-noise ratio as reported in Theorem 2. Part (b) of the theorem reminisces Part (a) except that the estimation of $\mathbf{\Omega}$ leads to the use of $\varrho_{\nu,\theta}(\beta)$ with more stringent conditions in order to control the error of estimation. Hall and Jin (2010) has shown that similar to part (a), the data transformation can lower the detection boundary for Gaussian data with known covariance matrix. Here, we demonstrate that a modified detection boundary written in terms of $\varrho_{\nu,\theta}(\beta)$, which is achieved by the transformed multi-level thresholding test for sub-Gaussian data with estimated precision matrix.

As $r = \varrho(\beta)$ is the detection boundary for the multi-level thresholding only test, and $\varrho(\beta) \le \varrho_{\nu,\theta}(\beta)$ with the two functions being identical for $\beta \ge 3/4 - \nu\theta/(4\nu + 4)$ (high sparsity), Theorem 4 indicates that for the high sparsity case, doing the data-transformation in the multi-thresholding test achieves a lower detection boundary than that of the multi-thresholding test without the transformation. This means the extra labor involved in estimating $\mathbf{\Omega}$ pays off with better power performance. However, for moderate sparsity $[1/2 \le \beta < 3/4 - \nu\theta/(4\nu + 4)]$, it is uncertain which has a lower detection limit due to the facts that despite $\varrho(\beta) < \varrho_{\nu,\theta}(\beta)$, both $\underline{\omega}$ and $\bar{\omega}$ are larger than 1. In this case, one may just do the multi-level thresholding without the data transformation.

5. Extension to ANOVA tests. The ANOVA hypotheses (1.2) can be equivalently written as

$$H_0^*: \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_{(m-1)} - \boldsymbol{\mu}_m = 0 \qquad \text{versus } H_1^*: \boldsymbol{\mu}_l \neq \boldsymbol{\mu}_{(l+1)}$$
$$\text{for some } 1 \le l \le m - 1.$$

A measure of difference between μ_{lk} and $\mu_{(l+1)k}$, the means in the kth dimension of the lth and $(l+1)$th population means is

$$T_{nk}^{l(l+1)} = \frac{1}{n_l(n_l-1)}\sum_{i\neq j}^{n_l} X_{li}^{(k)} X_{lj}^{(k)} + \frac{1}{n_{l+1}(n_{l+1}-1)}\sum_{i\neq j}^{n_{l+1}} X_{(l+1)i}^{(k)} X_{(l+1)j}^{(k)} - \frac{2}{n_l n_{l+1}}\sum_{i}^{n_l}\sum_{j}^{n_{l+1}} X_{li}^{(k)} X_{(l+1)j}^{(k)},$$

which is an unbiased estimator of $\{\mu_{lk}-\mu_{(l+1)k}\}^2$. Then a statistic measure of $\sum_{l=1}^{m-1}\{\mu_{lk}-\mu_{(l+1)k}\}^2$ among all the adjacent populations, is $T_{nk,m} = \sum_{l=1}^{m-1} T_{nk}^{l(l+1)}$. We observe that if $m=2$, $T_{nk,2}=T_{nk}$ which is the two-sample case defined by (2.1). Like the two-sample case, we assume that $\sigma_{l,kk}=1$ for $l\in\{1,\dots,m\}$ and the following condition analogous to (C1):

(C1′): As $n, p\to\infty$, $n_{l(l+1)}$ are of the same order as n and $n_l/(n_l+n_{l+1})\to\kappa_l$ where $n_{l(l+1)} = (n_l n_{l+1})/(n_l+n_{l+1})$.

The thresholding test statistic for the ANOVA hypotheses (1.2) is

$$L_1^*(s) = \sum_{k=1}^{p}\sum_{l=1}^{m-1} n_{l(l+1)} T_{nk}^{l(l+1)} I\{n_{l(l+1)} T_{nk}^{l(l+1)} + 1 > \lambda_n(s)\}.$$

We now provide the connection between the ANOVA test and the two-sample test. For any fixed coordinate $1\le k\le p$, we place the adjacent distance mean measure $T_{nk}^{l(l+1)}$ to a vector of length $m-1$ as $\tilde{\mathbf{T}}_{nk} = (n_{12}T_{nk}^{12}, n_{23}T_{nk}^{23},\dots, n_{m-1m}T_{nk}^{(m-1)m})^T$. We then stack $\{\tilde{\mathbf{T}}_{nk}\}_{k=1}^p$ column-wise one after another to form a $(m-1)p$-dim vector $\mathbf{Q}_n$. Specifically, for $1\le j\le (m-1)p$, let $Q_n^{(j)} = n_{l(l+1)}T_{nk}^{l(l+1)}$ for $j=(k-1)(m-1)+l$ with $1\le l\le m-1$. With this notation, an equivalent form of $L_1^*(s)$ is

$$L_1^*(s) = \sum_{j=1}^{(m-1)p} Q_n^{(j)} I\{Q_n^{(j)}+1>\lambda_n(s)\}. \tag{5.1}$$

Similarly, for $1\le j\le (m-1)p$, define $U^{(j)} = n_{l(l+1)}(\bar{X}_l^{(k)} - \bar{X}_{l+1}^{(k)})$ for $j=(k-1)(m-1)+l$ with $1\le k\le p$ and $1\le l\le m-1$. Let $\mathbf{U}=(U^{(1)},\dots,U^{(m-1)p})^T$ be the stacked $(m-1)p$-dim vector. A thresholding test statistic based on $\mathbf{U}$, analogous to (3.2), is

$$L_2^*(s) = \sum_{j=1}^{(m-1)p} (\{U^{(j)}\}^2-1) I(\{U^{(j)}\}^2 > \lambda_n(s)). \tag{5.2}$$

Both $L_1^*(s)$ and $L_2^*(s)$ maintain the forms of $L_1(s)$ and $L_2(s)$. This implies that both versions of the thresholding ANOVA test statistics can be treated essentially as two-sample thresholding statistics with increased "dimensions" $(m-1)p$. In the following, we use $L_n^*(s)$ to refer either $L_1^*(s)$ or $L_2^*(s)$.

To develop the transformed ANOVA test statistic from the stacked $(m-1)p$-dimensional random vector $\mathbf{U}$, we first define $(m-1)\times(m-1)$ matrices

$$\mathbf{V}_1 = \begin{pmatrix} 1-\kappa_1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \qquad \mathbf{V}_m = \begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & \kappa_{m-1} \end{pmatrix} \quad \text{and}$$

for $2 \le b \le m-1$, $\mathbf{V}_b = (v_{b,(l,k)})$ contains a submatrix

$$\begin{pmatrix} v_{b,(b-1,b-1)} & v_{b,(b-1,b)} \\ v_{b,(b-1,b)} & v_{b,(b,b)} \end{pmatrix} = \begin{pmatrix} \kappa_{b-1} & -\sqrt{\kappa_{b-1}(1-\kappa_b)} \\ -\sqrt{\kappa_{b-1}(1-\kappa_b)} & 1-\kappa_b \end{pmatrix}$$

and all the other elements of $\mathbf{V}_b$ are 0. Then the covariance matrix of $\mathbf{U}$ is $\mathbf{\Sigma}^* = \sum_{b=1}^m (\mathbf{\Sigma}_b \otimes \mathbf{V}_b)$, where $\mathbf{\Sigma}_b = \mathrm{Var}(\mathbf{X}_{bi})$ is the covariance matrix of the bth population, and $\otimes$ denotes the Kronecker product. If $m=2$, $\mathbf{\Sigma}^*$ is reduced to $\mathbf{\Sigma}_w$ in the two-sample case defined in Section 3.

Similar to the two-sample case, we assume that $\mathbf{\Sigma}_b$ for $b \in \{1,\dots,m\}$ belongs to the family of $V(\varepsilon_0, C, \nu)$ defined in Section 3. Let $\mathbf{\Omega}^* = \mathbf{\Sigma}^{*-1}$, which is unknown in practice and can be estimated by the Cholesky banding estimator $\hat{\mathbf{\Omega}}_\tau^*$ similar to $\hat{\mathbf{\Omega}}_\tau$ for the two-sample case. We now transform the vector $\mathbf{U}$ by $\hat{\mathbf{\Omega}}_\tau^*$ to $\hat{\mathbf{Z}}^* = \hat{\mathbf{\Omega}}_\tau^* \mathbf{U}$. Then the thresholding ANOVA statistic based on the transformed data is

$$\hat{J}_n^*(s,\tau) = \sum_{k=1}^{(m-1)p} \left(\frac{\{\widehat{Z}^{*(k)}\}^2}{\hat{\omega}_{kk}^*} - 1 \right) I\left(\frac{\{\widehat{Z}^{*(k)}\}^2}{\hat{\omega}_{kk}^*} > \lambda_n(s) \right), \tag{5.3}$$

where $\hat{\omega}_{kk}^*$ is the kth diagonal element of $\hat{\mathbf{\Omega}}_\tau^*$. The corresponding multi-level thresholding ANOVA test statistic, similar to (4.1), is

$$M_{\hat{J}_n^*} = \max_{s \in \Lambda_n^*} \frac{\hat{J}_n^*(s,\tau) - \hat{\mu}_{J_n^*(s,\tau),0}}{\hat{\sigma}_{J_n^*(s,\tau),0}}, \tag{5.4}$$

where $\Lambda_n^* = \{s_k : s_k = (\widehat{Z}^{*(k)})^2/\{2\log\{(m-1)p\}\hat{\omega}_{kk}^*\}$ for $k = 1,\dots,(m-1)p\} \cap (1 - \nu\theta/(\nu+1), 1-\eta^\star)$ for an arbitrarily small $\eta^\star$, and $\hat{\mu}_{J_n^*(s,\tau),0}$ and $\hat{\sigma}^2_{J_n^*(s,\tau),0}$ are estimates which can be developed similar to (3.7) by replacing p by $(m-1)p$. As given in (4.2) for the two-sample case, the multi-level thresholding ANOVA test based on the transformed data rejects H_0 in (1.2) at an α level if $M_{J_n^*} > (q_\alpha + b\{\log(m-1)p, \nu\theta/(\nu+1) - \eta^*\})/a\{\log(m-1)p\}$.

