Two-Sample Test of High Dimensional Means under Dependency *

T. Tony Cai, Weidong Liu and Yin Xia

Abstract

This paper considers in the high dimensional setting a canonical testing problem in multivariate analysis, namely testing the equality of two mean vectors. We introduce a new test statistic that is based on a linear transformation of the data by the precision matrix which incorporates the correlations among the variables. Limiting null distribution of the test statistic and the power of the test are analyzed. It is shown that the test is particularly powerful against sparse alternatives and enjoys certain optimality. A simulation study is carried out to examine the numerical performance of the test and compare with other tests given in the literature. The results show that the proposed test significantly outperforms those tests in a range of settings.

Keywords: Covariance matrix, extreme value distribution, high dimensional test, hypothesis testing, limiting null distribution, power, precision matrix, testing equality of mean vectors.

*Tony Cai is Dorothy Silberberg Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (Email:tcai@wharton.upenn.edu). The research was supported in part by NSF FRG Grant DMS-0854973. Weidong Liu is Professor, Department of Mathematics, Institute of Natural Sciences and MOE-LSC, Shanghai Jiao Tong University, Shanghai, China (Email: liuweidong99@gmail.com). Yin Xia is Ph.D student, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (Email: xiayin@sas.upenn.edu).

1 Introduction

A canonical testing problem in multivariate analysis is that of testing the equality of two mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ based on independent random samples, one from a distribution with mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}$ and another from a distribution with mean $\boldsymbol{\mu}_2$ and the same covariance matrix $\boldsymbol{\Sigma}$. This testing problem arises in many scientific applications, including genetics, econometrics, and signal processing. In the Gaussian setting where one observes $\boldsymbol{X}_k \overset{iid}{\sim} N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $k = 1, ..., n_1$, and $\boldsymbol{Y}_k \overset{iid}{\sim} N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, $k = 1, ..., n_2$, the classical test for testing the hypotheses

$$H_0 : \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \quad \text{versus} \quad H_1 : \quad \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \tag{1}$$

is Hotelling's T^2 test with the test statistic given by

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})' \widehat{\boldsymbol{\Sigma}}^{-1} (\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}),$$

where $\bar{\boldsymbol{X}} = n_1^{-1} \sum_{k=1}^{n_1} \boldsymbol{X}_k$ and $\bar{\boldsymbol{Y}} = n_2^{-1} \sum_{k=1}^{n_2} \boldsymbol{Y}_k$ are the sample means and $\widehat{\boldsymbol{\Sigma}}$ is the sample covariance matrix. The properties of Hotelling's T^2 test has been well studied in the conventional low-dimensional setting. It enjoys desirable properties when the dimension p is fixed. See, e.g., Anderson (2003).

In many contemporary applications, high dimensional data, whose dimension is often comparable to or even much larger than the sample size, are commonly available. Examples include genomics, medical imaging, risk management, and web search problems. In such high dimensional settings, classical methods designed for the low-dimensional case either perform poorly or are no longer applicable. For example, the performance of Hotelling's T^2 test is unsatisfactory when the dimension is high relative to the sample sizes.

Several proposals for correcting Hotelling's T^2 statistic have been introduced in the high dimensional settings. For example, Bai and Saranadasa (1996) proposed to remove $\widehat{\boldsymbol{\Sigma}}^{-1}$ in T^2 and introduced a new statistic based on the squared Euclidean norm $\|\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}\|_2^2$. Srivastava and Du (2008) and Srivastava (2009) constructed test statistics by replacing $\widehat{\boldsymbol{\Sigma}}^{-1}$ with the inverse of the diagonal of $\widehat{\boldsymbol{\Sigma}}$. Chen and Qin (2010) introduced a test statistic

by removing the cross-product terms $\sum_{i=1}^{n_1} \boldsymbol{X}_i' \boldsymbol{X}_i$ and $\sum_{i=1}^{n_2} \boldsymbol{Y}_i' \boldsymbol{Y}_i$ in $\|\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}\|_2^2$. All of the above test statistics are based on an estimator of $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{A} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for some given positive definite matrix $\boldsymbol{A}$. We shall call these test statistics sum of squares type statistics as they all aim to estimate the squared Euclidean norm $\|\boldsymbol{A}^{\frac{1}{2}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\|_2^2$.

