Two-Sample Tests for High Dimensional Means with Thresholding and Data Transformation

Chen, Song Xi and Li, Jun and Zhong, Pingshou

Peking University, Kent State Univeristy, Michgan State University

2014

Two-Sample Tests for High Dimensional Means with Thresholding and Data Transformation*

Song Xi Chen, Jun Li and Ping-Shou Zhong

Peking University and Iowa State University, Kent State University, and Michigan State University

Abstract

We consider testing for two-sample means of high dimensional populations by thresholding. Two tests are investigated, which are designed for better power performance when the two population mean vectors differ only in sparsely populated coordinates. The first test is constructed by carrying out thresholding to remove the non-signal bearing dimensions. The second test combines data transformation via the precision matrix with the thresholding. The benefits of the thresholding and the data transformations are showed by a reduced variance of the test thresholding statistics, the improved power and a wider detection region of the tests. Simulation experiments and an empirical study are performed to confirm the theoretical findings and to demonstrate the practical implementations.

KEYWORDS: Data Transformation; Large deviation; Large p small n; Sparse signals; Thresholding.

*Emails: csx@gsm.pku.edu.cn, junli@math.kent.edu, pszhong@stt.msu.edu

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" phenomenon. If the dimension p increases as the sample size n increases, many classical approaches originally designed for fixed dimension problems (Hotelling's test and the likelihood ratio tests for the covariances) may no longer be feasible. New methods are needed for the "large p, small n" setting.

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

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

Hotelling's T^2 test has been the classical test for the above hypotheses for fixed dimension p and is still applicable if $p \leq n_1 + n_2 - 2$. However, as shown in Bai and Saranadasa (1996), Hotelling's test suffers from a significant power loss when $p/(n_1 + n_2 - 2)$ approaches to 1 from below. When $p > n_1 + n_2 - 2$, the test is not applicable as the pooled sample covariance matrix, say $\boldsymbol{S}_n$, is no longer invertible.

There are proposals which modify Hotelling's T^2 statistic for high dimensional situations. Bai and Saranadasa (1996) proposed the following alteration

$$M_n = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^T(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - \mathrm{tr}(\boldsymbol{S}_n)/n, \tag{1.2}$$

by removing the inverse of the sample covariance matrix $\boldsymbol{S}_n^{-1}$ from the Hotelling's statistic, where $n = n_1 n_2/(n_1 + n_2)$. Chen and Qin (2010) considered a linear combi-

nation of U-statistics

$$T_n = \frac{1}{n_1(n_1-1)} \sum_{i \neq j}^{n_1} \boldsymbol{X}_{1i}^T \boldsymbol{X}_{1j} + \frac{1}{n_2(n_2-1)} \sum_{i \neq j}^{n_2} \boldsymbol{X}_{2i}^T \boldsymbol{X}_{2j} - \frac{2}{n_1 n_2} \sum_{i}^{n_1} \sum_{j}^{n_2} \boldsymbol{X}_{1i}^T \boldsymbol{X}_{2j}, \quad (1.3)$$

and showed that the corresponding test can operate under much relaxed regimes regarding the dimensionality and sample size constraint and without assuming $\Sigma_1 = \Sigma_2$. Srivastava, Katayama and Kano (2013) proposed using the diagonal matrix of the sample variance matrice to replace $\boldsymbol{S}_n$ under the normality. These three tests are basically all targeted on a weighted L_2 norms between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. In a development in another direction, Cai, Liu and Xia (2014) proposed a test based on the max-norm of marginal t-statistics. More importantly, they implemented a data transformation which is designed to increase the signal strength under sparsity as discovered early in Hall and Jin (2010) in their innovated higher criticism test for the one sample problem.

The L_2 norm based tests are known to be effective in detecting dense signals in the sense that the differences between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are populated over a large number of components. However, the tests will encounter power loss under the sparse signal settings where only a small portion of components of the two mean vectors are different. To improve the performance of these tests under the sparsity, we propose a thresholding test to remove the non-signal bearing dimensions. The idea of thresholding has been used in many applications, as demonstrated in Donoho and Johnstone (1994) for selecting significant wavelet coefficient and Fan (1996) for testing the mean of random vectors with IID normally distributed components. See also Ji and Jin (2012) for variable selection in high dimensional regression model. We find that the thresholding can reduce the variance of the Chen and Qin (2010) (CQ) test statistic, and hence increases the power of the test under sparsity for non-Gaussian data. We also confirm the effectiveness of the precision matrix transformation in increasing the signal strength of the CQ test. The transformation is facilitated by an estimator

of the precision matrix via the Cholesky decomposition with the banding approach (Bickel and Levina, 2008a, 2008b). It is shown that the test with the thresholding and the data transformation has a lower detection boundary than that without the data transformation, and can be lower than the detection boundary of an Oracle test without data transformation.

The rest of the paper is organized as follows. We analyze the thresholding test and its relative power performance to the CQ test and the Oracle test in Section 2. A multi-level thresholding test is proposed in Section 3 for detecting faint signals. Section 4 considers a data transformation with an estimated precision matrix. Simulation results are presented in Section 5. Section 6 reports an empirical study to select differentially expressed gene-sets for a human breast cancer data set. Section 7 concludes the paper with discussions. All technical details are relegated to the Appendix.

2. THRESHOLDING TEST

We first outline the CQ statistic before introducing the thresholding approach. The statistic (1.3) 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 $X_{ij}^{(k)}$ represents the k-th component of $\boldsymbol{X}_{ij}$. It can be readily shown that T_{nk} is unbiased to $(\mu_{1k} - \mu_{2k})^2$, which may be viewed as the amount of signal in the k-th dimension.

To facilitate simpler notations, we modify the test statistic T_n by standardizing each T_{nk} by $\sigma_{1,kk}/n_1 + \sigma_{2,kk}/n_2$, the variance of $\bar{X}_1^{(k)} - \bar{X}_2^{(k)}$, if both $\sigma_{1,kk}$ and $\sigma_{2,kk}$ are 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 variance estimates at the k-th dimension. This will make the CQ test invariant under the scale transformation; see Feng, Zou, Wang and Zhu (2013) for a related investigation. To expedite our discussion, we assume $\sigma^2_{i,kk}$ are known and equal to one without loss of generality. This leads to a modified version of the CQ statistic

$$\tilde{T}_n = n \sum_{k=1}^{p} T_{nk}, \tag{2.2}$$

where $n = n_1 n_2/(n_1 + n_2)$. Under the same setting, a modified version of the Bai and Saranadasa (BS) test statistic is

$$\tilde{M}_n = n \sum_{k=1}^{p} M_{nk} - p, \tag{2.3}$$

where $M_{nk} = (\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2$.

Let $\delta_k = \mu_{1k} - \mu_{2k}$ and $S_\beta = \{k : \delta_k \neq 0\}$ be the set of locations of the signals δ_k such that $|S_\beta| = p^{1-\beta}$ where $\beta \in (0, 1)$ is the sparsity parameter. Basically, the sparsity of the signal increases as β is closer to 1. Under the sparsity, an overwhelming number of T_{nk} carry no signals. However, including them increases the variance of the test statistic, and dilutes the signal to noise ratio of the test; and thus hampers the power of the test.

Let us now analyze the standardized CQ test under the sparsity. Define

$$\rho_{kl} = \mathrm{Cov}\left\{ \sqrt{n}(\bar{X}_1^{(k)} - \bar{X}_2^{(k)}), \sqrt{n}(\bar{X}_1^{(l)} - \bar{X}_2^{(l)}) \right\} = n(\sigma_{1,kl}/n_1 + \sigma_{2,kl}/n_2). \tag{2.4}$$

Similar to the derivation in Chen and Qin (2010), the variance of $\tilde{T}_n$ under H_0 is

$$\sigma^2_{\tilde{T}_n,0} = 2p + 2 \sum_{i \neq j} \rho_{ij}^2,$$

and that under H_1 is

$$\sigma^2_{\tilde{T}_n,1} = 2p + 2 \sum_{i \neq j} \rho_{ij}^2 + 4n \sum_{k,l \in S_\beta} \delta_k \, \delta_l \, \rho_{kl}. \tag{2.5}$$

It can be seen that $\sigma^2_{\tilde{T}_n,1} \geq \sigma^2_{\tilde{T}_n,0}$ since the last term of $\sigma^2_{\tilde{T}_n,1}$ is nonnegative due to $R = (\rho_{ij})_{p\times p}$ being non-negative definite.

Under a general multivariate model and some conditions on the covariance matrices, the asymptotic normality of $\tilde{T}_n$ can be established (Chen and Qin, 2010):

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

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

Let $\bar{\delta}^2 = \sum_{k\in S_\beta} n\,\delta_k^2/p^{1-\beta}$ represent the average standardized signal. The power of the test is

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

where $\Phi(\cdot)$ is the distribution function of $\mathrm{N}(0,1)$. Since $\sigma^2_{\tilde{T}_n,1} \geq \sigma^2_{\tilde{T}_n,0}$, the first term within $\Phi(\cdot)$ is bounded. Then, the power of the test is largely determined by the second term

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

which is called the signal to noise ratio of the test since the numerator is the average signal strength and the denominator is the standard deviation of the test statistic under H_1. An inspection 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 $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$,

$$\mathrm{SNR}_{\tilde{T}_n} = \frac{p^{1-\beta}\bar{\delta}^2}{\sqrt{2p + 4p^{1-\beta}\bar{\delta}^2}}.$$

Hence, if the sparsity $\beta > 1/2$ and the average signal $\bar{\delta} = o(p^{\beta/2-1/4})$, $\mathrm{SNR}_{\tilde{T}_n} = o(1)$. Then, the test has little power beyond the significant level. A reason for the power

loss is that the variance of $\tilde{T}_n$ is much inflated by including those non-signal bearing T_{nk}.

To put the above analysis in prospective, we consider an Oracle test which has the knowledge of the **possible** signal bearing set S_β (with slight abuse of notation), which is much smaller than the entire set of dimensions. The Oracle is only a semi-Oracle as he does not know the exact dimensions of the signals other than that they are within S_β.

The Oracle test statistic is

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

Similar to the derivation of (2.5), the variance of O_n under H_0 is

$$\sigma^2_{O_n,0} = 2p^{1-\beta} + 2 \sum_{i \neq j \in S_\beta} \rho^2_{ij},$$

and that under H_1 is

$$\sigma^2_{O_n,1} = 2p^{1-\beta} + 2 \sum_{i \neq j \in S_\beta} \rho^2_{ij} + 4n \sum_{k,l \in S_\beta} \delta_k \, \delta_l \, \rho_{kl}. \tag{2.8}$$

Comparing $\sigma^2_{O_n,1}$ with $\sigma^2_{\tilde{T}_n,1}$ in (2.5), we see that the first term of $\sigma^2_{O_n,1}$ is much smaller than that of $\sigma^2_{\tilde{T}_n,1}$. It may be shown that under the same conditions that establish the asymptotic normality of $\tilde{T}_n$,

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

which leads to the Oracle test that rejects H_0 if $O_n / \hat{\sigma}_{O_n,0} > z_\alpha$ where $\hat{\sigma}_{O_n,0}$ is a ratio consistent estimator of $\sigma_{O_n,0}$.