To discuss the detection boundary of the transformed multi-level thresholding ANOVA test, we define

$$\underline{\omega}^* = \varliminf_{p\to\infty} \Big(\min_{1\le k\le (m-1)p} \omega_{kk}^* \Big) \quad \text{and} \quad \bar{\omega}^* = \varlimsup_{p\to\infty} \Big(\max_{1\le k\le (m-1)p} \omega_{kk}^* \Big),$$

where ω_{kk}^* are the diagonal elements of $\mathbf{\Omega}^*$, and require the following conditions analogous to conditions (C2)–(C5) of the two-sample tests:

(C2$'$): The random samples $\{\mathbf{X}_{ij}\}_{j=1}^{n_i}$ $(i = 1, \dots, m)$ are mutually independent. For each $i \in \{1, \dots, m\}$, $\mathbf{X}_{ij}$ satisfies the sub-Gaussian condition outlined in (C2).

(C3$'$): For each $i \in \{1, \dots, m\}$, the sequence of random variables $\{X_{ij}^{(l)}\}_{l=1}^{p}$ satisfies conditions in (C3).

(C4$'$): $\mathbf{\Sigma}_1, \dots, \mathbf{\Sigma}_m$ belong to the matrix class $V(\varepsilon_0, C, \nu)$.

(C5$'$): The nonzero differences in adjacent pairs of means defined by $S_\beta^* = \{j : j = (k-1)(m-1) + l, \delta_{kl} \neq 0, \text{ for } k = 1, \dots, p \text{ and } l = 1, \dots, m-1\}$ are randomly distributed among $\{1, 2, \dots, (m-1)p\}$.

THEOREM 5. *Assume Conditions* (C1$'$)–(C5$'$).

(a) *When* $\mathbf{\Omega}^*$ *is known, if* $r < \bar{\omega}^{*-1}\varrho(\beta)$, *the sum of type I and II errors of the transformed multi-level thresholding ANOVA test converges to* 1 *as* $\alpha \to 0$ *and* $n \to \infty$; *if* $r > \underline{\omega}^{*-1}\varrho(\beta)$, *the sum of type I and II errors of the transformed multi-level thresholding ANOVA test converges to zero when* $\alpha = \bar{\Phi}(\{\log(m-1)p\}^{\varepsilon}) \to 0$ *for an arbitrarily small* $\varepsilon > 0$ *as* $n \to \infty$.

(b) *When* $\mathbf{\Omega}^*$ *is unknown and* $p = n^{1/\theta}$ *for* $0 < \theta < 1$, *then if* $r < \bar{\omega}^{*-1}\varrho_{\nu,\theta}(\beta)$, *the sum of type I and II errors of the transformed multi-level thresholding ANOVA test converges to* 1 *as* $\alpha \to 0$ *and* $n \to \infty$; *if* $r > \underline{\omega}^{*-1}\varrho_{\nu,\theta}(\beta)$, *the sum of type I and II errors of the transformed multi-level thresholding ANOVA test converges to zero when* $\alpha = \bar{\Phi}(\{\log(m-1)p\}^{\varepsilon}) \to 0$ *for an arbitrarily small* $\varepsilon > 0$ *as* $n \to \infty$.

Theorem 5 demonstrates the detection boundary for the transformed multi-level thresholding ANOVA test for sub-Gaussian data with estimated precision matrix. This is consistent with the results obtained for the two sample test given in Theorem 4, which is expected as the two-sample hypotheses (1.1) is a special case of (1.2).

Similar to $M_{\hat{J}_n^*}$, we can also construct the multi-level thresholding ANOVA statistic:

$$M_{L_n^*} = \max_{s \in \mathcal{S}_n^*} \frac{L_n^*(s) - \hat{\mu}_{L_n^*(s),0}}{\hat{\sigma}_{L_n^*(s),0}}, \tag{5.5}$$

where $\mathcal{S}_n^* = \{s_{kl} \in (0, 1-\eta) : s_{kl} = n_{l(l+1)}(\bar{X}_l^{(k)} - \bar{X}_{l+1}^{(k)})^2/\{2\log(m-1)p\}, \text{ for } k = 1, \dots, p; l = 1, \dots, m-1\}$ for a small $\eta > 0$, $\hat{\mu}_{L_n^*(s),0}$ and $\hat{\sigma}^2_{L_n^*(s),0}$ are the estimators of mean and variance analogous to $\hat{\mu}_{J_n^*(s,\tau),0}$ and $\hat{\sigma}^2_{J_n^*(s,\tau),0}$ in (5.4). Under (C1) and (C2$'$)–(C3$'$), we can show that the detection boundary of the multi-level thresholding ANOVA test based on $M_{L_n^*}$ can be lowered by the data transformation.

6. Simulation study. We report results from simulation experiments which were designed to evaluate the empirical performance of the two multi-level thresholding tests defined in (4.1) and (4.3) with and without transformation (Mult2 and Mult1). We also experimented the test of Chen and Qin (2010) (CQ), the Oracle test in (2.7) and two tests proposed by Cai, Liu and Xia (2014) (CLX 1 and CLX2). The latter tests are based on the max-norm statistics without and with transformation $G(I) = \max_{1\le k\le p} n(\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2$ and $G(\hat{\mathbf{\Omega}}) = \max_{1\le k\le p} n(\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)})^2/\hat{\omega}_{kk}$, where $\hat{\omega}_{kk}$ were estimates of the diagonal elements of $\mathbf{\Omega}$. Instead of the CLIME estimator used by Cai, Liu and Luo (2011), we used $\hat{\omega}_{kk}$ from the Cholesky decomposition with banding to estimate $\mathbf{\Omega}$.

The two random samples were generated according to the model

$$\mathbf{X}_{ij} = \mathbf{\Sigma}_i^{1/2}\mathbf{Z}_{ij} + \boldsymbol{\mu}_i, \tag{6.1}$$

where the innovations $\mathbf{Z}_{ij}$ are IID p-dimensional random vectors with independent components such that $\mathrm{E}(\mathbf{Z}_{ij}) = 0$ and $\mathrm{Var}(\mathbf{Z}_{ij}) = I_p$. We considered two types of innovations: the Gaussian $\mathbf{Z}_{ij} \sim N(0, I_p)$ and the Gamma where each component of $\mathbf{Z}_{ij}$ is the standardized Gamma$(4, 0.5)$ such that it has zero mean and unit variance. We assigned $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$ under H_0 and under H_1, $\boldsymbol{\mu}_1 = 0$ and $\boldsymbol{\mu}_2$ had $[p^{1-\beta}]$ nonzero entries of equal value that were uniformly allocated among $\{1, \dots, p\}$. Here, $[a]$ denotes the integer part of a. The values of the nonzero entries were $\sqrt{2r\log p/n}$ with $r > 0$. The covariance matrices $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 =: \mathbf{\Sigma} = (\sigma_{ij})$ where $\sigma_{ij} = 0.4^{|i-j|}$ for $1 \le i, j \le p$.

In the simulation, the dimension p was chosen to be 200 and 600, and the two sample sizes (n_1, n_2) to be $(30, 40)$, $(60, 80)$ and $(90, 120)$, respectively. The sparsity parameter β was ranged from 0.3 to 0.8. To gain perspectives on the level of sparsity in the simulation, we note that for $p = 200$ with $\beta = 0.7$, there were $200^{1-0.7} \approx 5$ signals, and for $p = 600$, there were $600^{1-0.7} \approx 7$ signals, which were sparse indeed.

To select the banding width τ in the estimation of $\mathbf{\Omega}$, we used the cross-validation approach by Bickel and Levina (2008a). We divided a given dataset into two subsamples by repeated (N times) random data split. For the lth split, let $\hat{\mathbf{\Sigma}}_\tau^{(l)} = \{(I - \hat{A}_\tau^{(l)})'\}^{-1}\hat{D}_\tau^{(l)}(I - \hat{A}_\tau^{(l)})^{-1}$ be the Cholesky decomposition of $\mathbf{\Sigma}$ obtained from the first subsample and let $\mathbf{S}_n^{(l)}$ be the sample covariance obtained from the second subsample. Then τ is selected as

$$\hat{\tau} = \min_\tau \frac{1}{N}\sum_{l=1}^{N} \|\hat{\mathbf{\Sigma}}_\tau^{(l)} - \mathbf{S}_n^{(l)}\|_F, \tag{6.2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Table 1 reports the empirical sizes of the multi-thresholding tests with the data transformation (Mult2) and without the data transformation (Mult1), and Cai, Liu and Xia's max-norm tests with (CLX2) and without (CLX1) the data transformation. It also provides the empirical sizes for Mult1 and Mult2 with the bootstrap