It is known that tests based on the sum of squares type statistics can have good power against the "dense" alternatives. That is, under the alternative hypothesis H_1 the signals in $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ spread out over a large number of coordinates. For a range of applications including anomaly detection, medical imaging, and genomics, however, the means of the two populations are typically either identical or are quite similar in the sense that they only possibly differ in a small number of coordinates. In other words, under the alternative H_1, the difference of the two means $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is sparse. For example, for ultrasonic flaw detection in highly-scattering materials, many scattering centers such as grain boundaries produce echoes and the ensemble of these echoes is usually defined as background noise, while small cracks, flaws, or other metallurgical defects would be defined as signals. See, for example, Zhang, Zhang and Wang (2000). In this case, it is natural to take $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ to be sparse when the metallurgical defects exist. Similarly, for detection of hydrocarbons in materials, instantaneous spectral analysis is often used to detect hydrocarbons through low-frequency shadows, which is usually considered as sparse signals. See Castagna, Sun and Siegfried (2003). In medical imaging, MRI is commonly used for breast cancer detection. It is used to visualize microcalcifications, which can be an indication of breast cancer. The signals are rare in such applications, see James, Clymer and Schmalbrock (2001). Another application is the shape analysis of brain structures, in which the shape differences, if any, are commonly assumed to be confined to a small number of isolated regions inside the whole brain. This is equivalent to the sparse alternative. See Cao and Worsley (1999) and Taylor and Worsley (2008). In these sparse settings, tests based on the sum of squares type statistics are not powerful. For example, the three tests mentioned earlier all require $(n_1+n_2)\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 / p^{\frac{1}{2}} \to \infty$ in order for any of the tests to be able to distinguish between the null and the alternative with probability tending to 1.

The goal of this paper is to develop a test that performs well in general and is particularly powerful against sparse alternatives in the high dimensional setting under dependency. To explore the advantages of the dependence between the variables, we introduce a new test statistic that is based on a linear transformation of the observations by the precision matrix $\mathbf{\Omega}$. Suppose for the moment the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ is known. For testing the null hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, we first transform the samples $\{\boldsymbol{X}_k; 1 \le k \le n_1\}$ and $\{\boldsymbol{Y}_k; 1 \le k \le n_2\}$ by multiplying with $\mathbf{\Omega}$ to obtain the transformed samples $\{\mathbf{\Omega}\boldsymbol{X}_k; 1 \le k \le n_1\}$ and $\{\mathbf{\Omega}\boldsymbol{Y}_k; 1 \le k \le n_2\}$. The new test statistic is then defined to be the maximum of the squared two sample t-statistics of the transformed observations $\{\mathbf{\Omega}\boldsymbol{X}_k; 1 \le k \le n_1\}$ and $\{\mathbf{\Omega}\boldsymbol{Y}_k; 1 \le k \le n_2\}$. We shall first show that the limiting null distribution of this test statistic is the extreme value distribution of type I, and then construct an asymptotically α level test based on the limiting distribution. It is shown that this test enjoys certain optimality and uniformly outperforms two other natural tests against sparse alternatives. The asymptotic properties including the power of the tests are investigated in Section 3.

The covariance matrix $\mathbf{\Sigma}$ and the precision matrix $\mathbf{\Omega}$ are typically unknown in practice and thus need to be estimated. Estimation of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ in the high dimensional setting has been well studied in the last few years. See, for example, Yuan and Lin (2007), Bickel and Levina (2008), Rothman et al. (2008), Ravikumar et al. (2008), Cai, Zhang and Zhou (2010), Yuan (2010), Cai and Liu (2011), Cai, Liu, and Luo (2011), and Cai and Yuan (2012). In particular, when $\mathbf{\Omega}$ is sparse, it can be well estimated by the constrained ℓ_1 minimization method proposed in Cai, Liu, and Luo (2011). When such information is not available, the adaptive thresholding procedure introduced in Cai and Liu (2011) can be applied to estimate $\mathbf{\Sigma}$ and its inverse is then used to estimate $\mathbf{\Omega}$. The estimate of $\mathbf{\Omega}$ is then plugged into the test statistic mentioned above to yield a data driven procedure. In principle, other "good" estimators of $\mathbf{\Omega}$ can also be used. It is shown that, under regularity conditions, the data-driven test performs asymptotically as well as the test based on the oracle statistic and thus shares the same optimality.