The asymptotic normality implies that the power of the Oracle test is

$$\beta_{O_n}(||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||) = \Phi\bigg(- \frac{\sigma_{O_n,0}}{\sigma_{O_n,1}} z_\alpha + \frac{p^{1-\beta}\bar{\delta}^2}{\sigma_{O_n,1}} \bigg).$$

It is largely determined by

$$\mathrm{SNR}_{O_n} =: \frac{p^{1-\beta}\bar{\delta}^2}{\sqrt{2p^{1-\beta} + 2\sum_{i\neq j\in S_\beta}\rho_{ij}^2 + 4n\sum_{k,l\in S_\beta}\delta_k\,\delta_l\,\rho_{kl}}}, \tag{2.9}$$

which is much larger than $\mathrm{SNR}_{\tilde{T}_n}$ since $\sigma^2_{O_n,1} \ll \sigma^2_{\tilde{T}_n,1}$. If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{I}_p$,

$$\mathrm{SNR}_{O_n} = \frac{p^{1-\beta}\bar{\delta}^2}{\sqrt{2p^{1-\beta} + 4p^{1-\beta}\bar{\delta}^2}} = \frac{p^{\frac{1-\beta}{2}}\bar{\delta}^2}{\sqrt{2 + 4\bar{\delta}^2}}, \tag{2.10}$$

that tends to infinity for $\beta > 1/2$ as long as $\bar{\delta}$ is a large order of $p^{\beta/4-1/4}$, which is much smaller than $p^{\beta/2-1/4}$ for the CQ test, indicating the test is able to detect much fainter signal.

The reason that the Oracle test has better power is that all the excluded dimensions are definitely non-signal bearing and those included have much smaller dimensions. In reality, the locations of those non-signal bearing dimensions are unknown. However, thresholding can be carried out to exclude those non-signal 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 non-signal bearing T_{nk} while maintaining those with signals. The thresholding test statistic is

$$L_1(s) = \sum_{k=1}^{p} n\,T_{nk} I\Big\{n\,T_{nk} + 1 > \lambda_n(s)\Big\}, \tag{2.11}$$

where $I(\cdot)$ is the indicator function.

We can also carry out the thresholding on BS test statistic (2.3), which leads to

$$L_2(s) = \sum_{k=1}^{p}\Big\{n(\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2 - 1\Big\} I\Big\{n(\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2 > \lambda_n(s)\Big\}. \tag{2.12}$$

As we will show later, both $L_1(s)$ and $L_2(s)$ have very similar properties. Therefore, we choose $L_n(s)$ to refer to either $L_1(s)$ or $L_2(s)$.

Before we show that the thresholding can reduce the variance contributed from those non-signal bearing dimensions without harming the signals, we introduce the

notion of α-mixing to quantify the dependence among the components of the random vector $\boldsymbol{X} = (X^{(1)}, \cdots, X^{(p)})^T$.

For any integers $a < b$, define $\mathcal{F}_{\boldsymbol{X},(a,b)}$ to be the σ-algebra generated by $\{X^{(m)} : m \in (a,b)\}$ and define the α-mixing coefficient

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

The following conditions are assumed in our analysis.

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

(C2): Let $\boldsymbol{X}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{W}_{ij}$. 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): The sequence of random variables $\{X_{ij}^{(l)}\}_{l=1}^p$ is α-mixing such that $\alpha_{\boldsymbol{X}}(k) \leq C\alpha^k$ for some $\alpha \in (0,1)$ and a positive constant C, and ρ_{kl} defined in (2.4) are summable such that $\sum_{l=1}^p |\rho_{kl}| < \infty$ for any $k \in \{1, \cdots, p\}$.

Condition (C1) specifies the growth rate of dimension p relative to n under which the large deviation results can be applied to derive the means and variances of the test statistics. Condition (C2) assumes that $(X_{ij}^{(k)}, X_{ij}^{(l)})$ has a bivariate sub-Gaussian distribution, which is more general than the Gaussian distribution. Condition (C3) prescribes weak dependence among the column components of the random vector, which is commonly assumed in time series analysis.

Derivations given in Appendix leading to (A.1) and (A.2) show that the mean of the thresholding test statistic $L_n(s)$ is

$$\begin{aligned} \mu_{L_n(s)} \;=\; & \Bigg(\frac{2}{\sqrt{2\pi}} (2s\log p)^{\frac{1}{2}} p^{1-s} + \sum_{k \in S_\beta} \{ n\,\delta_k^2 I(n\,\delta_k^2 > 2s\log p) \\ & + (2s\log p) \bar{\Phi}(\eta_k^-) I(n\,\delta_k^2 < 2s\log p) \} \Bigg) \{1 + o(1)\}, \end{aligned} \tag{2.13}$$

and the variance is

$$\begin{aligned}\sigma^2_{L_n(s)} &= \Big(\frac{2}{\sqrt{2\pi}}\{(2s\log p)^{\frac{3}{2}}+(2s\log p)^{\frac{1}{2}}\}p^{1-s}+\sum_{k,l\in S_\beta}(4n\delta_k\delta_l\rho_{kl}+2\rho^2_{kl})\\ &\times\ I(n\delta_k^2>2s\log p)I(n\,\delta_l^2>2s\log p)+\sum_{k\in S_\beta}(2s\log p)^2\bar{\Phi}(\eta_k^-)\\ &\times\ I(n\,\delta_k^2<2s\log p)\Big)\{1+o(1)\},\end{aligned}\tag{2.14}$$

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

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

$$\sigma^{-1}_{L_n(s)}\big\{L_n(s)-\mu_{L_n(s)}\big\}\xrightarrow{d}\mathrm{N}(0,1).$$

Let $\mu_{L_n(s),0}$ and $\sigma_{L_n(s),0}$ be the mean and variance under H_0 which can be obtained by ignoring the summation terms in (2.13) and (2.14). Then, Theorem 1 implies an asymptotic α level test that rejects H_0 if

$$L_n(s)>z_\alpha\hat{\sigma}_{L_n(s),0}+\hat{\mu}_{L_n(s),0},\tag{2.15}$$

where $\hat{\mu}_{L_n(s),0}$ and $\hat{\sigma}_{L_n(s),0}$ are consistent estimators of $\mu_{L_n(s),0}$ and $\sigma_{L_n(s),0}$ satisfying

$$\mu_{L_n(s),0}-\hat{\mu}_{L_n(s),0}=o\{\sigma_{L_n(s),0}\}\quad\text{and}\quad\hat{\sigma}_{L_n(s),0}/\sigma_{L_n(s),0}\xrightarrow{p}1.\tag{2.16}$$

If all the signals δ_k^2 are strong such that $n\,\delta_k^2>2\log p$, choosing $s=1^-$ such that $(1-s)\log(p)=o(1)$ leads to

$$\mu_{L_n(s)} = \Big\{\frac{2}{\sqrt{2\pi}}(2\log p)^{\frac{1}{2}}+\sum_{k\in S_\beta}n\,\delta_k^2\Big\}\{1+o(1)\},$$

and

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

Except for a slowly varying logarithm function of p, $\sigma^2_{L_n(s)}$ has the same leading order variance of the Oracle statistic

$$\sigma^2_{O_n,1}=\sum_{k,l\in S_\beta}\Big(4n\,\delta_k\,\delta_l\,\rho_{kl}+2\rho^2_{kl}\Big),$$

indicating the effectiveness of the thresholding under the strong signal situation. With the same choice of s for strong signals case, $\mu_{L_n(s),0}$ and $\sigma^2_{L_n(s),0}$ can be respectively estimated by

$$\hat{\mu}_{L_n,0} = \frac{2}{\sqrt{2\pi}}(2\log p)^{\frac{1}{2}} \quad \text{and} \quad \hat{\sigma}^2_{L_n,0} = \frac{2}{\sqrt{2\pi}}\{(2\log p)^{\frac{3}{2}} + (2\log p)^{\frac{1}{2}}\}.$$

It can be shown that (2.16) is satisfied under (C1) and thus can be employed in the formulation of a test procedure.

The asymptotic power of the thresholding test (2.15) is

$$\beta_{L_n}(||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||) = \Phi\left(-\frac{z_\alpha \sigma_{L_n(s),0}}{\sigma_{L_n(s),1}} + \frac{\mu_{L_n(s),1} - \mu_{L_n(s),0}}{\sigma_{L_n(s),1}}\right),$$

which, similar to the CQ and the Oracle tests, is largely determined by

$$\begin{aligned} \mathrm{SNR}_{L_n} &=: \frac{\mu_{L_n(s),1} - \mu_{L_n(s),0}}{\sigma_{L_n(s),1}} \\ &= \frac{p^{1-\beta}\bar{\delta}^2}{\sqrt{2L_p + 2p^{1-\beta} + 2\sum_{k\neq l\in S_\beta}\rho_{kl}^2 + 4n\sum_{k,l\in S_\beta}\delta_k\,\delta_l\,\rho_{kl}}}, \end{aligned} \tag{2.17}$$

which is much larger than that of the CQ test in (2.6) and differs from that of the Oracle test given in (2.9) only by a slowly varying multi-log$p$ function L_p. This echoes that established in Fan (1996) for Gaussian data with no dependence among the column components of the data.

3. MULTI-LEVEL THRESHOLDING

It is shown in Section 2 that if all the signals are strong such that $n\delta_k > 2\log p$, a single thresholding with $s = 1^-$ improves significantly the power of the test and attains nearly the power of the Oracle test. However, if some signals are weak such that $n\,\delta_k^2 = 2r\log p$ with $r < 1$ for some $k \in S_\beta$, the thresholding has to be administrated at smaller levels $2s\log p$ for $s \in (0, 1)$. In this case, the single-level thresholding does not work well. One approach that provides a solution to such situation is the

higher criticism test (Donoho and Jin, 2004) which effectively combines many levels of thresholding together to formulate a higher criticism (HC) criterion. Zhong, Chen and Xu (2013) proposed a more powerful test procedure than the HC test under sparsity and data dependence. Both Donoho and Jin (2004)'s HC test and the test proposed in Zhong et al. (2013) are for one sample, and both did not provide much details on the power performance.

The multi-level thresholding statistic is

$$M_{L_n} = \max_{s\in(0,1-\eta)} \frac{L_n(s) - \hat{\mu}_{L_n(s),0}}{\hat{\sigma}_{L_n(s),0}}. \tag{3.1}$$

Maximizing over the thresholding statistics at multiple levels allows faint and unknown signals to be captured. Since both $\hat{\mu}_{L_n(s),0}$ and $\hat{\sigma}_{L_n(s),0}$ are monotonically decreasing and $L_n(s)$ contains indicator functions, provided (2.16) is satisfied, it can be shown that the maximization in (3.1) is attained over $\mathcal{S}_n = \{s_k : s_k = n(\bar{X}_1^{(k)} - \bar{X}_1^{(k)})^2/(2\log p), \text{for } k = 1, \cdots, p\} \cap (0, 1-\eta)$ so that

$$M_{L_n} = \max_{s\in\mathcal{S}_n} \frac{L_n(s) - \hat{\mu}_{L_n(s),0}}{\hat{\sigma}_{L_n(s),0}}. \tag{3.2}$$

The following theorem shows that M_{L_n} is asymptotically Gumbel distributed.

Theorem 2. Assume Conditions **(C1)-(C3)** and condition (2.16) is satisfied. Then under H_0,

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

where functions $a(y) = (2\log y)^{\frac{1}{2}}$ and $b(y,\eta) = 2\log y + 2^{-1}\log\log y - 2^{-1}\log\{\frac{4\pi}{(1-\eta)^2}\}$.