TABLE 1

Empirical sizes of the proposed multi-thresholding tests with (*Mult*2) *and without data transformation* (*Mult*1), *Cai, Liu and Xia's max-norm tests with* (*CLX*2) *and without* (*CLX*1) *data transformation, Chen and Qin's test* (*CQ*) *and the Oracle test for the Gaussian and Gamma data. The numbers inside the parentheses are the size obtained via the parametric bootstrap procedure for the Mult*1 *and Mult*2 *tests, respectively*

p	(n_1, n_2)	Oracle	CQ	CLX1	CLX2	Mult1 (Mult1*)	Mult2 (Mult2*)
				Gaussian			
200	(30, 40)	0.068	0.052	0.039	0.022	0.094 (0.057)	0.044 (0.049)
	(60, 80)	0.067	0.065	0.048	0.026	0.099 (0.059)	0.033 (0.035)
	(90, 120)	0.066	0.063	0.042	0.032	0.103 (0.064)	0.063 (0.037)
400	(30, 40)	0.059	0.055	0.040	0.031	0.091 (0.063)	0.082 (0.058)
	(60, 80)	0.059	0.064	0.040	0.023	0.093 (0.051)	0.046 (0.052)
	(90, 120)	0.062	0.066	0.038	0.027	0.093 (0.071)	0.051 (0.041)
600	(30, 40)	0.058	0.053	0.037	0.054	0.095 (0.057)	0.129 (0.112)
	(60, 80)	0.050	0.049	0.047	0.033	0.080 (0.064)	0.061 (0.051)
	(90, 120)	0.054	0.054	0.043	0.036	0.098 (0.066)	0.072 (0.042)
				Gamma			
200	(30, 40)	0.068	0.062	0.034	0.027	0.097 (0.064)	0.056 (0.056)
	(60, 80)	0.065	0.063	0.036	0.022	0.103 (0.069)	0.031 (0.029)
	(90, 120)	0.061	0.055	0.040	0.027	0.084 (0.057)	0.046 (0.035)
400	(30, 40)	0.065	0.053	0.051	0.032	0.108 (0.050)	0.092 (0.078)
	(60, 80)	0.057	0.055	0.042	0.036	0.110 (0.051)	0.064 (0.043)
	(90, 120)	0.073	0.049	0.038	0.038	0.092 (0.047)	0.055 (0.042)
600	(30, 40)	0.068	0.054	0.041	0.059	0.114 (0.054)	0.134 (0.121)
	(60, 80)	0.057	0.056	0.039	0.031	0.090 (0.052)	0.061 (0.060)
	(90, 120)	0.059	0.052	0.041	0.037	0.099 (0.059)	0.073 (0.058)

approximation described in Section 4. We observe that the empirical sizes of the two thresholding tests tended to be larger than the nominal 5% level due to a slow convergence to the extreme value distribution. The bootstrap calibration can significantly improve the size. To make the power comparison fair, we preadjusted the nominal significant levels of all the tests such that their empirical sizes were all close to 0.05. We obtain the average empirical power curves with respect to r and β under each of the settings outlined above based on 1000 simulations.

Figure 1 displays the empirical power profiles of the proposed multi-thresholding tests with data transformation (Mult2) and without data transformation (Mult1), and Cai, Liu and Xia's max-norm tests with (CLX2) and without (CLX1) data transformation with respect to the signal strength r at a given level of sparsity $\beta = 0.7$ for Gaussian data. The power profile for the Gamma innovations are given in the Supplementary Material. Figures 2–3 provide alternative views of the power profiles of these tests where the powers are displayed with respect to the sparsity β at four levels of signal strength $r = 0.1, 0.2, 0.3$ and 0.4 for Gaussian data. The

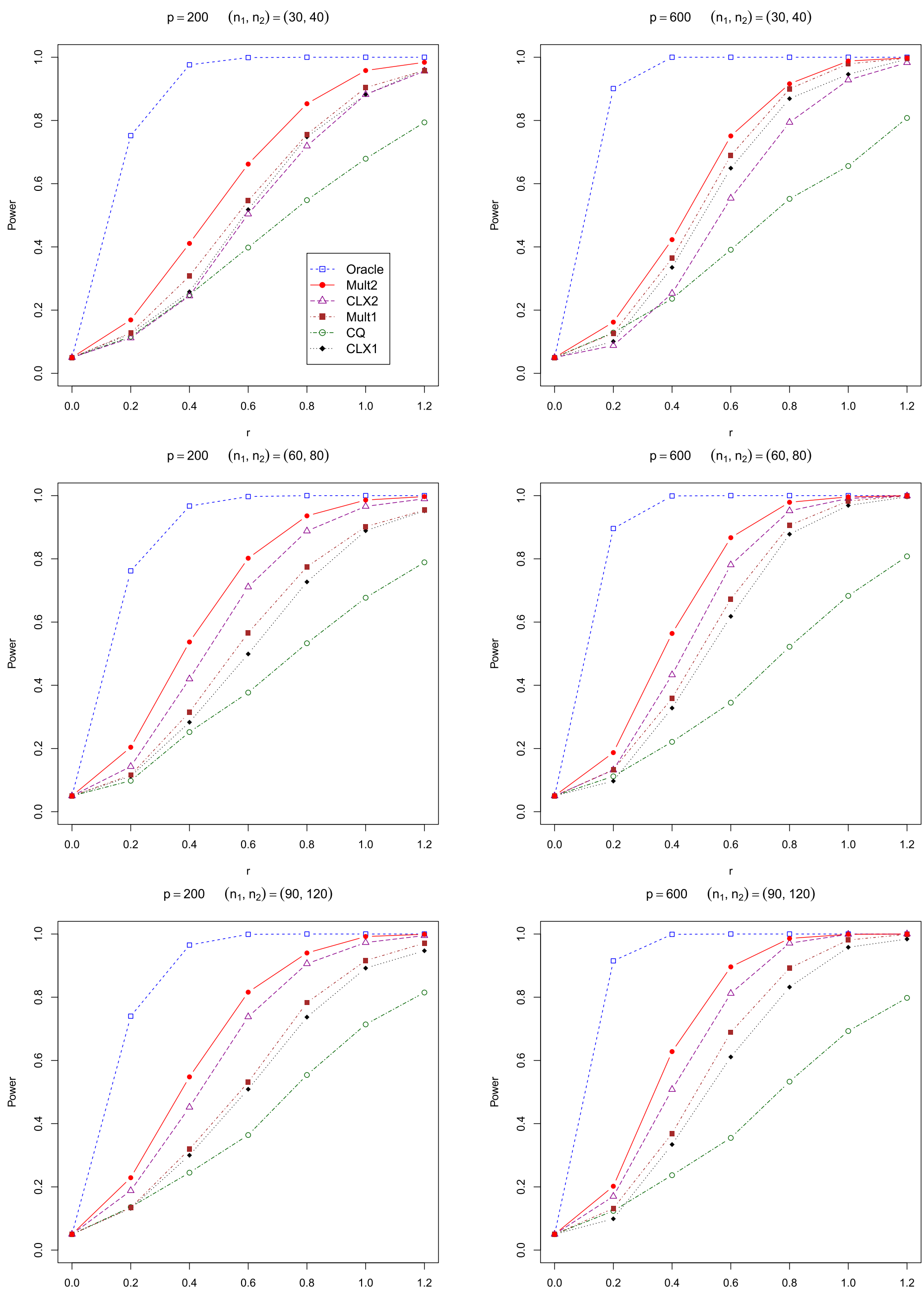

FIG. 1. *Average Power with respect to the signal strength r of the proposed multi-thresholding tests with* (*Mult*2) *and without data transformation* (*Mult*1), *Cai, Liu and Xia's max-norm tests with* (*CLX*2) *and without* (*CLX*1) *data transformation*, *Chen and Qin's test* (*CQ*) *and the Oracle test for the Gaussian data with the sparsity $\beta = 0.7$.*

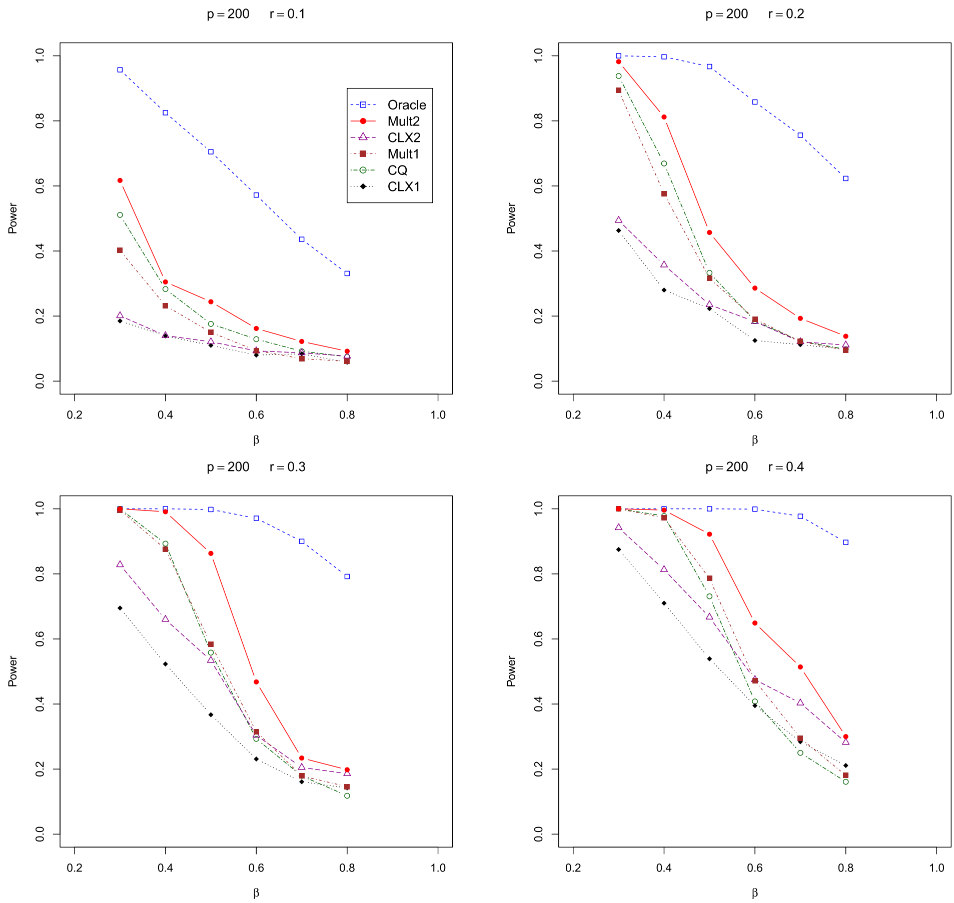

FIG. 2. *Average power with respect to the sparsity β of the proposed multi-thresholding tests with* (*Mult*2) *and without data transformation* (*Mult*1), *Cai, Liu and Xia's max-norm tests with* (*CLX*2) *and without* (*CLX*1) *data transformation, Chen and Qin's test* (*CQ*) *and the Oracle test for the Gaussian data with* $p = 200$, $n_1 = 60$ *and* $n_2 = 80$.

figures also report the powers of Chen and Qin's (2010) test (CQ) and the Oracle test to provide some bench marks for the performance.