A simulation study is carried out to investigate the numerical performance of the pro-

posed test in a wide range of settings. The numerical results show that the power of proposed test uniformly and significantly dominates those of the tests based on the sum of squares type statistics when either Σ or Ω is sparse. When both Σ and Ω are non-sparse, the proposed test with the inverse of the adaptive thresholding estimator of Σ still significantly outperforms the sum of squares type tests.

The rest of the paper is organized as follows. After reviewing basic notation and definitions, Section 2 introduces the new test statistics. Theoretical properties of the proposed tests are investigated in Section 3. Limiting null distributions of the test statistics and the power of the tests, both for the case the precision matrix $\mathbf{\Omega}$ is known and the case $\mathbf{\Omega}$ is unknown, are analyzed. Extensions to the non-Gaussian distributions are given in Section 4. A simulation study is carried out in Section 5 to investigate the numerical performance of the tests. Discussions of the results and other related work are given in Section 6. The proofs of the main results are delegated in Section 7. Additional simulation results and theoretical analysis are given in the supplement Cai, Liu and Xia (2013b).

2 Methodology

This section considers the testing problem in the setting of Gaussian distributions. Extensions to the non-Gaussian case will be discussed in Section 4. We shall first present our testing procedure in the oracle setting in Section 2.1 where the precision matrix $\mathbf{\Omega}$ is assumed to be known. In addition, two other natural testing procedures are introduced in this setting. A data-driven procedure is given in Section 2.2 for the general case of the unknown precision matrix $\mathbf{\Omega}$ by using an estimator of the precision matrix $\mathbf{\Omega}$.

We begin with basic notation and definitions. For a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$, define the ℓ_q norm by $|\boldsymbol{\beta}|_q = (\sum_{i=1}^p |\beta_i|^q)^{1/q}$ for $1 \le q \le \infty$ with the usual modification for $q = \infty$. A vector $\boldsymbol{\beta}$ is called k-sparse if it has at most k nonzero entries. For a matrix $\mathbf{\Omega} = (\omega_{i,j})_{p\times p}$, the matrix 1-norm is the maximum absolute column sum, $\|\mathbf{\Omega}\|_{L_1} = \max_{1\le j\le p} \sum_{i=1}^p |\omega_{i,j}|$, the matrix elementwise infinity norm is defined to be $|\mathbf{\Omega}|_\infty = \max_{1\le i,j\le p} |\omega_{i,j}|$ and the elementwise ℓ_1 norm is $\|\mathbf{\Omega}\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\omega_{i,j}|$. For a matrix $\mathbf{\Omega}$, we say $\mathbf{\Omega}$ is k-sparse if

each row/column has at most k nonzero entries. We shall denote the difference $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ by $\boldsymbol{\delta}$ so the null hypothesis can be equivalently written as $H_0 : \boldsymbol{\delta} = 0$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, write $a_n = O(b_n)$ if there exists a constant C such that $|a_n| \leq C|b_n|$ holds for all sufficiently large n, write $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$, and write $a_n \asymp b_n$ if there are positive constants c and C such that $c \leq a_n/b_n \leq C$ for all $n \geq 1$.

2.1 Oracle Procedures

Suppose we observe independent p-dimensional random samples

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_{n_1} \overset{iid}{\sim} N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \text{ and } \boldsymbol{Y}_1, ..., \boldsymbol{Y}_{n_2} \overset{iid}{\sim} N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

where the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is known. In this case, the null hypothesis $H_0 : \boldsymbol{\delta} = 0$ is equivalent to $H_0 : \boldsymbol{\Omega\delta} = 0$. An unbiased estimator of $\boldsymbol{\Omega\delta}$ is the sample mean vector $\boldsymbol{\Omega}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}) =: \bar{\boldsymbol{Z}} = (\bar{Z}_1, \ldots, \bar{Z}_p)^T$. We propose to test the null hypothesis $H_0 : \boldsymbol{\delta} = 0$ based on the test statistic

$$M_{\boldsymbol{\Omega}} = \frac{n_1 n_2}{n_1 + n_2} \max_{1\leq i\leq p} \frac{\bar{Z}_i^2}{\omega_{i,i}}. \tag{2}$$