The theorem implies that a two-sample multi-level thresholding test of asymptotic α level rejects H_0 if

$$M_{L_n} \ge G_\alpha = \{q_\alpha + b(\log p, \eta)\}/a(\log p), \tag{3.3}$$

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

Define

$$\varrho(\beta) = \begin{cases} \beta - \frac{1}{2}, & \frac{1}{2} \leq \beta \leq \frac{3}{4}; \\ (1 - \sqrt{1-\beta})^2, & \frac{3}{4} < \beta < 1. \end{cases} \tag{3.4}$$

Ingster (1997) shows that $r = \varrho(\beta)$ is the optimal detection boundary for uncorrelated Gaussian data in the sense that when (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 $\boldsymbol{X}_i$ are IID $\mathrm{N}(\mu, I_p)$ data. Zhong et al. (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) further above the detection boundary.

Theorem 3. Assume Conditions **(C1)-(C3)** and $\hat{\mu}_{L_n(s),0}$ and $\hat{\sigma}_{L_n(s),0}$ satisfy (2.16). If $r > \varrho(\beta)$, the sum of type I and II errors of the multi-level thresholding test converges to zero when $\alpha = \bar{\Phi}\{(\log p)^\epsilon\} \to 0$ for an arbitrarily small $\epsilon > 0$ as $n \to \infty$. If $r < \varrho(\beta)$, the sum of type I and II errors of the multi-level thresholding test converges to 1 as $\alpha \to 0$ and $n \to \infty$.

Theorem 3 implies that the two-sample multi-level thresholding test also attains $r = \varrho(\beta)$ as the detection boundary in the current two-sample test setting of non-parametric distributional assumption. This means that the test can asymptotically distinguish H_1 from H_0 for any (r, β) above the detection boundary. If the mean and variance estimators $\hat{\mu}_{L_n(s),0}$ and $\hat{\sigma}_{L_n(s),0}$ do not satisfy (2.16), the detection boundary will be higher just like what will happen in Theorem 6 given in Section 4 when we consider testing via data transformation with estimated precision matrix.

4. TEST WITH DATA TRANSFORMATION

We consider in this section another way for power improvement, which involves enhancing the signal strength by data rotation, inspired by the works of Hall and Jin (2010) and Cai, Liu and Xia (2014). We will show in this section that the signal enhancement can be achieved by transforming the data via an estimate of the inverse of a mixture of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. Transforming data to achieve better power has been considered in Hall and Jin (2010) in their innovated higher criticism test under dependence and Cai, Liu and Xia (2014) in their max-norm based test. The transformation used in Hall and Jin (2010) was via a banded Cholesky factor, and that adopted in Cai, Liu and Xia (2014) was via the CLIME estimator of the inverse of the covariance matrix (the precision matrix) proposed in Cai, Liu and Luo (2011).

Consider a bandable covariance matrix class

$$\begin{aligned} V(\epsilon_0, C, \alpha) \quad = \quad & \Big\{ \mathbf{\Sigma} : 0 < \epsilon_0 \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq \epsilon_0^{-1}, \alpha > 0, \\ & |\sigma_{ij}| \leq C(1 + |i-j|)^{-(\alpha+1)} \quad \text{for all} \quad i, j : |i-j| \geq 1 \Big\}. \end{aligned}$$

This class of matrices satisfies both the banding and thresholding conditions of Bickel and Levina (2008b). Hall and Jin (2010) also considered this class when they proposed the innovated higher criticism test under dependence.

(C4): Both $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ belong to the matrix class $V(\epsilon_0, C, \alpha)$.

Although both **(C3)** and **(C4)** assume the weak dependence among the column components of the random vector X_{ij}, imposing **(C4)** ensures that the banding estimation of the covarianvce matrix which makes the transformed data are still weakly dependent. To appreciate this, let $\mathbf{\Omega} = \{(1-\kappa)\mathbf{\Sigma}_1 + \kappa\mathbf{\Sigma}_2\}^{-1} = (\omega_{ij})_{p\times p}$. 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

$$\mathbf{\Omega}(\tau) = \Big\{ \omega_{ij} \mathrm{I}(|i-j| \leq \tau) \Big\}_{p\times p},$$

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 the 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 their off-diagonal entries decaying to zero at polynomial rates, $\mathbf{\Omega}$ has the same rate of decay as well (Jaffard, 1990; Sun, 2005; Gröchenig, and Leinert, 2006), which ensures that the transformed data are still weakly dependent.

The two transformed samples are

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

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 $\bar{Z}_i^{(k)}(\tau) = n_i^{-1}\sum_{j=1}^{n_j} Z_{ij}^{(k)}(\tau)$ for $i = 1, 2$. Lemmas 5 and 7 in Appendix show that there exists a constant $C > 1$ such that

$$\varpi_{kk}(\tau) = \omega_{kk} + O(\tau^{-C}) \quad \text{and} \quad \omega_{kk} > 1. \tag{4.1}$$

We have two ways to construct the transformed thresholding test statistic by replacing $\boldsymbol{X}_{ij}$ with $\boldsymbol{Z}_{ij}(\tau)$ in either (2.11) or (2.12). Alhough both have similar properties, the latter which has the form

$$J_n(s,\tau) = \sum_{k=1}^{p}\left\{\frac{n(\bar{Z}_1^{(k)}(\tau) - \bar{Z}_2^{(k)}(\tau))^2}{\varpi_{kk}(\tau)} - 1\right\} I\left\{\frac{n(\bar{Z}_1^{(k)}(\tau) - \bar{Z}_2^{(k)}(\tau))^2}{\varpi_{kk}(\tau)} > \lambda_n(s)\right\} \tag{4.2}$$

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

Let $\delta_{\mathbf{\Omega}(\tau)} = (\delta_{\mathbf{\Omega}(\tau),1}, \cdots, \delta_{\mathbf{\Omega}(\tau),p})^T$ where

$$\delta_{\mathbf{\Omega}(\tau),k} = \sum_l \mathbf{\Omega}_{kl}(\tau)\delta_l = \sum_{l \in S_\beta} \omega_{kl}\delta_l \mathrm{I}(|k-l| \le \tau) \tag{4.3}$$

denotes the difference between the transformed means in the k-th dimension. Similar

to (2.13) and (2.14), the mean and variance of the transformed statistic $J_n(s,\tau)$ are

$$\mu_{J_n(s,\tau)} = \Bigg(\frac{2}{\sqrt{2\pi}}(2s\log p)^{\frac{1}{2}}p^{1-s} + \sum_{k\in S_{\mathbf{\Omega}(\tau),\beta}}\{n\frac{\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)}I(n\frac{\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)} > 2s\log p) + (2s\log p)\bar{\Phi}(\eta^-_{\mathbf{\Omega}(\tau)k})I(n\frac{\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)} < 2s\log p)\}\Bigg)\{1+o(1)\}, \tag{4.4}$$

and

$$\sigma^2_{J_n(s,\tau)} = \Bigg(\frac{2}{\sqrt{2\pi}}\{(2s\log p)^{\frac{3}{2}} + (2s\log p)^{\frac{1}{2}}\}p^{1-s} + \sum_{k,l\in S_{\mathbf{\Omega}(\tau),\beta}}(4n\frac{\delta_{\mathbf{\Omega}(\tau),k}}{\varpi^{1/2}_{kk}(\tau)}\frac{\delta_{\varpi(\tau),l}}{\varpi^{1/2}_{ll}(\tau)}\rho_{\mathbf{\Omega},kl} + 2\rho^2_{\mathbf{\Omega},kl})I(n\frac{\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)} > 2s\log p)I(n\frac{\delta^2_{\mathbf{\Omega}(\tau),l}}{\varpi_{ll}(\tau)} > 2s\log p) + \sum_{k\in S_{\mathbf{\Omega}(\tau),\beta}}(2s\log p)^2\bar{\Phi}(\eta^-_{\mathbf{\Omega}(\tau)k})I(n\frac{\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)} < 2s\log p)\Bigg)\{1+o(1)\}, \tag{4.5}$$

where $S_{\mathbf{\Omega}(\tau),\beta} = \{k : \delta_{\mathbf{\Omega}(\tau),k} \neq 0\}$ is the set of locations of the non-zero 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

$$\rho_{\mathbf{\Omega},kl} = \mathrm{Cov}\Bigg\{\frac{\sqrt{n}(\bar{Z}^{(k)}_1(\tau) - \bar{Z}^{(k)}_2(\tau))}{\sqrt{\varpi_{kk}(\tau)}}, \frac{\sqrt{n}(\bar{Z}^{(l)}_1(\tau) - \bar{Z}^{(l)}_2(\tau))}{\sqrt{\varpi_{ll}(\tau)}}\Bigg\}.$$

In practice, the precision matrix $\mathbf{\Omega}$ is unknown and needs to be estimated. We consider the Cholesky decomposition and the banding approach similar to that in Bickel and Levina (2008a). Define $\boldsymbol{Y}_{kl} = \boldsymbol{X}_{1k} - \sqrt{\frac{\kappa}{1-\kappa}}\boldsymbol{X}_{2l}$ for $k = 1,\cdots,n_1$ and $l = 1,\cdots,n_2$, where $\kappa = \lim_{n\to\infty} n_1/(n_1+n_2)$. Then $\mathrm{Var}(Y_{kl}) = \mathbf{\Sigma}_w \equiv \mathbf{\Sigma}_1 + \frac{\kappa}{1-\kappa}\mathbf{\Sigma}_2$. Thus, to estimate $\mathbf{\Omega} = (1-\kappa)^{-1}\mathbf{\Sigma}_w^{-1}$, we only need to estimate $\mathbf{\Sigma}_w^{-1}$.

Let $\boldsymbol{Y}$ be an IID copy of $\boldsymbol{Y}_{kl}$ for any fixed k and l such that $\boldsymbol{Y} = (Y^{(1)},\cdots,Y^{(p)})^T$. For $j = 1,\cdots,p$, define $\hat{Y}^{(j)} = \mathbf{a}_j^T\mathbf{W}^{(j)}$ where $\mathbf{a}_j = \{\mathrm{Var}(\mathbf{W}^{(j)})\}^{-1}\mathrm{Cov}(\hat{Y}^{(j)},\mathbf{W}^{(j)})$ and $\mathbf{W}^{(j)} = (Y^{(1)},\cdots,Y^{(j-1)})^T$. Let $\epsilon_j = Y^{(j)} - \hat{Y}^{(j)}$ and $d_j^2 = \mathrm{Var}(\epsilon_j)$, and $\boldsymbol{A}$ be the lower triangular matrix with the j-th row being $(\mathbf{a}_j^T, \mathbf{0}_{p-j+1})$ and $\boldsymbol{D} = \mathrm{diag}(d_1^2,\cdots,d_p^2)$ where $\mathbf{0}_s$ means a vector of 0 with length s. Then, the population version of Cholesky decomposition is $\mathbf{\Sigma}_w^{-1} = (I-\boldsymbol{A})^T\boldsymbol{D}^{-1}(I-\boldsymbol{A})$.

The banded estimators for $\boldsymbol{A}$ and $\boldsymbol{D}$ (Bickel and Levina, 2008a) can be used in the case of $p > \min\{n_1,n_2\}$. Specifically, let $\boldsymbol{Y}_{n,kl} = \boldsymbol{X}_{1k} - \sqrt{\frac{n_1}{n_2}}\boldsymbol{X}_{2l} := (Y^{(1)}_{n,kl},\cdots,Y^{(p)}_{n,kl})^T$.