The basic trend of Figure 1 was that the powers of all the tests were increasing as the signal strength r was increased, and that of Figures 2–3 is that the powers were decreasing as the sparsity was increased. These are all expected. It is also expected to see in each figure that the Oracle test had the best power among all the tests since all the dimensions bearing noise were removed in advance. A careful examination of the power profiles reveals that the two tests that employed data transformation (Mult2 and CLX2) were the top two performers among the non-Oracle tests especially for large sample sizes, indicating the effectiveness of the data transformation. Under the moderate sparsity, the thresholding test with data

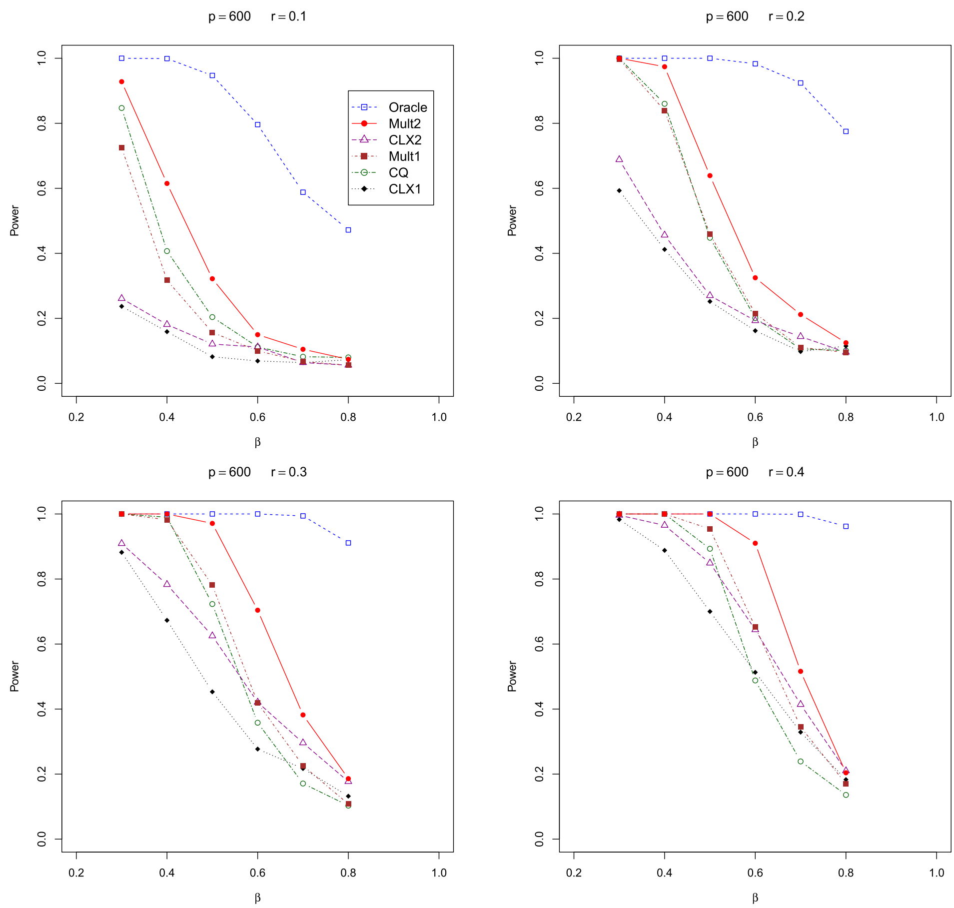

FIG. 3. *Average power with respect to the sparsity β of the proposed multi-thresholding tests with* (*Mult*2) *and without data transformation* (*Mult*1), *Cai, Liu and Xia's max-norm tests with* (*CLX*2) *and without* (*CLX*1) *data transformation, Chen and Qin's test* (*CQ*) *and the Oracle test for the Gaussian data with* $p = 600$, $n_1 = 60$ *and* $n_2 = 80$.

transformation (Mult2) had the best performance among all the non-Oracle test. Under the high sparsity with $\beta = 0.8$, the power of Mult2 was higher than that of the max-norm (CLX2) test with data transformation under faint signals, but only slightly lower than that of the CLX2 under strong signals. The CQ test and the CLX1 had the least power among the tests, with the CLX1 being more powerful than the CQ for the more sparse situation (large β) and vice versa for the faint signal case (smaller r). The CQ test was not designed for the sparse and faint signal settings of the simulation. The above features became more pronounced when we increase the dimensionality to $p = 600$ as shown in Figures 1 and 3.

Simulation studies were also conducted to demonstrate the performance of the multi-level thresholding ANOVA tests defined in (5.5) and (5.4) without and with transformation. For simplicity, we considered testing the equality of the mean vectors among three populations. Three random samples $\{\mathbf{X}_{1j}\}_{j=1}^{n_1}$, $\{\mathbf{X}_{2j}\}_{j=1}^{n_2}$ and $\{\mathbf{X}_{3j}\}_{j=1}^{n_3}$ were generated according to the multivariate model (6.1). We chose $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = 0$ under H_0, and under H_1, $\boldsymbol{\mu}_1 = 0$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ in total had $[(2p)^{0.4}]$ nonzero entries of equal value $\sqrt{2r\log(2p)/n}$, which were uniformly allocated among the $2p$ components of $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$. The covariance matrices were assigned such that $\boldsymbol{\Sigma}_1 = (0.4^{|i-j|})$, $\boldsymbol{\Sigma}_2 = (0.5^{|i-j|})$ and $\boldsymbol{\Sigma}_3 = (0.6^{|i-j|})$ for $1 \le i, j \le p$.

Table 2 displays the empirical sizes and powers of the multi-thresholding ANOVA tests without (Mult-A1) and with (Mult-A2) data transformation subject to different values of p, n_1, n_2, n_3 and r when $\mathbf{Z}_{ij} \sim \mathrm{N}(0, I_p)$ in the multivariate model (6.1). Similar to the two-sample test, the bootstrap calibration was implemented to improve the sizes of the testing procedures. Except slightly conservative sizes at the sample size of 40, others were quite close to the nominal significance level of 0.05. Despite the fact that the powers of both ANOVA tests were increased as the signal strength r was increased, the ANOVA test with data transformation had better performance than that without, which again confirms the advantageous of the transformation. Here, we only report the empirical sizes and powers based

TABLE 2
Empirical sizes and powers of the multi-thresholding ANOVA tests with (Mult-A2) and without (Mult-A1) data transformation for Gaussian data with $\boldsymbol{\Sigma}_1 = (0.4^{|i-j|})$, $\boldsymbol{\Sigma}_2 = (0.5^{|i-j|})$ and $\boldsymbol{\Sigma}_3 = (0.6^{|i-j|})$

			Power		
(p, n_1, n_2, n_3)	**Methods**	**Size**	$r = 0.1$	$r = 0.2$	$r = 0.4$
(100, 40, 40, 40)	Mult-A1	0.025	0.054	0.099	0.231
	Mult-A2	0.043	0.151	0.278	0.883
(100, 80, 80, 80)	Mult-A1	0.034	0.059	0.122	0.327
	Mult-A2	0.040	0.232	0.552	0.979
(100, 100, 100, 100)	Mult-A1	0.050	0.083	0.124	0.357
	Mult-A2	0.049	0.199	0.483	0.991
(200, 40, 40, 40)	Mult-A1	0.022	0.042	0.078	0.249
	Mult-A2	0.020	0.133	0.545	0.965
(200, 80, 80, 80)	Mult-A1	0.041	0.057	0.126	0.409
	Mult-A2	0.050	0.278	0.669	0.990
(200, 100, 100, 100)	Mult-A1	0.037	0.070	0.137	0.429
	Mult-A2	0.050	0.212	0.752	0.995
(400, 40, 40, 40)	Mult-A1	0.020	0.042	0.058	0.251
	Mult-A2	0.027	0.127	0.402	0.960
(400, 80, 80, 80)	Mult-A1	0.047	0.059	0.124	0.451
	Mult-A2	0.041	0.255	0.700	0.992
(400, 100, 100, 100)	Mult-A1	0.033	0.076	0.133	0.727
	Mult-A2	0.041	0.277	0.492	0.999

on the Gaussian data. Results based on other distributions such as the Gamma are similar and thus omitted due to the space limitation.

7. Discussion. This paper investigates the benefits of multi-level thresholding in a L_2 formulation of the test statistics with or without the data transformation via the precision matrix. It shows that the thresholding combined with the data transformation leads to a very powerful test procedure for the high sparsity case. In this case, thresholding with the data transformation has better power than the thresholding alone formulation. Our study confirms the benefit of the transformation discovered by Hall and Jin (2010) for the higher criticism test and Cai, Liu and Xia (2014) for the max-norm based test. The proposed thresholding tests can be viewed as improvements of the test of Chen and Qin (2010) when the signals are sparse and faint. The CQ test is similar to the max-norm test without data transformation, except that it is based on the L_2-norm. Generally speaking, the max-norm test works better for more sparse and stronger signals whereas the CQ test is for denser but fainter signals. These aspects were confirmed by our simulations. A reason for the proposed test having better power than the CLX test is that the proposed test has both thresholding and data transformation whereas CLX test has only the data transformation.

Hall and Jin (2010) discovered that by transforming data with $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, the signal strength of the high dimensional testing problem can be enhanced. However, they only considered the case of a *known* covariance matrix with Gaussian data. There is much uncertainty if the signal and the eventual power enhancement would be maintained when $\mathbf{\Omega}$ has to be estimated for non-Gaussian data. We embark on this task by estimating $\mathbf{\Omega}$, and shows that with the estimated precision matrix in high dimension, a modified version of the detection boundary established in Hall and Jin (2010) can be reached. Moreover, the effect of estimating precision matrix on the detection boundary is also considered. Cai, Liu and Xia (2014) (CLX) studied the relative performance of several forms of data transformation for testing two-sample means based on the maximum norm. Although they confirmed the advantage of data transformation via the precision matrix discovered in Hall and Jin (2010), CLX did not have results on the detection boundary. In relation to Zhong, Chen and Xu (2013) (ZCX), this paper studies a new test that combines data transformation and thresholding, which was not considered in ZCX. The current paper also extends the proposed method to ANOVA test, which was not considered in the one-sample study of ZCX.

APPENDIX: TECHNICAL DETAILS

We provide the technical detail in the proofs of Theorems 1 and 4. Proofs of other theorems are relegated to the Supplementary Material to this paper (Chen, Li and Zhong (2018)).

A.1. Proof of Theorem 1. Recall that $\mathbf{\Omega} = \mathbf{\Sigma}_w^{-1} = (\omega_{ij})_{p\times p}$ where $\mathbf{\Sigma}_w = (1-\kappa)\mathbf{\Sigma}_1 + \kappa\mathbf{\Sigma}_2$ and $\kappa = \lim_{n\to\infty} n_1/(n_1+n_2)$. We first assume $\mathbf{\Omega}$ is known to gain insight on the test. Rather than transforming the data via $\mathbf{\Omega}$, we transform it via $\mathbf{\Omega}(\tau) = \{\omega_{ij}\mathrm{I}(|i-j|\le\tau)\}_{p\times p}$, a banded version of $\mathbf{\Omega}$ for an integer τ between 1 and $p-1$. There are two reasons to use $\mathbf{\Omega}(\tau)$. One is that the signal enhancement is facilitated mainly by elements of $\mathbf{\Omega}$ close to the main diagonal. Another is that banding maintains the α-mixing structure of the transformed data provided $k-2\tau\to\infty$. Since both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ have off-diagonal entries decaying to zero at polynomial rates, $\mathbf{\Omega}$ has the same decay rate (Gröchenig and Leinert (2006), Jaffard (1990), Sun (2005)), so the transformed data are still weakly dependent. Two transformed samples are

$$\{\mathbf{Z}_{1j}(\tau) = \mathbf{\Omega}(\tau)\mathbf{X}_{1j} : 1\le j\le n_1\} \quad \text{and} \quad \{\mathbf{Z}_{2j}(\tau) = \mathbf{\Omega}(\tau)\mathbf{X}_{2j} : 1\le j\le n_2\}.$$

Let $\varpi_{kk}(\tau) = \mathrm{Var}\{\sqrt{n}(\bar{Z}_1^{(k)}(\tau) - \bar{Z}_2^{(k)}(\tau))\}$ be the counterpart of $n(\sigma_{1,kk}/n_1 + \sigma_{2,kk}/n_2)$ for the transformed data where $\bar{Z}_i^{(k)}(\tau) = n_i^{-1}\sum_{j=1}^{n_i} Z_{ij}^{(k)}(\tau)$ for $i=1,2$. Lemmas 5 and 7 in Chen, Li and Zhong (2018) show that there exists a constant $C>1$ such that $\varpi_{kk}(\tau) = \omega_{kk} + O(\tau^{-C})$ and $\omega_{kk}>1$.

The transformed thresholding statistic can be constructed by replacing $\mathbf{X}_{ij}$ with $\mathbf{Z}_{ij}(\tau)$ in either (3.1) or (3.2). Although both have similar properties, the latter of which has the form

$$\begin{aligned} J_n(s,\tau) = \sum_{k=1}^{p}&\left\{\frac{n(\bar{Z}_1^{(k)}(\tau) - \bar{Z}_2^{(k)}(\tau))^2}{\varpi_{kk}(\tau)} - 1\right\} \\ &\times I\left\{\frac{n(\bar{Z}_1^{(k)}(\tau) - \bar{Z}_2^{(k)}(\tau))^2}{\varpi_{kk}(\tau)} > \lambda_n(s)\right\}\end{aligned}$$

and is easier to work with, which we will present in the following.

Let $\delta_{\mathbf{\Omega}(\tau)} = (\delta_{\mathbf{\Omega}(\tau),1},\dots,\delta_{\mathbf{\Omega}(\tau),p})^T$ where $\delta_{\mathbf{\Omega}(\tau),k} = \sum_l \mathbf{\Omega}_{kl}(\tau)\delta_l = \sum_{l\in S_\beta}\omega_{kl}\times\delta_l\mathrm{I}(|k-l|\le\tau)$ is the difference between the transformed means in the kth dimension. Using $\mathbf{Z}_{ij}(\tau) = \mathbf{\Omega}(\tau)\mathbf{X}_{ij}$ and $\sum_l|\omega_{kl}|<\infty$, for a given constant C, $Z_{ij}^{(k)}(\tau) = \sum_l \omega_{kl}X_{ij}^{(l)}\mathrm{I}(|k-l|<\tau)$. Since $X_{ij}^{(l)}$ is sub-Gaussian for any $l = 1,\dots,p$, $Z_{ij}^{(l)}(\tau)$ is sub-Gaussian by Hölder inequality and mathematical induction. Hence, the large derivation results can be applied to derive the mean and variance of $J_n(s,\tau)$ by following the similar steps for the mean and variance of the thresholding statistic. Derivations in Lemmas 2–3 show that the mean and variance of the thresholding test statistic $L_n(s)$ is

$$\begin{aligned}\mu_{L_n(s)} = \Bigg(&\frac{2}{\sqrt{2\pi}}(2s\log p)^{\frac{1}{2}}p^{1-s} + \sum_{k\in S_\beta}\{n\delta_k^2 I(n\delta_k^2 > 2s\log p) \\ &+ (2s\log p)\bar{\Phi}(\eta_k^-)I(n\delta_k^2 < 2s\log p)\}\Bigg)\{1+o(1)\} \quad \text{and}\end{aligned} \tag{A.1}$$

$$\sigma^2_{L_n(s)} = \bigg(\frac{2}{\sqrt{2\pi}}\{(2s\log p)^{\frac{3}{2}} + (2s\log p)^{\frac{1}{2}}\}p^{1-s} + \sum_{k,l\in S_\beta}(4n\delta_k\delta_l\rho_{kl} + 2\rho_{kl}^2)$$

$$\times I(n\delta_k^2 > 2s\log p)I(n\delta_l^2 > 2s\log p) + \sum_{k\in S_\beta}(2s\log p)^2\bar{\Phi}(\eta_k^-) \tag{A.2}$$

$$\times I(n\delta_k^2 < 2s\log p)\bigg)\{1+o(1)\},$$

where $\bar{\Phi} = 1 - \Phi$ and $\eta_k^- = (2s\log p)^{1/2} - n^{1/2}\delta_k$.