Given a τ, regress $Y_{n,kl}^{(j)}$ on $\mathbf{Y}_{n,kl,-\tau}^{(j)} = (Y_{n,kl}^{(j-\tau)}, \cdots, Y_{n,kl}^{(j-1)})^T$ to obtain the least square estimate of $\mathbf{a}_{j,\tau} = (a_{j-\tau}, \cdots, a_{j-1})^T$:

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

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 j-th row of a lower triangular matrix $\hat{A}_\tau$ and $\hat{D}_\tau = \text{diag}(d_{1,\tau}^2, \cdots, d_{p,\tau}^2)$ where $d_{j,\tau}^2 = \frac{1}{n_1 n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2}(Y_{n,kl}^{(j)} - \hat{\mathbf{a}}_{j,\tau}^T \mathbf{Y}_{n,kl,-\tau}^{(j)})^2$. Thus, the estimator of $\mathbf{\Sigma}_w^{-1}$ is

$$\widehat{\mathbf{\Sigma}_w^{-1}} = (I - \hat{A}_\tau)^T \hat{D}_\tau^{-1} (I - \hat{A}_\tau), \tag{4.6}$$

which results in $\hat{\mathbf{\Omega}}_\tau = \{1 - n_1/(n_1+n_2)\}^{-1}\widehat{\mathbf{\Sigma}_w^{-1}}$.

The consistency of $\hat{\mathbf{\Omega}}_\tau$ to $\mathbf{\Omega}$ basically follows the proof of Theorem 3 in Bickel and Levina (2008a) with a main difference that replaces the exponential tail inequality for a sample mean in Lemma A.3 of their paper to an exponential inequality of a two-sample U-statistics. Moreover, if the banding parameter $\tau \asymp (n^{-1}\log p)^{-\frac{1}{2(\alpha+1)}}$ and $n^{-1}\log p = o(1)$, it can be shown that

$$||\hat{\mathbf{\Omega}}_\tau - \mathbf{\Omega}|| = O_p\left\{(\log p/n)^{\frac{\alpha}{2(\alpha+1)}}\right\},$$

where $\|\cdot\|$ is the spectral norm.

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

$$\hat{J}_n(s,\tau) = \sum_{k=1}^{p}\left\{\frac{n(\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)})^2}{\hat{\omega}_{kk}} - 1\right\} I\left\{\frac{n(\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)})^2}{\hat{\omega}_{kk}} > \lambda_n(s)\right\}. \tag{4.7}$$

To consistently estimate $\mathbf{\Omega}$, we require that $\tau \asymp (n^{-1}\log p)^{-\frac{1}{2(\alpha+1)}}$. This requirement leads to a modification on the range of the thresholding level s as shown in the next theorem.

Theorem 4. Assume Conditions **(C1)-(C4)**. If $p = n^{1/\theta}$ for $0 < \theta < 1$ and $\tau \asymp (n^{-1}\log p)^{-\frac{1}{2(\alpha+1)}}$, then for any $s \in (1-\theta, 1)$,

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

The restriction on the thresholding level s in Theorem 4 is to ensure the estimation error of $\hat{\mathbf{\Omega}}_\tau$ is negligible. Similar restriction is provisioned in Delaigle et al. (2011) and Zhong et al. (2013). Note that if θ is arbitrarily close to 0, p will grow exponentially fast with n.

A single-level thresholding test based on the transformed data rejects H_0 if

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

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

From Theorem 4, the asymptotic power of the transformed thresholding test is

$$\beta_{\hat{J}_n(s,\tau)}(||\mu_1 - \mu_2||) = \Phi\left(-\frac{z_\alpha \sigma_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),1}} + \frac{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),1}} \right),$$

which is determined by

$$\mathrm{SNR}_{\hat{J}_n(s,\tau)} =: \frac{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),1}}.$$

Therefore, to compare with the thresholding test without transformation, it is equivalent to compare $\mathrm{SNR}_{\hat{J}_n(s,\tau)}$ to SNR_{L_n}. To this end, we assume the following regarding the distribution of the non-zero δ_k in S_β.

(C5): The elements of S_β are randomly distributed among $\{1, 2, \cdots, p\}$.

Under Conditions (C1)-(C5), Lemma 8 in the Appendix shows that with probability approaching to 1,

$$\mathrm{SNR}_{\hat{J}_n(s,\tau)} \geq \mathrm{SNR}_{L_n}, \tag{4.8}$$

which holds for both strong and weak signals. Hence, the transformed thresholding test is more powerful regardless of the underlying signal strength for randomly allocated signals.

Similar to M_{L_n} defined in (3.2) for weaker signals, a multi-level thresholding statistic for transformed data is

$$M_{\hat{J}_n} = \max_{s \in T_n} \frac{\hat{J}_n(s,\tau) - \hat{\mu}_{J_n(s,\tau),0}}{\hat{\sigma}_{J_n(s,\tau),0}}, \tag{4.9}$$

where $T_n = \{s_k : s_k = n(\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)})^2/(2\log p\hat{\omega}_{kk}) \text{ for } k = 1, \cdots, p\} \cap (1-\theta, 1-\eta^\star)$ for arbitrarily small $\eta^\star$. The asymptotic distribution of $M_{\hat{J}_n}$ is given in the following Theorem.

Theorem 5. Assume Conditions **(C1)-(C4)**, $p = n^{1/\theta}$ for $0 < \theta < 1$ and $\tau \asymp (n^{-1}\log p)^{-\frac{1}{2(\alpha+1)}}$. Then under H_0,

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

where functions $a(\cdot)$ and $b(\cdot,\cdot)$ are defined in Theorem 2.

The theorem implies an asymptotically α level test that rejects H_0 if

$$M_{\hat{J}_n} \ge \{q_\alpha + b(\log p, \theta - \eta^\star)\}/a(\log p). \tag{4.10}$$

It is expected that the above test as well as the thresholding test without the data transformation will encounter size distortion. The size distortion is caused by the generally slow convergence to the extreme value distribution. It may be also due to the second order effects of the data dependence. Our analyses have shown that the data dependence has no leading order effect on the asymptotic variance of the thresholding test statistics. However, a closer examination on the variance shows that the second order term is not that smaller than the leading order variance. This can create a discrepancy when approximating the distribution of the multi-level thresholding statistics by the Gumbel distribution. To remedy the problem, we proposed a

parametric bootstrap approximation to the null distribution of the multi-level thresholding statistics with and without the data transformation. We first estimate Σ_i by $\hat{\Sigma}_i$ for $i = 1, 2$ through the Cholesky decomposition which can be obtained by inverting the one-sample version of (4.6) based on the samples $\{\boldsymbol{X}_{1j}\}_{j=1}^{n_1}$ and $\{\boldsymbol{X}_{2j}\}_{j=1}^{n_2}$, respectively. Bootstrap resamples are generated repeatedly from $\mathrm{N}(0, \hat{\Sigma}_i)$ which allows us to obtain the bootstrap copies of the statistic M_{L_n} defined in (3.2), namely $M_{L_n}^{*,(1)}(s), \cdots, M_{L_n}^{*,(B)}(s)$, after B repetitions. We use $\{M_{L_n}^{*,(b)}\}_{b=1}^{B}$ to obtain the empirical null distribution of the multi-level thresholding statistic. The same parametric bootstrapping method can be also applied to the transformed multi-level thresholding statistic.

We have shown that the transformed thresholding test has a better power performance than the thresholding test without the transformation. We are to show that the transformed multi-level thresholding test has lower detection boundary than the multi-level thresholding test without transformation.

To define the detection boundary of the transformed multi-level thresholding test, let

$$\underline{\omega} = \underline{\lim}_{p\to\infty}\left(\min_{1\le k\le p}\omega_{kk}\right) \quad \text{and} \quad \bar{\omega} = \overline{\lim}_{p\to\infty}\left(\max_{1\le k\le p}\omega_{kk}\right).$$

Results in (4.1) imply that $\underline{\omega}$ and $\bar{\omega} \ge 1$. Define

$$\varrho_\theta(\beta) = \begin{cases} (\sqrt{1-\theta} - \sqrt{1-\beta-\frac{\theta}{2}})^2, & \frac{1}{2} \le \beta \le \frac{3-\theta}{4}; \\ \beta - \frac{1}{2}, & \frac{3-\theta}{4} \le \beta \le \frac{3}{4}; \\ (1-\sqrt{1-\beta})^2, & \frac{3}{4} < \beta < 1. \end{cases} \tag{4.11}$$

Theorem 6. Assume Conditions **(C1)**-**(C5)**.

(a) When $\boldsymbol{\Omega}$ is known, if $r < \bar{\omega}^{-1} \cdot \varrho(\beta)$, 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 $r > \underline{\omega}^{-1} \cdot \varrho(\beta)$, the sum of type I and II errors of the transformed multi-level

thresholding test converges to zero when $\alpha = \bar{\Phi}\{(\log p)^{\epsilon}\} \to 0$ for an arbitrarily small $\epsilon > 0$ as $n \to \infty$.

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

Hall and Jin (2010) has shown that utilizing the dependence can lower the detection boundary $r = \varrho(\beta)$ for Gaussian data with known covariance matrix. We demonstrate in Theorem 6 that the detection boundary can be lowered respectively for the transformed multi-level thresholding test with $\mathbf{\Omega}$ being known or unknown for sub-Gaussian data with estimated precision matrix. The theorem shows that there is a cost associated with using the estimated precision matrix in terms of a higher detetction boundary and more restriction on the p and n relationship.

5. SIMULATION STUDY

In this section, the simulation was designed to confirm the performance of the two multi-level thresholding tests defined in (3.2) and (4.9) without and with transformation. We also experimented the test of Chen and Qin (2010) given in (1.3), the Oracle test in (2.7), and two tests proposed by Cai, Liu and Xia (2014). The latter tests are based on the max-norm statistics

$$G(I) = \max_{1 \le k \le p} n(\bar{X}_1^{(k)} - \bar{X}_2^{(k)})^2 \quad \text{and} \quad G(\hat{\mathbf{\Omega}}) = \max_{1 \le k \le p} \frac{n(\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)})^2}{\hat{\omega}_{kk}},$$

without and with transformation, where $\hat{\omega}_{kk}$ were estimates of the diagonal elements of $\mathbf{\Omega}$. Cai, Liu and Xia (2014) showed that $G(I)$ and $G(\hat{\mathbf{\Omega}})$ converge to the type I extreme value distribution with cumulative distribution function $\exp(-\frac{1}{\sqrt{\pi}}\exp(-x/2))$, which

was used to formulate the test procedures based on the two max-norm statistics. Cai, Liu and Xia (2014) employed the CLIME estimator based on a constrained l_1 minimization estimator of Cai, Liu and Luo (2011) to estimate $\mathbf{\Omega}$. Since we use the Cholesky decomposition with banding to estimate $\mathbf{\Omega}$ in the transformed thresholding test, we used the estimated $\hat{\omega}_{kk}$ from the approach in the formulation of the max-norm statistics.