By replacing δ_k by $\delta_{\mathbf{\Omega}(\tau),k}$ and S_β by $S_{\mathbf{\Omega}(\tau),\beta}$ in (A.1) and (A.2) where after the transformation, δ_k becomes $\delta_{\mathbf{\Omega}(\tau),k}$ and the set S_β including nonzero signals becomes $S_{\mathbf{\Omega}(\tau),\beta}$, the mean and variance of $J_n(s,\tau)$ are

$$\mu_{J_n(s,\tau)} = \bigg(\frac{2}{\sqrt{2\pi}}(2s\log p)^{\frac{1}{2}}p^{1-s} + \sum_{k\in S_{\mathbf{\Omega}(\tau),\beta}}\bigg\{\frac{n\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)}I\bigg(\frac{n\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)} > 2s\log p\bigg)$$

$$+ (2s\log p)\bar{\Phi}(\eta^-_{\mathbf{\Omega}(\tau)k})I\bigg(n\frac{\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)} < 2s\log p\bigg)\bigg\}\bigg)\{1+o(1)\} \quad \text{and} \tag{A.3}$$

$$\sigma^2_{J_n(s,\tau)} = \bigg(\frac{2}{\sqrt{2\pi}}\{(2s\log p)^{\frac{3}{2}} + (2s\log p)^{\frac{1}{2}}\}p^{1-s}$$

$$+ \sum_{k,l\in S_{\mathbf{\Omega}(\tau),\beta}}\bigg(4n\frac{\delta_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}^{1/2}(\tau)}\frac{\delta_{\varpi(\tau),l}\rho_{\mathbf{\Omega},kl}}{\varpi_{ll}^{1/2}(\tau)} + 2\rho^2_{\mathbf{\Omega},kl}\bigg)$$

$$\times I\bigg(\frac{n\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)} > 2s\log p\bigg)I\bigg(\frac{n\delta^2_{\mathbf{\Omega}(\tau),l}}{\varpi_{ll}(\tau)} > 2s\log p\bigg) \tag{A.4}$$

$$+ \sum_{k\in S_{\mathbf{\Omega}(\tau),\beta}}(2s\log p)^2\bar{\Phi}(\eta^-_{\mathbf{\Omega}(\tau)k})I\bigg(n\frac{\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)} < 2s\log p\bigg)\bigg)\{1+o(1)\},$$

where $S_{\mathbf{\Omega}(\tau),\beta} = \{k : \delta_{\mathbf{\Omega}(\tau),k} \neq 0\}$ is the set of locations of the nonzero signals $\delta_{\mathbf{\Omega}(\tau),k}$, $\eta^-_{\mathbf{\Omega}(\tau)k} = (2s\log p)^{1/2} - n^{1/2}\delta_{\mathbf{\Omega}(\tau),k}/\varpi_{kk}(\tau)^{1/2}$ and $\rho_{\mathbf{\Omega},kl} = \mathrm{Cov}\{\sqrt{n}\times(\bar{Z}_1^{(k)}(\tau) - \bar{Z}_2^{(k)}(\tau))/\sqrt{\varpi_{kk}(\tau)}, \sqrt{n}(\bar{Z}_1^{(l)}(\tau) - \bar{Z}_2^{(l)}(\tau))/\sqrt{\varpi_{ll}(\tau)}\}$.

We first establish the asymptotic normality of $J_n(s,\tau)$ where the banding parameter τ is chosen to be a slowly varying function. To this end, we first show that both $\{Z_{1i}^{(k)}(\tau)\}_{k=1}^p$ and $\{Z_{2i}^{(k)}(\tau)\}_{k=1}^p$ are α-mixing sequences. By condition (C3), $\{X_{1j}^{(k)}\}_{k=1}^p$ and $\{X_{2j}^{(k)}\}_{k=1}^p$ are α-mixing sequences. Then any event $A \in \sigma(\mathcal{F}^{(1)}_{\mathbf{X},(1,a)}, \mathcal{F}^{(2)}_{\mathbf{X},(1,a)})$ and $B \in \sigma(\mathcal{F}^{(1)}_{\mathbf{X},(a+k,\infty)}, \mathcal{F}^{(2)}_{\mathbf{X},(a+k,\infty)})$, $|P(A\cap B) - P(A)P(B)| \to 0$ as $k \to \infty$. By the relationship between $\mathbf{Z}_{1i}(\tau)$ and $\mathbf{X}_{1i}$, for any τ,

$$Z_{1i}^{(a)}(\tau) \in \sigma(\mathcal{F}^{(1)}_{\mathbf{X},(a-\tau,a+\tau)}) \quad \text{and} \quad Z_{1i}^{(a+k)}(\tau) \in \sigma(\mathcal{F}^{(1)}_{\mathbf{X},(a+k-\tau,a+k+\tau)}).$$

Then as long as $k-2t\to\infty$, $|P(A'\cap B')-P(A')P(B')|\to 0$ for any $A'\in\sigma(\mathcal{F}^{(1)}_{\mathbf{Z},(1,a)},\mathcal{F}^{(2)}_{\mathbf{Z},(1,a)})$ and $B'\in\sigma(\mathcal{F}^{(1)}_{\mathbf{Z},(a+k,\infty)},\mathcal{F}^{(2)}_{\mathbf{Z},(a+k,\infty)})$. It follows that $\alpha_{\mathbf{Z}_1(\tau)}(k)=\alpha_{\mathbf{X}_1}(k-2t)$ if $k>2t$. Therefore, $\alpha_{\mathbf{Z}_1(\tau)}(k)\to 0$ as $k-2t\to\infty$ where $\alpha_{\mathbf{Z}_1(\tau)}$ is the α-mixing coefficient for the sequence $\{Z^{(k)}_{1j}(\tau)\}_{k=1}^p$. Similarly, it can be shown that $\alpha_{Z_2(\tau)}(k)\to 0$ as $k-2t\to\infty$. Thus, both $\{Z^{(k)}_{1i}(\tau)\}_{k=1}^p$ and $\{Z^{(k)}_{2i}(\tau)\}_{k=1}^p$ are α-mixing sequences. Then the asymptotic normality of $J_n(s,\tau)$ can be established by applying the Bernstein's blocking method as we have done in the proof of Proposition 1. To further establish the normality of $\hat{J}_n(s,\tau)$, we note that our $\hat{J}_n$ can be written as

$$\begin{aligned}\hat{J}_n &= J_n+\sum_{k=1}^p\left(\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}}-\frac{S_{nk}}{\varpi_{kk}}\right)I\left(\frac{S_{nk}}{\varpi_{kk}}>\lambda_n\right)\\ &\quad+\sum_{k=1}^p\left(\frac{S_{nk}}{\varpi_{kk}}+1\right)\left[I\left(\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}}>\lambda_n\right)-I\left(\frac{S_{nk}}{\varpi_{kk}}>\lambda_n\right)\right]\\ &\quad+\sum_{k=1}^p\left(\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}}-\frac{S_{nk}}{\varpi_{kk}}\right)\left[I\left(\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}}>\lambda_n\right)-I\left(\frac{S_{nk}}{\varpi_{kk}}>\lambda_n\right)\right]\\ &= J_n+\mathrm{I}+\mathrm{II}+\mathrm{III},\end{aligned}$$

where $\hat{S}_{nk}=n(\bar{\hat{Z}}^{(k)}_1-\bar{\hat{Z}}^{(k)}_2)^2$ and $S_{nk}=n(\bar{Z}^{(k)}_1(\tau)-\bar{Z}^{(k)}_2(\tau))^2$. To show the asymptotic normality of $\hat{J}_n$ under H_0, we only need to show that $\mathrm{I}/\sigma_{J_n,0}=o_p(1)$ and $\mathrm{II}/\sigma_{J_n,0}=o_p(1)$ since III is smaller order of I or II.

We first consider I, which can be bounded by

$$\mathrm{I}\le\max_{1\le k\le p}\left|\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}}-\frac{S_{nk}}{\varpi_{kk}}\right|\sum_{k=1}^p\mathrm{I}\left(\frac{S_{nk}}{\varpi_{kk}}>\lambda_n\right).$$

Using $\mathrm{E}\{\sum_{k=1}^p\mathrm{I}(\frac{S_{nk}}{\varpi_{kk}}>\lambda_n)\}=\sum_{k=1}^p\mathrm{P}(\frac{S_{nk}}{\varpi_{kk}}>\lambda_n)=O(\frac{p^{1-s}}{\sqrt{2s\log p}})$ by Lemma 1 in Chen, Li and Zhong (2018), we have $\sum_{k=1}^p\mathrm{I}(\frac{S_{nk}}{\varpi_{kk}}>\lambda_n)=O_p(\frac{p^{1-s}}{\sqrt{2s\log p}})$. Recall that $\hat{S}_{nk}=n\{\sum_l\hat{\omega}_{kl}(\bar{X}^{(l)}_1-\bar{X}^{(l)}_2)\}^2$ and $S_{nk}=n\{\sum_l\omega_{kl}(\tau)(\bar{X}^{(l)}_1-\bar{X}^{(l)}_2)\}^2$. Then it can be derived that

$$\begin{aligned}&\max_k\left|\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}}-\frac{S_{nk}}{\varpi_{kk}}\right|\\ &\quad\le M\max_l n\big(\bar{X}^{(l)}_1-\bar{X}^{(l)}_2\big)^2\max_k\left\{\sum_{l=1}^p|\hat{\omega}_{kl}-\omega_{kl}|+\tau^{-a}+O\big(\tau^{-C}\big)\right\},\end{aligned}$$

where $M>0$, $a>0$ and we use the fact that $\varpi_{kk}=\omega_{kk}+O(\tau^{-C})$ from Lemma 5 in the Supplementary Material. From the fact that $\max_l n(\bar{X}^{(l)}_1-\bar{X}^{(l)}_2)^2=$

$O_p(\log p)$ and $\max_k \sum_{l=1}^p |\hat{\omega}_{kl} - \omega_{kl}| = O_p[(\frac{\log p}{n})^{\nu/(2\nu+2)}]$ (see Bickel and Levina (2008a)),

$$\max_k \left| \frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}} \right| \sum_{k=1}^p \mathrm{I}\left(\frac{S_{nk}}{\varpi_{kk}} > \lambda_n \right) = O_p\{L_p p^{1-s} n^{-\nu/(2\nu+2)} + \log p(\tau^{-a} + \tau^{-C})\},$$

where L_p and τ are slowly varying functions. We can choose τ such that $\log p(\tau^{-a} + \tau^{-C}) = o(1)$. Therefore, we have $\mathrm{I} = O_p(L_p p^{1-s} n^{-\nu/(2\nu+2)})$. By assumption that $p = n^{1/\theta}$ and $s > 1 - \nu\theta/(\nu+1)$, then $\mathrm{I}/\sigma_{J_n,0} = o_p(1)$.