In the simulation experiments, the two random samples $\{\boldsymbol{X}_{1j}\}_{j=1}^{n_1}$ and $\{\boldsymbol{X}_{2j}\}_{j=1}^{n_2}$ were generated according to the following multivariate model

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

where the innovations $\boldsymbol{Z}_{ij}$ are IID p-dimensional random vectors with independent components such that $\mathrm{E}(\boldsymbol{Z}_{ij}) = 0$ and $\mathrm{Var}(\boldsymbol{Z}_{ij}) = I_p$. We considered two types of innovations: the Gaussian where $\boldsymbol{Z}_{ij} \sim N(0, I_p)$ and the Gamma where each component of $\boldsymbol{Z}_{ij}$ is standardized Gamma$(4, 0.5)$ such that it has zero mean and unit variance. For simplicity, 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 $[p^{1-\beta}]$ non-zero entries of equal value, which were uniformly allocated among $\{1, \cdots, p\}$. Here $[a]$ denotes the integer part of a. The values of the nonzero entries were $\sqrt{2r\log p/n}$ for a set of r-values ranging evenly from 0.1 to 0.4. The covariance matrices $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 =: \mathbf{\Sigma} = (\sigma_{ij})$ where $\sigma_{ij} = \rho^{|i-j|}$ for $1 \le i, j \le p$ and $\rho = 0.6$. The dimension p was 200 and 600, respectively and the sample sizes $n_1 = 30$ and $n_2 = 40$.

The banding width parameter τ in the estimation of $\mathbf{\Omega}$ was chosen according to the data-driven procedure proposed by Bickel and Levina (2008a), which is described as follows. For a given data set, we divided it into two subsamples by repeated (N times) random data split. For the l-th split, $l \in \{1, \cdots, N\}$, we 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 by taking the same approach described in previous section for $\hat{A}_\tau^{(l)}$

and $\hat{D}_\tau^{(l)}$. Also we let $\boldsymbol{S}_n^{(l)}$ be the sample covariance matrix obtained from the second subsample. Then the banding parameter τ is selected as

$$\hat{\tau} = \min_{\tau} \frac{1}{N} \sum_{l=1}^{N} ||\hat{\boldsymbol{\Sigma}}_\tau^{(l)} - \boldsymbol{S}_n^{(l)}||_F, \tag{5.1}$$

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 approximation of the critical values as described in Section 4. We observe that the empirical sizes of the two threshodling tests tended to be larger than the nominal 5% level due to a slow convergence to the extreme value distribution. The proposed parametric bootstrap calibration can significantly improve the size.

To make the power comparison fair, we pre-adjusted the nominal significant levels of all tests such that their empirical sizes were all close to 0.05. We obtain the average empirical power curves (called power profiles) plotted with respect to r and β under each of the simulation settings outlined above based on 1000 simulations. We observed only some very small change in the power profiles when the underlying distribution was switched from the Gaussian to the Gamma, which confirmed the nonparametric nature of the tests considered. Due to the space limitation, we only display in the following the power profiles based on the Gaussian data, and those for the Gamma innovations are given in the supplementary material.

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 two given level of sparsity ($\beta = 0.5$ and 0.6) and $\rho = 0.6$ for Gaussian data. 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. These figures also report the powers of Chen and Qin (2010)'s 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, indicating the effectiveness of the data transformation. The thresholding test with data transformation (Mult2) had the best performance among all the non-Oracle tests. This together with the observed performance of the thresholding test without transformation (Mult1) and the CLX2 shows that the combining the data transformation with the thresholding leads to a quite powerful test performance. The CQ test and the max-norm test without data transformation (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, although it is a proper test under ultra high dimensionality in the sense that the size of the test can be attained and with reasonable power in non-sparse settings. The above features became more pronounced when we increase the dimensionality to $p = 600$ as shown in Figures 1 and 3.

6. EMPIRICAL STUDY

In this section, we demonstrate the performance of the multi-level thresholding test defined in (4.9) on a human breast cancer dataset, available at http://www.ncbi.nlm.nih.gov. The data have been analyzed by Richardson et al. (2006) to provide insight into the molecular pathogenesis of Sporadic basal-like cancers (BLC), a distinct class of human breast cancers. The original microarray gene expression data consist of 7 normal specimens, 2 BRCA-associated breast cancer specimens, 18 sporadic BLC specimens and 20 non-BLC specimens. Since the most of interests on this data set is to display the unique characteristics of BLC relative to non-BLC specimens, we formed two samples. One consists of $n_1 = 18$ BLC cases and another consists of $n_2 = 20$ non-BLC specimens for analysis which form two samples respectively.

Biologically speaking, each gene does not function individually in isolation. Rather, genes tend to work collectively to perform their biological functions. Gene-sets are technically defined in Gene Ontology (GO) system that provides structured vocabularies which produce names of gene-sets (also called GO terms), see Ashburner et al. (2000) for more details.

There were 9918 GO terms, which were obtained from the original data set after we excluded some GO terms with missing information. To accommodate high dimensionality, we further removed those GO terms with the number of genes less than 20 and the number of remaining GO terms varied by chromosomes. In order to take advantage of the inter-gene correlation, we first selected genes from one of 23 chromosomes and then ordered them by their locations on the chromosome. By doing this, genes with adjacent locations are more strongly correlated than genes far away from each other. This would also facilitate the bandable assumption for the covariance matrices. A major motivation in our analysis is to identify sets of genes which are significantly different between the BLC and the non-BLC specimens.

As discussed in Richardson et al. (2006), BLC specimens display X chromosome abnormalities in the sense that most of the BLC cases lack markers of a normal inactive X chromosome, which are rare in non-BLC specimens. Moreover, single nucleotide polymorphism array analysis demonstrated loss of heterozygosity (loss of a normal and functional allele at a heterozygous locus) in chromosome 14 and 17 was quite frequent in BLC specimens, a phenomenon largely missing among non-BLC specimens. Therefore, our main interest was on chromosomes X, 14 and 17.

We applied the multi-level thresholding test based on the data transformation on each of gene-sets in chromosomes X, 14 and 17 by first transforming the data with estimated $\mathbf{\Omega}$ through the Cholesky decomposition discussed in Section 4. We also applied the CQ test to serve as contrasts. By controlling the false discovery rate (Benjamini and Hochberg, 1995) at 0.05, the CQ test declared 81 GO terms significant on chromosome X, 80 out of which were also declared significant by the multi-level thresholding test. However, the multi-thresholding test found 4 more significant GO terms not found significant by the CQ test. Similarly, on chromosome 14, CQ test declared 76 GO terms significant which were all included by the 86 GO terms declared significant by the multi-level thresholding test. On chromosome 17, 5 out of 166 GO terms declared significant by the CQ test were not declared significant by the multi-level thresholding test. On the other hand, 14 out of 175 GO terms declared significant by the multi-level thresholding test were not declared significant by the CQ test.

Table 2 lists the top ten most significant GO terms declared by the multi-level thresholding test on the three chromosomes, respectively. The table also marks those gene-sets which were not tested significant by the CQ test. There were three gene-sets in the top ten which were not declared significant by the CQ test in chromosomes X and 14, and two gene-sets in chromosomes 17. These empirical results support our

theoretically findings that the multi-level thresholding test with data transformation is more powerful than the CQ test by conducting both thresholding and utilizing data dependence.

7. DISCUSSION

Our analysis in this paper shows that the thresholding combined with the data transformation via the estimated precision matrix leads to a very powerful test procedure. The analysis also shows that thresholding alone is not sufficient in lifting the power when there is sufficient amount of dependence in the covariance, and the data transformation is quite crucial. The latter 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 test of thresholding with data transformation can be viewed as a significant improvement of the test of Chen and Qin (2010) for sparse and faint signals. 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 (with both thresholding and data transformation) having better power than the test of Cai, Liu and Xia (2014) with data transformation is due to the thresholding conducted on the L_2 formulation of the test statistics since the proposed test has both thresholding and data transformation whereas CLX test has only the data transformation. The max-norm formulation does not accommodate the need to threshold. This reveals an adavntage of the L_2 formulation.

The results that the proposed test with the estimated covariance can produce lower detection boundary than that of the standard higher criticism test using asymptotic

p-values (Delaigle et al., 2011) is another advantage of the proposal. We want to point out that the study carried out in this paper is not a direct extension from that in Zhong, Chen and Xu (2013). Zhong et al. (2013) considered an alternative L_2-formulation to the higher criticism (HC) test of Donoho and Jin (2004) for one-sample hypotheses. They showed that, although the L_2 formulation attains the same detection boundary as the HC test, the L_2 formulation is more advantageous to the HC when the sparsity and signal strength combination (β, r) is above the detection boundary. However, Zhong et al. (2013) did not study the specific benefits of the thresholding in improving the power of the high dimensional multivariate test and the relative performance to the Oracle test; nor did they considered the data transformation via the precision matrix.

APPENDIX: TECHNICAL DETAILS.

Throughout the Appendix, we assume $n_1 \to \infty$, $n_2 \to \infty$ and let $n = \frac{n_1 n_2}{n_1+n_2}$.

A.1. Lemmas

Lemma 1. We denote $\delta_k = \mu_{1k} - \mu_{2k}$. As $x = o(n^{\frac{1}{3}})$, T_{nk} in (2.1) satisfies

$$\begin{aligned} \mathrm{P}(n\,T_{nk} + 1 > x) &= \{1 + o(1)\} I(\sqrt{n}|\delta_k| > \sqrt{x}) \\ &+ \left[\bar{\Phi}(\sqrt{x} - \sqrt{n}|\delta_k|) + \bar{\Phi}(\sqrt{x} + \sqrt{n}|\delta_k|)\right]\{1 + O(n^{-1/6}) \\ &+ O(\frac{x^{3/2}}{n^{1/2}})\} I(\sqrt{n}|\delta_k| < \sqrt{x}). \end{aligned}$$

Lemma 2. Assume Conditions (C1)-(C2). The mean of the thresholding test statistic $L_n(s)$ is

$$\begin{aligned} \mu_{L_n(s)} &= \sum_{i \in S_\beta^c} \mathrm{E}\{L_{n,i}(s)\} + \sum_{k \in S_\beta} \mathrm{E}\{L_{n,k}(s)\} \\ &= \left(\frac{2}{\sqrt{2\pi}}\sqrt{2s\log p} p^{1-s} + \sum_{k \in S_\beta}\left[n\,\delta_k^2 I\left\{n\,\delta_k^2 > \lambda_n(s)\right\}\right.\right. \\ &+ \left.\left.(2s\log p)\bar{\Phi}(\eta_k^-) I\left\{n\,\delta_k^2 < \lambda_n(s)\right\}\right]\right)\{1 + o(1)\}. \end{aligned} \tag{A.1}$$

Lemma 3. Assume Conditions (C1)-(C3). The variance of the thresholding test statistic $L_n(s)$ is

$$\begin{aligned}\sigma^2_{L_n(s),1} &= \Bigg(\frac{2}{\sqrt{2\pi}}[(2s\log p)^{\frac{3}{2}} + (2s\log p)^{\frac{1}{2}}]p^{1-s} + 4p\bar{\Phi}(\sqrt{2s\log p}) \\ &+ \sum_{k,l\in S_\beta}(4n\,\delta_k\,\delta_l\,\rho_{kl} + 2\rho^2_{kl})I\Big\{n\,\delta^2_k > \lambda_n(s)\Big\}I\Big\{n\,\delta^2_l > \lambda_n(s)\Big\} \\ &+ \sum_{k\in S_\beta}(2s\log p)^2\bar{\Phi}(\eta^-_k)I\Big\{n\,\delta^2_k < \lambda_n(s)\Big\}\Bigg)\{1+o(1)\}. \qquad \text{(A.2)}\end{aligned}$$

Lemma 4. Suppose $\{\boldsymbol{Z}_i\}_{i=1}^p$ is a sequence of α-mixing random variables with zero mean and satisfying

$$M_{2l+\delta} = \sup_i\Big[\mathrm{E}(\boldsymbol{Z}_i)^{2l+\delta}\Big]^{1/(2l+\delta)} < \infty,$$

for $\delta > 0$ and $l \geq 1$. Let $\alpha(i)$ be α-mixing coefficient. Then,

$$\mathrm{E}(\sum_{i=1}^p \boldsymbol{Z}_i)^{2l} \leq Cp^l\Big[M^{2l}_{2l} + M^{2l}_{2l+\delta}\sum_{i=1}^\infty i^{l-1}\alpha(i)^{\delta/(2l+\delta)}\Big],$$

where C is a finite constant positive constant depending only on l.