For the second term II, we have

$$\begin{aligned} \mathrm{II} \leq{}& \max_k \left| \frac{S_{nk}}{\varpi_{kk}} + 1 \right| \max_k I\left\{ \frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} > \lambda_n \right\} \sum_{k=1}^p I\left\{ \left| \frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}} \right| > \left| \frac{S_{nk}}{\varpi_{kk}} - \lambda_n \right| \right\} \\ &+ \max_k \left| \frac{S_{nk}}{\varpi_{kk}} + 1 \right| \max_k I\left\{ \frac{S_{nk}}{\varpi_{kk}} > \lambda_n \right\} \sum_{k=1}^p I\left\{ \left| \frac{S_{nk}}{\varpi_{kk}} - \frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} \right| > \left| \frac{S_{nk}}{\varpi_{kk}} - \lambda_n \right| \right\} \\ :={}& \mathrm{II}_1 + \mathrm{II}_2. \end{aligned}$$

Because the proofs for II_1 and II_2 are similar, we only show II_2. Note that $\max_k |\frac{S_{nk}}{\varpi_{kk}} + 1| \leq 1 + \max_k \frac{(\sum_l \omega_{kl}(\tau))^2}{\varpi_{kk}} \max_l n(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})^2 = O_p(\log p)$. And

$$\begin{aligned} &\sum_{k=1}^p I\left\{ \left| \frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}} \right| > \left| \frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s) \right| \right\} \\ &\quad \leq \sum_{k=1}^p I\left(\left| \frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}} \right| > h \right) + \sum_{k=1}^p I\left(\left| \frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s) \right| < h \right), \end{aligned} \tag{A.5}$$

where the second indicator function on the right-hand side satisfies $\mathrm{E}\{\sum_{k=1}^p I(|\frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s)| < h)\} = \frac{h}{\sqrt{2s\log p}} p^{1-s}$. So, in (A.5), $\sum_{k=1}^p I(|\frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s)| < h) = O_p(\frac{h}{\sqrt{2s\log p}} p^{1-s})$. For $\sum_{k=1}^p I(|\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}}| > h)$ in (A.5), we first notice that

$$\left| \frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}} \right| \leq M \max_l n(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})^2 \left\{ \sum_{l=1}^p |\hat{\omega}_{kl} - \omega_{kl}| \right\} + o(1).$$

Then

$$\begin{aligned} &\mathrm{E}\left\{ \sum_{k=1}^p I\left(\left| \frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}} \right| > h \right) \right\} \\ &\quad \leq \sum_{k=1}^p \mathrm{P}\left(\sum_{l=1}^p |\hat{\omega}_{kl} - \omega_{kl}| > \frac{h}{MnT^2} \right) + \sum_{k=1}^p \mathrm{P}\left(\max_l |\bar{X}_1^{(l)} - \bar{X}_2^{(l)}| > T \right), \end{aligned}$$

where, if $T = C\sqrt{\log p/n}$, then $\sum_{k=1}^{p} \mathrm{P}(\max_l |\bar{X}_1^{(l)} - \bar{X}_2^{(l)}| > T) \le p^{2-C} \to 0$, for sufficient large C. If $h = C^* \log p (\frac{\log p}{n})^{\nu/(2\nu+2)}$, there exists a $a > 0$ such that $\sum_{k=1}^{p} \mathrm{P}(\sum_{l=1}^{p} |\hat{\omega}_{kl} - \omega_{kl}| > \frac{h}{MnT^2}) = \sum_{k=1}^{p} \mathrm{P}(\sum_{l=1}^{p} |\hat{\omega}_{kl} - \omega_{kl}| > M'(\frac{\log p}{n})^{\nu/(2\nu+2)}) \le p^{1-a}$. Therefore, by choosing C^* large enough such that $a > \nu\theta/(2\nu+2)$, $\sum_{k=1}^{p} I(|\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}}| > h) = O_p(p^{1-a}) = o_p(pn^{-\nu/(2\nu+2)})$ for $p = n^{1/\theta}$. With $h = C^* \log p (\frac{\log p}{n})^{\nu/(2\nu+2)}$, $\sum_{k=1}^{p} I(|\frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s)| < h) = O_p(L_p n^{-\nu/(2\nu+2)} p^{1-s})$. In addition, $\max_k I\{\frac{S_{nk}}{\varpi_{kk}} > \lambda_n(s)\} = O_p(p^{-s})$. Therefore, $\mathrm{II}_2 = o_p(L_p p^{1-s} n^{-\nu/(2\nu+2)})$. Similarly, one can show that $\mathrm{II}_1 = o_p(L_p \times p^{1-s} n^{-\nu/(2\nu+2)})$. In summary, $\mathrm{II}/\sigma_{J_n,0} = o_p(\mathrm{I}/\sigma_{J_n,0}) = o_p(1)$. The asymptotic normality of $\hat{J}_n$ under H_1 can be established based on similar derivations. This completes the proof of Theorem 1.

A.2. Proof of Theorem 4. We first consider $\mathbf{\Omega}$ is known. We know that the power of the transformed thresholding test is determined by $\mathrm{SNR}_{J_n(s,\tau)}$. Recall that for $k \in S_\beta$, $\underline{\omega}\delta_k^2 \le \delta_{\mathbf{\Omega}(\tau),k}^2/\varpi_{kk}(\tau) \le \bar{\omega}\delta_k^2$. If $M_1 = \sum_{k \in S_\beta} \{n\underline{\omega}\delta_k^2 I(n\underline{\omega}\delta_k^2 > 2s\log p) + (2s\log p)\bar{\Phi}(\eta_k^-) I(n\underline{\omega}\delta_k^2 < 2s\log p)\}$,

$$\mathrm{SNR}_{J_n(s,\tau)} \ge M_1/V_1 \tag{A.6}$$

where

$$\begin{aligned} V_1^2 = {} & \frac{2}{\sqrt{2\pi}}\{(2s\log p)^{\frac{3}{2}} + (2s\log p)^{\frac{1}{2}}\} p^{1-s} \\ & + \sum_{k \in S_\beta} (2s\log p)^2 \bar{\Phi}(\eta_k^-) I(n\underline{\omega}\delta_k^2 < 2s\log p) \\ & + \sum_{k,l \in S_\beta} (4n\underline{\omega}^2 \delta_k \delta_l \rho_{\mathbf{\Omega},kl} + 2\rho_{\mathbf{\Omega},kl}^2) I(n\underline{\omega}\delta_k^2 > 2s\log p) I(n\underline{\omega}\delta_l^2 > 2s\log p). \end{aligned}$$

Note that M_1/V_1 is the signal-to-noise ratio of the thresholding test without the transformation with $n\underline{\omega}\delta_k^2 = 2\underline{\omega}r\log p$. From the proof of Proposition 3 in the Supplementary Material, $M_1/V_1 \to \infty$ as long as s is properly chosen and $\underline{\omega}r > \varrho(\beta)$. Therefore, $\{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}\}/\sigma_{J_n(s,\tau),1} \to \infty$, as long as $\underline{\omega}r > \varrho(\beta)$. This establishes the upper bound of the detectable region.

To show the second statement in part (a) of Theorem 4, we notice that the maximal transformed thresholding test is of asymptotic α level. Therefore, it is sufficient to show that its power tends to 1 above the detection boundary as $n \to \infty$ and $\alpha \to 0$. To this end, we notice that

$$\begin{aligned} \mathrm{P}(M_{J_n} \ge G_\alpha | H_1) & \ge \mathrm{P}\left(\frac{J_n(s,\tau) - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),0}} \ge G_\alpha | H_1\right) \\ & \ge \Phi\left(-\frac{\sigma_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),1}} G_\alpha + \frac{M_1}{V_1}\right), \end{aligned}$$

where $G_\alpha = \{q_\alpha + b(\log p, \eta)\}/a(\log p)$. We choose $\alpha_n = \bar{\Phi}\{(\log p)^\varepsilon\} \to 0$ as $p \to \infty$ for $\varepsilon > 0$ such that $G_\alpha = O\{(\log\log p)^{1/2}\}$. If $\underline{\omega} r > \varrho(\beta)$, we can find a s satisfying one of cases given in the proof of Proposition 3 in the Supplementary Material such that the second term in $\Phi(\cdot)$ dominates and tends to infinity, which leads to $\Phi(\cdot) \to 1$.

Then we consider the first statement in part (a) of Theorem 4. Note that $\mathrm{SNR}_{J_n(s,\tau)} \le M_2/V_2$, where $M_2 = \sum_{k\in S_\beta}\{n\bar{\omega}\delta_k^2 I(n\bar{\omega}\delta_k^2 > 2s\log p) + (2s\log p) \times \bar{\Phi}(\eta_k^-) I(n\bar{\omega}\delta_k^2 < 2s\log p)\}$ and

$$\begin{aligned}V_2^2 = {}& \frac{2}{\sqrt{2\pi}}\{(2s\log p)^{\frac{3}{2}} + (2s\log p)^{\frac{1}{2}}\}p^{1-s}\\ &+ \sum_{k\in S_\beta}(2s\log p)^2\bar{\Phi}(\eta_k^-)I(n\bar{\omega}\delta_k^2 < 2s\log p)\\ &+ \sum_{k,l\in S_\beta}(4n\bar{\omega}^2\delta_k\delta_l\rho_{\mathbf{\Omega},kl} + 2\rho^2_{\mathbf{\Omega},kl})I(n\bar{\omega}\delta_k^2 > 2s\log p)I(n\bar{\omega}\delta_l^2 > 2s\log p).\end{aligned}$$

Note that M_2/V_2 is the signal-to-noise ratio of the thresholding test with $n\bar{\omega}\delta_k^2 = 2\bar{\omega} r\log p$, which converges to 0 for any s if $\bar{\omega} r < \varrho(\beta)$, that is, $\{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}\}/\sigma_{J_n(s,\tau),1} \to 0$. Similar to the proof of Proposition 3, we can show that $M_{J_n} = \max_{s\in T_n}\tilde{J}_n(s)\{1 + o_p(1)\}$, where $\tilde{J}_n(s) = (J_n(s) - \mu_{J_n(s,\tau),1})/\sigma_{J_n(s,\tau),1}$. Since $\mathrm{P}\{a(\log p)\max_{s\in T_n}\tilde{J}_n(s) - b(\log p, c) \le x\} \to \exp(-e^{-x})$, where $c = \max(\eta - r + 2r\sqrt{1-\eta} - \beta, \eta)I(r < 1-\eta) + \max(1-\beta, \eta)I(r > 1-\eta)$. Then, similar to the proof in Proposition 2, we have $\mathrm{P}(M_{J_n} \ge G_\alpha | H_1) = \alpha\{1 + o(1)\} \to 0$, which implies that the type II error tends to 1 as $\alpha \to 0$.