Lemma 5. Under condition (C4), the following relationship holds:

$$\varpi_{kk}(\tau) = \omega_{kk} + O(\tau^{-C}), \quad \text{for} \quad C > 1,$$

where $\omega_{kk} = \{(1-\kappa)\boldsymbol{\Sigma}_1 + \kappa\boldsymbol{\Sigma}_2\}^{-1}_{kk}$ and $\varpi_{kk}(\tau) = \mathrm{Var}\{\sqrt{n}(\bar{\boldsymbol{Z}}^{(k)}_1(\tau) - \bar{\boldsymbol{Z}}^{(k)}_2(\tau))\}$.

Lemma 6. If $\beta > 1/2$ and the banding parameter $\tau = L_p$ for a slowly varying function L_p, then under condition (C5), with probability approaching 1,

$$\delta_{\boldsymbol{\Omega}(\tau),k} \approx \omega_{kk}\delta_k \quad \text{for} \;\; k \in S_\beta,$$

Lemma 7. For any positive definite matrix $A_{p,p} = (a_{ij})_{p\times p}$ and its inverse $B_{p,p} = (b_{ij})_{p\times p}$, the following inequality holds

$$a_{ii}\cdot b_{ii} \geq 1 \quad i = 1, \cdots, p.$$

Lemma 8. Under the conditions assumed in Theorem 4 and condition (C5), with probability approaching to 1,

$$\beta_{\hat{J}_n(s,\tau)} \geq \beta_{L_n(s)}.$$

A.2. Proofs of Theorems 1, 2 and 3

The proofs are two-sample extension of Theorems 1, 2 and 3 in Zhong, Chen and Xu (2013). We include them in the supplementary material.

A.3. Proof of Theorem 4

We first derive the mean and variance of the transformed thresholding test statistics. Note that by the relationship $\boldsymbol{Z}_{ij}(\tau) = \boldsymbol{\Omega}(\tau)\boldsymbol{X}_{ij}$ 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 $l = 1, \cdots, p$, $Z_{ij}^{(l)}(\tau)$ is sub-Gaussian by using Hölder inequality and mathematical induction. Hence, the large derivation results can be applied to derive the mean and variance of the transformed thresholding test statistic defined in (4.2) by following the similar steps when we derive the mean and variance of the thresholding step. Therefore, to obtain the mean and variance of the transformed thresholding test, we can simply replace δ_k by $\delta_{\boldsymbol{\Omega}(\tau),k}$ and S_β by $S_{\boldsymbol{\Omega}(\tau),\beta}$ in (2.13) and (2.14), respectively, where after the transformation, nonzero signals δ_k becomes $\delta_{\boldsymbol{\Omega}(\tau),k}$ and the set S_β including these nonzero signals becomes $S_{\boldsymbol{\Omega}(\tau),\beta}$.

We first establish the asymptotic normality of transformed thresholding test defined in (4.2) where the banding parameter t is chosen to be a slowly varying function. To this end, we first show that both $\{Z_{1i}^{(k)}(t)\}_{k=1}^p$ and $\{Z_{2i}^{(k)}(t)\}_{k=1}^p$ 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 $A \in \sigma(\mathcal{F}_{\boldsymbol{X},(1,a)}^{(1)}, \mathcal{F}_{\boldsymbol{X},(1,a)}^{(2)})$ and $B \in \sigma(\mathcal{F}_{\boldsymbol{X},(a+k,\infty)}^{(1)}, \mathcal{F}_{\boldsymbol{X},(a+k,\infty)}^{(2)})$,

$$|P(A \cap B) - P(A)P(B)| \to 0 \quad \text{as} \quad k \to \infty.$$

By the relationship between $\boldsymbol{Z}_{1i}(t)$ and $\boldsymbol{X}_{1i}$, for any t,

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

Then as long as $k-2t \to \infty$, $|P(A' \cap B') - P(A')P(B')| \to 0$ for any $A' \in \sigma(\mathcal{F}^{(1)}_{\boldsymbol{Z},(1,a)}, \mathcal{F}^{(2)}_{\boldsymbol{Z},(1,a)})$ and $B' \in \sigma(\mathcal{F}^{(1)}_{\boldsymbol{Z},(a+k,\infty)}, \mathcal{F}^{(2)}_{\boldsymbol{Z},(a+k,\infty)})$. It follows that

$$\alpha_{\boldsymbol{Z}_1(t)}(k) = \alpha_{\boldsymbol{X}_1}(k-2t) \quad \text{if} \quad k > 2t.$$

Therefore, $\alpha_{\boldsymbol{Z}_1(t)}(k) \to 0$ as $k-2t \to \infty$ where $\alpha_{\boldsymbol{Z}_1(t)}$ is the α-mixing coefficient for the sequence $\{Z_{1j}^{(k)}(t)\}_{k=1}^{p}$. Similarly, it can be shown that $\alpha_{Z_2(t)}(k) \to 0$ as $k-2t \to \infty$. Thus, both $\{Z_{1i}^{(k)}(t)\}_{k=1}^{p}$ and $\{Z_{2i}^{(k)}(t)\}_{k=1}^{p}$ are α-mixing sequences. Then the asymptotic normality of $J_n(s,t)$ can be established by applying the Bernstein's blocking method as we have done in the proof of Theorem 1. To further establish the normality of $\hat{J}_n(s,\tau)$, we note that our $\hat{J}_n$ can be written as

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

where $\hat{S}_{nk} = n(\bar{\hat{Z}}_1^{(k)} - \bar{\hat{Z}}_2^{(k)})^2$ and $S_{nk} = n(\bar{Z}_1^{(k)}(t) - \bar{Z}_2^{(k)}(t))^2$. 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

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

Using $\mathrm{E}\{\sum_{k=1}^{p} \mathrm{I}(\frac{S_{nk}}{\varpi_{kk}} > \lambda_n)\} = \sum_{k=1}^{p} \mathrm{P}(\frac{S_{nk}}{\varpi_{kk}} > \lambda_n)$, and from Lemma 1,

$$\sum_{k=1}^{p} \mathrm{P}(\frac{S_{nk}}{\varpi_{kk}} > \lambda_n) = O(\frac{p^{1-s}}{\sqrt{2s\log p}}),$$

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 $\hat{S}_{nk} = n\{\sum_l \hat{\omega}_{kl}(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})\}^2$ and $S_{nk} = n\{\sum_l \omega_{kl}(t)(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})\}^2$. Then,

$$\begin{aligned}
\max_k |\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}}| &\leq \max_l n(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})^2 \max_k \frac{(\sum_l \omega_{kl})^2}{\omega_{kk}^2} |\hat{\omega}_{kk} - \omega_{kk}|\{1 + o(1)\} \\
&+ \max_l n(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})^2 \max_k \frac{\sum_l |\omega_{kl} + \sqrt{\frac{\omega_{kk}}{\varpi_{kk}}}\omega_{kl}(t)|}{\omega_{kk}} \Big\{ \sum_l |\hat{\omega}_{kl} - \omega_{kl}| \\
&+ \sum_l |\omega_{kl} - \sqrt{\frac{\omega_{kk}}{\varpi_{kk}}}\omega_{kl}(t)| \Big\} \\
&\leq M \max_l n(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})^2 \max_k \Big\{ \sum_{l=1}^{p} |\hat{\omega}_{kl} - \omega_{kl}| + t^{-a} + O(t^{-C}) \Big\},
\end{aligned}$$

where $M > 0$, $a > 0$ and we use the fact that $\varpi_{kk} = \omega_{kk} + O(t^{-C})$ from Lemma 5. From the fact that $\max_l n(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})^2 = O_p(\log p)$ and $\max_k \sum_{l=1}^{p} |\hat{\omega}_{kl} - \omega_{kl}| = O_p[(\frac{\log p}{n})^{q/2}]$ for any q such that $1/(\alpha + 1) < q < 1$ (See Bickel and Levina, 2008b), we know

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

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

For the second term II, we have

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

Because the proofs for II_1 and II_2 are similar, we only show II_2 in the following.

First, we note that

$$\max_k |\frac{S_{nk}}{\varpi_{kk}} + 1| \leq 1 + \max_k \frac{(\sum_l \omega_{kl}(t))^2}{\varpi_{kk}} \max_l n(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})^2 = O_p(\log p).$$

And,

$$\begin{aligned}
&\sum_{k=1}^{p} I\Big\{|\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}}| > |\frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s)|\Big\} \\
\leq\ & \sum_{k=1}^{p} I(|\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}}| > h) + \sum_{k=1}^{p} I(|\frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s)| < h). \qquad \text{(A.3)}
\end{aligned}$$

The second indicator function on the right above can be evaluated by the following:

$$\begin{aligned}
\mathrm{E}\Big\{\sum_{k=1}^{p} I(|\frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s)| < h)\Big\} &= \sum_{k=1}^{p} \mathrm{P}(|\frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s)| < h) \\
&= \sum_{k=1}^{p}\Big\{\bar{\Phi}(\sqrt{\lambda_n(s) - h}) - \bar{\Phi}(\sqrt{\lambda_n(s) + h})\Big\} \\
&= \frac{h}{\sqrt{2s\mathrm{log}p}} p^{1-s}.
\end{aligned}$$

Therefore, in (A.3), $\sum_{k=1}^{p} I(|\frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s)| < h) = O_p(\frac{h}{\sqrt{2s\mathrm{log}p}} p^{1-s})$. To evaluate $\sum_{k=1}^{p} I(|\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}}| > h)$ in (A.3), we use the same approach. First, notice that

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

Then,

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

where, if we choose $T = C\sqrt{\mathrm{log}p/n}$, $\sum_{k=1}^{p} \mathrm{P}(\max_l |\bar{X}_1^{(l)} - \bar{X}_2^{(l)}| > T) \leq p^{2-C} \to 0$, for sufficient large C. If $h = C^* \mathrm{log}p(\frac{\mathrm{log}p}{n})^{q/2}$, there exists a $a > 0$ such that

$$\sum_{k=1}^{p} \mathrm{P}(\sum_{l=1}^{p} |\hat{\omega}_{kl} - \omega_{kl}| > \frac{h}{MnT^2}) = \sum_{k=1}^{p} \mathrm{P}(\sum_{l=1}^{p} |\hat{\omega}_{kl} - \omega_{kl}| > M'(\frac{\mathrm{log}p}{n})^{q/2}) \leq p^{1-a}.$$

Therefore, by choosing C^* large enough such that $a > q\theta/2$, $\sum_{k=1}^{p} I(|\frac{\hat{S}_{nk}}{\hat{\omega}_{kk}} - \frac{S_{nk}}{\varpi_{kk}}| > h) = O_p(p^{1-a}) = o_p(pn^{-q/2})$ for $p = n^{1/\theta}$. Under the choice $h = C^* \mathrm{log}p(\frac{\mathrm{log}p}{n})^{q/2}$,

$\sum_{k=1}^{p} I(|\frac{S_{nk}}{\varpi_{kk}} - \lambda_n(s)| < h) = O_p(L_p n^{-\frac{q}{2}} p^{1-s})$. In addition, we know that $\max_k I\{\frac{S_{nk}}{\varpi_{kk}} > \lambda_n(s)\} = O_p(p^{-s})$. Therefore, we know that $\mathrm{II}_2 = o_p(L_p p^{1-s} n^{-q/2})$. Similarly, one can show that $\mathrm{II}_1 = o_p(L_p p^{1-s} n^{-q/2})$. In summary, $\mathrm{II}/\sigma_{J_n,0} = o_p(\mathrm{I}/\sigma_{J_n,0}) = o_p(1)$. This completes the proof of Theorem 4.