Next, we consider $\mathbf{\Omega}$ is unknown. Let $G_\alpha^\star = \{q_\alpha + b(\log p, \nu\theta/(\nu+1) - \eta^\star)\}/a(\log p)$. If we choose $\alpha_n = \bar{\Phi}\{(\log p)^\varepsilon\} \to 0$ as $p \to \infty$ for any small number $\varepsilon > 0$, $G_\alpha^\star = O\{(\log\log p)^{1/2}\}$. We only show that if $r > \underline{\omega}^{-1}\varrho_{\nu,\theta}(\beta)$, the sum of type I and II of $M_{\hat{J}_n}$ converges to 0, since the proof that the sum of type I and II of $M_{\hat{J}_n}$ tends to 1 if $r < \bar{\omega}^{-1}\varrho_{\nu,\theta}(\beta)$ is similar to that for M_{J_n}. Notice that

$$\begin{aligned}\mathrm{P}(M_{\hat{J}_n} \ge G_\alpha^\star | H_1) &\ge \mathrm{P}\biggl(\frac{\hat{J}_n(s,\tau) - \hat{\mu}_{J_n(s,\tau),0}}{\hat{\sigma}_{J_n(s,\tau),0}} \ge G_\alpha^\star | H_1\biggr)\\ &\ge \mathrm{P}\biggl\{\biggl(\frac{J_n(s,\tau) - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),0}} + \frac{\mu_{J_n(s,\tau),0} - \hat{\mu}_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),0}}\\ &\qquad + o_p(1)\biggr)\frac{\sigma_{J_n(s,\tau),0}}{\hat{\sigma}_{J_n(s,\tau),0}} \ge G_\alpha^\star | H_1\biggr\},\end{aligned} \tag{A.7}$$

where we have used the fact that if $p = n^{1/\theta}$ for $0 < \theta < 1$, $(\hat{J}_n(s,\tau) - \mu_{J_n(s,\tau),0})/\sigma_{J_n(s,\tau),0} = (J_n(s,\tau) - \mu_{J_n(s,\tau),0})/\sigma_{J_n(s,\tau),0} + o_p(1)$ given in the proof of Theorem 3. Moreover, as shown in Zhong, Chen and Xu (2013), with $p = n^{1/\theta}$

for $0 < \theta < 1$, $\{\mu_{J_n(s,\tau),0} - \hat{\mu}_{J_n(s,\tau),0}\}/\sigma_{J_n(s,\tau),0} \to 0$, and $\sigma_{J_n(s,\tau),0}/\hat{\sigma}_{J_n(s,\tau),0} \to 1$. Then the probability in (A.7) is determined by

$$\frac{J_n(s,\tau) - \mu_{J_n(s,\tau),0}}{G^\star_\alpha \sigma_{J_n(s,\tau),0}} = \left(\frac{J_n(s,\tau) - \mu_{J_n(s,\tau),1}}{G^\star_\alpha \sigma_{J_n(s,\tau),1}} + \frac{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}}{G^\star_\alpha \sigma_{J_n(s,\tau),1}}\right) \times \frac{\sigma_{J_n(s,\tau),1}}{\sigma_{J_n(s,\tau),0}},$$

where $(J_n(s,\tau) - \mu_{J_n(s,\tau),1})/(G^\star_\alpha \sigma_{J_n(s,\tau),1}) = o_p(1)$, and $\sigma_{J_n(s,\tau),1} > \sigma_{J_n(s,\tau),0}$. As long as we can show $(\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0})/(G^\star_\alpha \sigma_{J_n(s,\tau),1}) \to \infty$, (A.7) tends 1. From (A.6), we only need to show that with properly chosen s, $M_1/(G^\star_\alpha V_1) \to \infty$. As shown in Theorem 3, we need to choose $s \in (1 - \nu\theta/(\nu+1), 1)$ if $\mathbf{\Omega}$ is unknown such that the asymptotic normality of the transformed thresholding test with $\hat{\mathbf{\Omega}}$ can be established. The modification on the detection boundary can be derived by adding the additional restriction $s > 1 - \nu\theta/(\nu+1)$ on the four cases in the proof of Proposition 3. Similar to Delaigle, Hall and Jin (2011), and Zhong, Chen and Xu (2013), the modified detection boundary is given by (4.6). As a result, we know that $M_1/(G^\star_\alpha V_1) \to \infty$ if $\underline{\omega} r > \varrho_{\nu,\theta}(\beta)$. This shows that if $r > \underline{\omega}^{-1}\varrho_{\nu,\theta}(\beta)$, the power of $M_{\hat{J}_n}$ tends to 1.

Acknowledgments. We thank the joint Editors, an Associate Editor and two referees for their constructive comments and suggestions which have improved the presentation of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Two-sample and ANOVA tests for high dimensional means" (DOI: 10.1214/18-AOS1720SUPP; .pdf). The Supplementary Material provides the proofs of lemmas, propositions and Theorems 2, 3 and 5. It also includes extra simulation results and an empirical study.

REFERENCES

BAI, Z. and SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329. MR1399305

BICKEL, P. J. and LEVINA, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BICKEL, P. J. and LEVINA, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CAI, T. T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. MR3164870

CHEN, S. X., LI, J. and ZHONG, P.-S. (2019). Supplement to "Two-sample and ANOVA tests for high dimensional means." DOI:10.1214/18-AOS1720SUPP.

CHEN, S. X. and QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. MR2604697

DELAIGLE, A., HALL, P. and JIN, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 283–301. MR2815777

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

FAN, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Amer. Statist. Assoc.* **91** 674–688. MR1395735

FENG, L., ZOU, C., WANG, Z. and ZHU, L. (2015). Two-sample Behrens–Fisher problem for high-dimensional data. *Statist. Sinica* **25** 1297–1312. MR3409068

GREGORY, K. B., CARROLL, R. J., BALADANDAYUTHAPANI, V. and LAHIRI, S. N. (2015). A two-sample test for equality of means in high dimension. *J. Amer. Statist. Assoc.* **110** 837–849. MR3367268

GRÖCHENIG, K. and LEINERT, M. (2006). Symmetry and inverse-closedness of matrix algebras and functional calculus for infinite matrices. *Trans. Amer. Math. Soc.* **358** 2695–2711. MR2204052

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357

HE, J. and CHEN, S. X. (2016). Testing super-diagonal structure in high dimensional covariance matrices. *J. Econometrics* **194** 283–297. MR3536981

HOTELLING, H. (1931). The generalization of student's ratio. *Ann. Math. Stat.* **2** 360–378.

INGSTER, YU. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.* **6** 47–69. MR1456646

JAFFARD, S. (1990). Propriétés des matrices "bien localisées" près de leur diagonale et quelques applications. *Ann. Inst. H. Poincaré Anal. Non Linéaire* **7** 461–476. MR1138533

JI, P. and JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40** 73–103. MR3013180

PETROV, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables. Oxford Studies in Probability* **4**. The Clarendon Press, New York. MR1353441

QIU, Y. and CHEN, S. X. (2015). Bandwidth selection for high-dimensional covariance matrix estimation. *J. Amer. Statist. Assoc.* **110** 1160–1174. MR3420692

SRIVASTAVA, M. S., KATAYAMA, S. and KANO, Y. (2013). A two sample test in high dimensional data. *J. Multivariate Anal.* **114** 349–358. MR2993891

SUN, Q. (2005). Wiener's lemma for infinite matrices with polynomial off-diagonal decay. *C. R. Math. Acad. Sci. Paris* **340** 567–570. MR2138705

WAGAMAN, A. S. and LEVINA, E. (2009). Discovering sparse covariance structures with the isomap. *J. Comput. Graph. Statist.* **18** 551–572. MR2751641

WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215

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. *Ann. Statist.* **41** 2820–2851. MR3161449

S. X. CHEN
DEPARTMENT OF BUSINESS STATISTICS
AND ECONOMETRICS
GUANGHUA SCHOOL OF MANAGEMENT AND
CENTER FOR STATISTICAL SCIENCE
PEKING UNIVERSITY
BEIJING 100871
CHINA
E-MAIL: csx@gsm.pku.edu.cn

J. LI
DEPARTMENT OF MATHEMATICAL SCIENCES
KENT STATE UNIVERSITY
KENT, OHIO 44242
USA
E-MAIL: junli@math.kent.edu

P.-S. ZHONG
DEPARTMENT OF STATISTICS AND PROBABILITY
MICHIGAN STATE UNIVERSITY
EAST LANSING, MICHIGAN 48823
USA
E-MAIL: pszhong@stt.msu.edu