A.4. Proof of Theorem 5

The proof of Theorem 5 is similar to that of Theorem 2. We omit it.

A.5. Proof of Theorem 6

We first consider $\mathbf{\Omega}$ is known. We know that the power of the transformed thresholding test is determined by

$$\mathrm{SNR}_{J_n(s,\tau)} = \frac{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),1}}.$$

Recall that for $k \in S_\beta$, $\underline{\omega}\delta_k^2 \le \frac{\delta^2_{\mathbf{\Omega}(\tau),k}}{\varpi_{kk}(\tau)} \le \bar{\omega}\delta_k^2$. Then, we have the following inequality

$$\frac{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),1}} \ge \frac{M_1}{V_1}, \tag{A.4}$$

where $M_1 = \sum_{k\in S_\beta} \left\{ n\underline{\omega}\delta_k^2 I(n\underline{\omega}\delta_k^2 > 2s\log p) + (2s\log p)\bar{\Phi}(\eta_k^-) I(n\underline{\omega}\delta_k^2 < 2s\log p) \right\}$ and

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

Note that M_1/V_1 is the signal-to-noise ratio of the thresholding test without the transformation. But the signal $n\underline{\omega}\delta_k^2 = 2\underline{\omega} r \log p$. From the proof of Theorem 3, we know that $M_1/V_1 \to \infty$ as long as s is properly chosen and $\underline{\omega} r > \varrho(\beta)$. Therefore,

$$\frac{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),1}} \to \infty,$$

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 6, 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

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

Then, we can choose $\alpha_n = \bar{\Phi}\{(\log p)^\epsilon\} \to 0$ as $p \to \infty$ for any small number $\epsilon > 0$ 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 Theorem 3 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 6. Note that

$$\frac{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),1}} \leq \frac{M_2}{V_2},$$

where $M_2 = \sum_{k \in S_\beta} \bigg\{ n\bar{\omega}\delta_k^2 I(n\bar{\omega}\delta_k^2 > 2s\log p) + (2s\log p)\bar{\Phi}(\eta_k^-) I(n\bar{\omega}\delta_k^2 < 2s\log p) \bigg\}$ and

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

We also 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 $\bar{\omega} r < \varrho(\beta)$, i.e.,

$$\frac{\mu_{J_n(s,\tau),1} - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),1}} \to 0.$$

Similar to the proof for the second statement of Theorem 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 Theorem 3, 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 $G^\star_\alpha = \{q_\alpha + b(\log p, \eta^\star - \theta)\}/a(\log p)$. If we choose $\alpha_n = \bar{\Phi}\{(\log p)^\epsilon\} \to 0$ as $p \to \infty$ for any small number $\epsilon > 0$, $G^\star_\alpha = O\{(\log\log p)^{1/2}\}$. We only show that if $r > \underline{\omega}^{-1}\varrho_\theta(\beta)$, 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 $r < \bar{\omega}^{-1}\varrho_\theta(\beta)$ is similar to the proof for M_{J_n}. We notice that

$$\begin{aligned}
\mathrm{P}(M_{\hat{J}_n} \ge G^\star_\alpha | H_1) &\ge \mathrm{P}\left(\frac{\hat{J}_n(s,\tau) - \hat{\mu}_{J_n(s,\tau),0}}{\hat{\sigma}_{J_n(s,\tau),0}} \ge G^\star_\alpha | H_1\right) \\
&= \mathrm{P}\Bigg\{\left(\frac{J_n(s,\tau) - \mu_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),0}} + \frac{\mu_{J_n(s,\tau),0} - \hat{\mu}_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),0}}\right. \\
&+ \left. o_p(1)\right)\frac{\sigma_{J_n(s,\tau),0}}{\hat{\sigma}_{J_n(s,\tau),0}} \ge G^\star_\alpha | H_1\Bigg\}, \qquad \text{(A.5)}
\end{aligned}$$

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 5. Moreover, as shown in Zhong, Chen and Xu (2013), with $p = n^{1/\theta}$ for $0 < \theta < 1$,

$$\frac{\mu_{J_n(s,\tau),0} - \hat{\mu}_{J_n(s,\tau),0}}{\sigma_{J_n(s,\tau),0}} \to 0, \quad \text{and} \quad \frac{\sigma_{J_n(s,\tau),0}}{\hat{\sigma}_{J_n(s,\tau),0}} \to 1.$$

Then the probability in (A.5) is determined by

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

where $(J_n(s,\tau) - \mu_{J_n(s,\tau),1})/(G^\star_\alpha \sigma_{J_n(s,\tau),1}) = o_p(1)$, and $\sigma_{J_n(s,\tau),1} > \sigma_{J_n(s,\tau),0}$. Therefore, 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.5) tends 1. From

inequality (A.4), we only need to show that with properly chosen s, $M_1/(G_\alpha^\star V_1) \to \infty$. As we have shown in Theorem 5, we need to choose the level of the threshold $s \in (1-\theta, 1)$ 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 $s > 1-\theta$ on the four cases in the proof of Theorem 3. Similar to the result in Delaigle, Hall and Jin (2011), and Zhong, Chen and Xu (2013), the modified detection boundary is given by (4.11). As a result, we know that $M_1/(G_\alpha^\star V_1) \to \infty$ if $\underline{\omega} r > \varrho_\theta(\beta)$. This shows that if $r > \underline{\omega}^{-1}\varrho_\theta(\beta)$, the power of $M_{\hat{J}_n}$ tends to 1.

REFERENCE

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25-29.

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

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

Bickel, P. and Levina, E. (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics*, **36**, 199-227.

Bickel, P. and Levina, E. (2008b). Covariance regularization by thresholding. *The Annals of Statistics*, **36**, 2577-2604.

Cai, T., Liu, W. and Luo, X. (2011). A constrained l_1 minimization approach to

sparse precision matrix estimation. *Journal of the American Statistical Association*, **106**, 594-607.

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

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

Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Students t-statistic. *Journal of the Royal Statistical Society: Series B*, **73**, 283-301.

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

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425-455.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, **91**, 674-688.

Feng, L., Zou, C., Wang, Z. and Zhu, L. (2013). Two-sample Behrens-Fisher problem for high-dimensional data. Manuscript.

Gröchenig, K. and Leinert, M. (2006). Symmetry and inverse-closedness of matrix algebra and functional calculus for infinite matrices. *Transactions of the American Mathematical Society*, **358**, 2695-2711.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, **38**, 1686-1732.

Ingster, Y. I. (1997). Some problems of hypothesis testing leading to innitely divisible distributions. *Mathematical Methods of Statistics*, **6**, 47-69.

Jaffard, S. (1990). Propriétés des matrices "bien localisées" près de leur diagonale et quelques applications. *Annales de l Institut Henri Poincare (C) Non Linear*

Analysis, **7**, 461-476.

Ji, P. and Jin, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics*, **40**, 73-103.

Kim, T. Y. (1994). Moment bounds for non-stationary dependent sequences. *Journal of Applied Probability*, **31**, 731-742.

Petrov, V. V. (1995). *Limit theorems of probability theory: sequences of independent random variables*. Clarendon Press, London.

Richardson, A., Wang, Z., Nicolo, A., Lu, X., Brown, M., Miron, A., Liao, X., Iglehart, J., Livingston, D. and Ganesan, S. (2006). X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, **9**, 121-132.

Srivastava, M., Katayama, S. and Kano, Y. (2013). A two sample test in high dimensional data. *Journal of Multivariate Analysis*, **114**, 349-358.

Sun, Q. (2005). Wiener's lemma for infinite matrices with polynomial off-diagonal decay. *Comptes Rendus Mathematique*, **340**, 567-570.

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

Table 1: Empirical sizes of the proposed multi-thresholding tests with (Mult2) and without data transformation (Mult1), Cai, Liu and Xia's max-norm tests with (CLX2) and without (CLX1) data transformation, Chen and Qin's test (CQ) and the Oracle test for Gaussian and Gamma data. Parametric bootstrapping was conducted in Mult1* and Mult2* to calibrate the size distortion of Mult1 and Mult2.

p	(n_1, n_2)	Oracle	CQ	CLX1	CLX2	Mult1 (Mult1*)	Mult2 (Mult2*)
				Normal			
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)

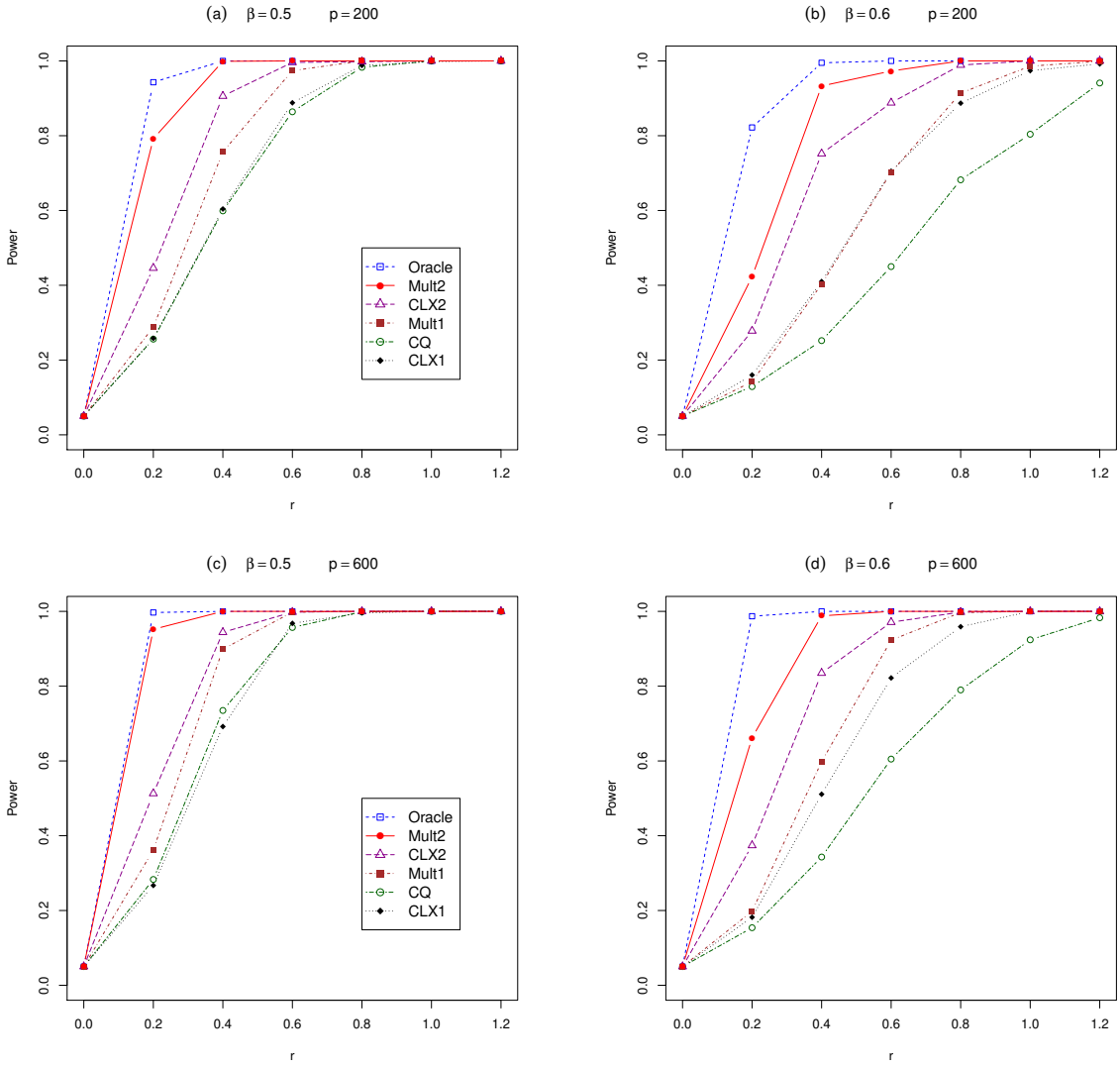

Figure 1: Average Power with respect to the signal strength r of the proposed multi-thresholding tests with (Mult2) and without data transformation (Mult1), Cai, Liu and Xia's max-norm tests with (CLX2) and without (CLX1) data transformation, Chen and Qin's test (CQ) and the Oracle test for Gaussian data with $n_1 = 30$ and $n_2 = 40$.

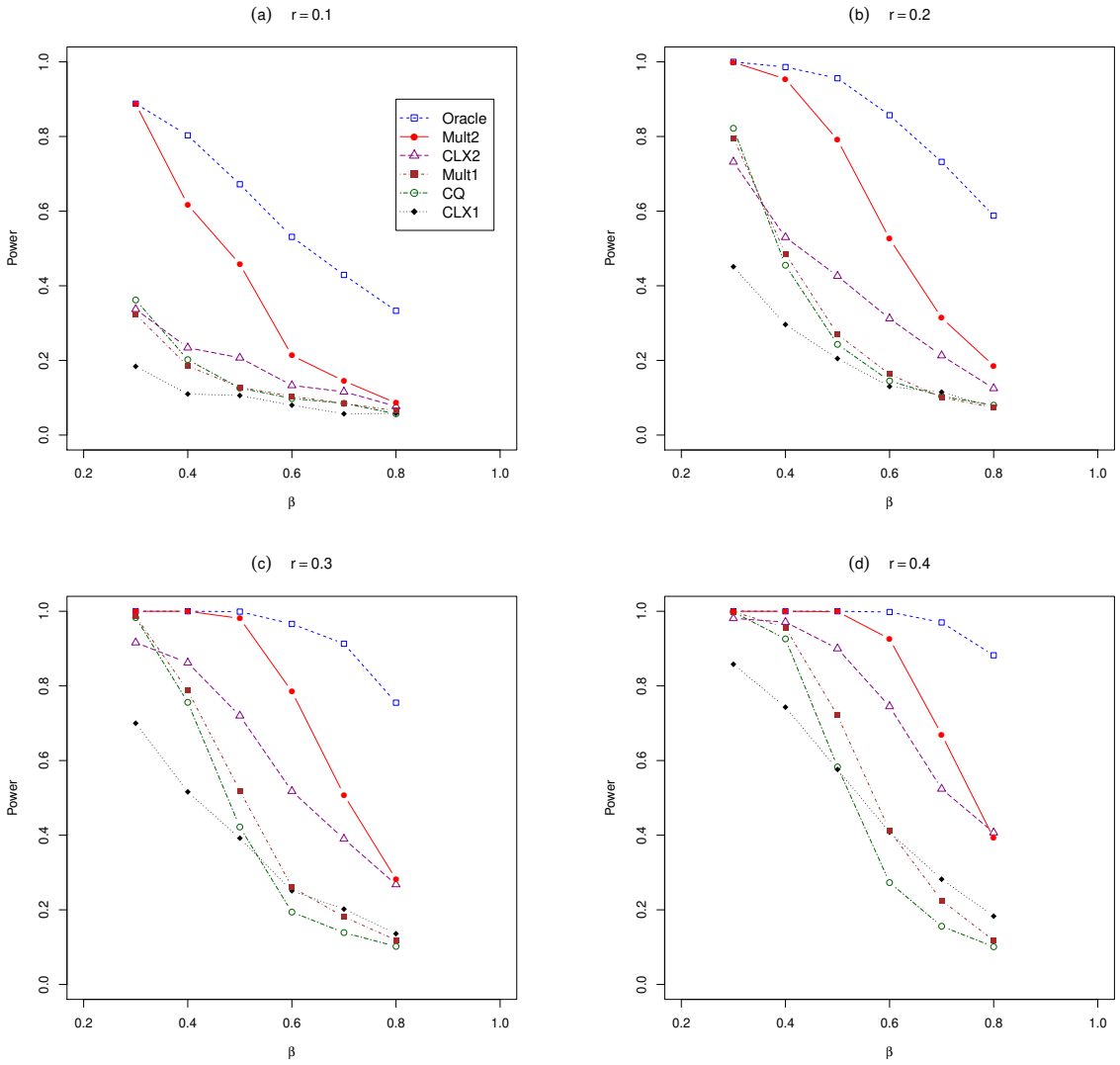

Figure 2: Average Power with respect to the sparsity β of the proposed multi-thresholding tests with (Mult2) and without data transformation (Mult1), Cai, Liu and Xia's max-norm tests with (CLX2) and without (CLX1) data transformation, Chen and Qin's test (CQ) and the Oracle test for Gaussian data with $p = 200$, $n_1 = 30$ and $n_2 = 40$.

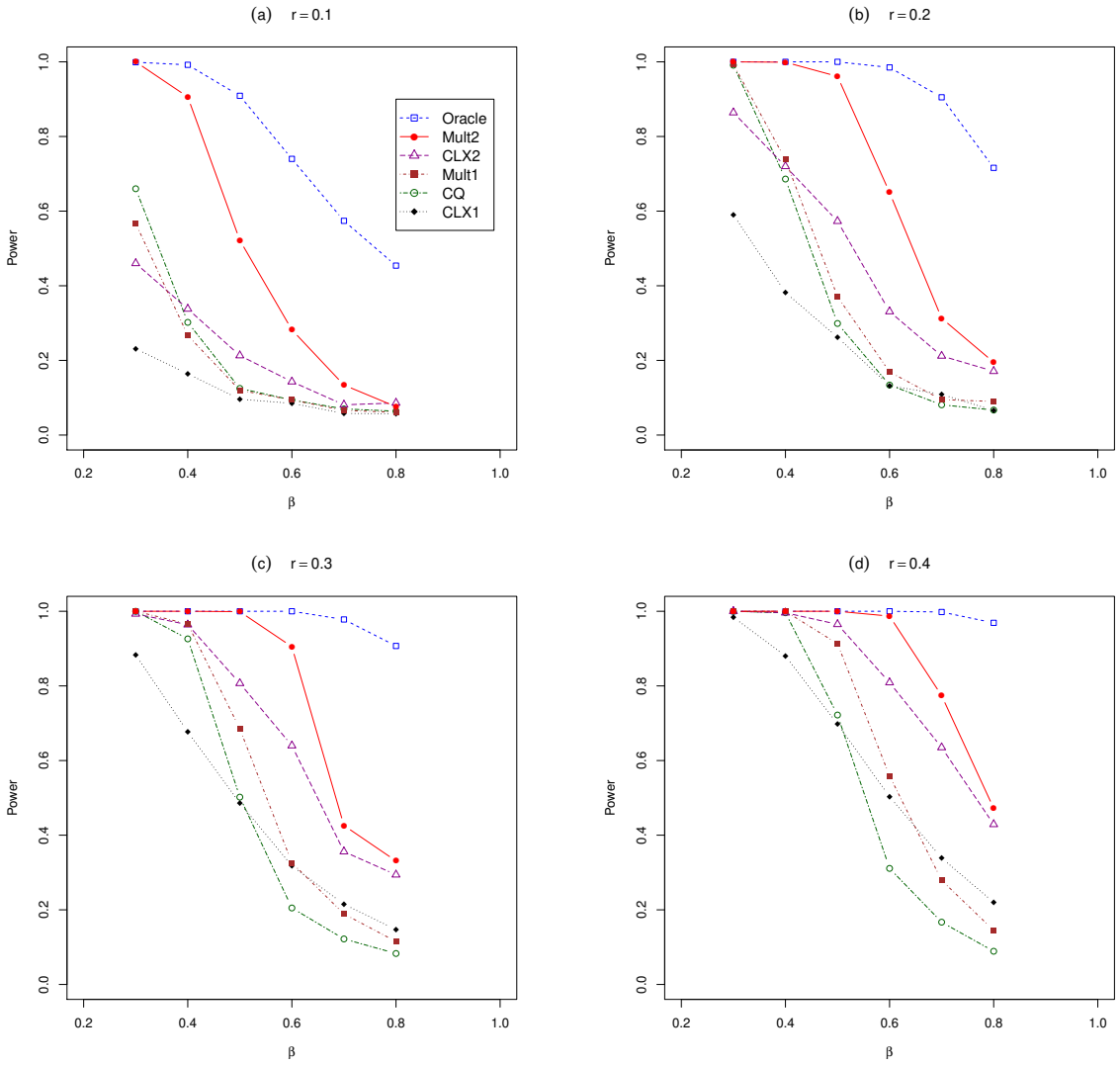

Figure 3: Average Power with respect to the sparsity β of the proposed multi-thresholding tests with (Mult2) and without data transformation (Mult1), Cai, Liu and Xia's max-norm tests with (CLX2) and without (CLX1) data transformation, Chen and Qin's test (CQ) and the Oracle test for Gaussian data with $p = 600$, $n_1 = 30$ and $n_2 = 40$.

Table 2: Top ten most significant GO terms by the multi-level thresholding test on chromosomes X, 14 and 17 with false discovery rate at 0.05, where $*$ refers to GO terms not being declared significant by the CQ test.

Chromosome X		Chromosome 14		Chromosome 17	
GO ID	GO term name	GO ID	GO term name	GO ID	GO term name
0005524	ATP binding	0005524	ATP binding	0005524	ATP binding
0000166	nucleotide binding	0000166	nucleotide binding	0000166	nucleotide binding
0005515	protein binding	0005515	protein binding*	0005515	protein binding
0004872	receptor activity	0016740	transferase activity*	0004872	receptor activity*
0016020	membrane	0006468	protein amino acid phosphorylation	0016740	transferase activity
0005634	nucleus	0005576	extracellular region*	0006468	protein amino acid phosphorylation
0003677	DNA binding	0016020	membrane	0005576	extracellular region
0003700	transcription factor activity *	0005634	nucleus	0005887	integral to plasma membrane
0007165	signal transduction*	0003677	DNA binding	0016020	membrane
0046872	metal ion binding*	0003700	transcription factor activity	0005654	nucleoplasm*