At first sight, the test statistic $M_{\boldsymbol{\Omega}}$ is not the most intuitive choice for testing $H_0 : \boldsymbol{\delta} = 0$. We first briefly illustrate the motivation on the linear transformation of the data by the precision matrix. Under a sparse alternative, the power of a test mainly depends on the magnitudes of the signals (nonzero coordinates of $\boldsymbol{\delta}$) and the number of the signals. It will be shown in Section 7 that $(\boldsymbol{\Omega\delta})_i$ is approximately equal to $\delta_i\omega_{i,i}$ for all i in the support of $\boldsymbol{\delta}$. The magnitudes of the nonzero signals δ_i are then transformed to $|\delta_i|\omega_{i,i}^{\frac{1}{2}}$ after normalized by the standard deviation of the transformed variable $(\boldsymbol{\Omega X})_i$. In comparison, the magnitudes of the signals in the original data are $|\delta_i|/\sigma_{i,i}^{\frac{1}{2}}$. It can be seen from the elementary inequality $\omega_{i,i}\sigma_{i,i} \geq 1$ for $1 \leq i \leq p$ that $|\delta_i|\omega_{i,i}^{\frac{1}{2}} \geq |\delta_i|/\sigma_{i,i}^{\frac{1}{2}}$. That is, such a linear transformation magnifies the signals and the number of the signals due to the dependence in the data. The transformation thus helps to distinguish the null and alternative hypotheses. The advantage of this linear transformation will be proved rigorously in Section 7. In the context of signal detection under a Gaussian mixture model, Hall and Jin (2010) introduced

the innovated higher criticism procedure which is also based on the transformation of precision matrix. We should note that the innovated higher criticism procedure is only for the detection purpose, and it does not provide an asymptotically α-level test.

The asymptotic null distribution of $M_{\boldsymbol{\Omega}}$ will be studied in Section 3. Note that $M_{\boldsymbol{\Omega}}$ is the maximum of p dependent normal random variables. It is well known that the limiting distribution of the maximum of p independent χ^2_1 random variables after normalization is the extreme value distribution of type I. This result was generalized by Berman (1964) to the dependent case, where the limiting distribution for the maximum of a stationary sequence was considered. In the setting of the present paper, the precision matrix $\boldsymbol{\Omega}$ does not have any natural order and the result in Berman (1964) thus does not apply. We shall prove by using different techniques that $M_{\boldsymbol{\Omega}}$ still converges to the extreme value distribution of type I under the null H_0.

More generally, for a given invertible $p \times p$ matrix $\boldsymbol{A}$, the null hypothesis $H_0 : \boldsymbol{\delta} = 0$ is equivalent to $H_0 : \boldsymbol{A\delta} = 0$. Set $\boldsymbol{\delta^A} = (\delta_1^{\boldsymbol{A}}, ..., \delta_p^{\boldsymbol{A}})' := \boldsymbol{A}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$. Denote the covariance matrix of $\boldsymbol{AX}$ by $\boldsymbol{B} = (b_{i,j})$ and define the test statistic

$$M_{\boldsymbol{A}} = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \le i \le p} \frac{(\delta_i^{\boldsymbol{A}})^2}{b_{i,i}}. \tag{3}$$

The most natural choices of $\boldsymbol{A}$ are arguably $\boldsymbol{A} = \boldsymbol{\Omega}^{\frac{1}{2}}$ and $\boldsymbol{A} = \boldsymbol{I}$. In the case of $\boldsymbol{A} = \boldsymbol{\Omega}^{\frac{1}{2}}$, the components of $\boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{X}$ and $\boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{Y}$ are independent. Set $\bar{\boldsymbol{W}} = (\bar{W}_1, ..., \bar{W}_p)^T := \boldsymbol{\Omega}^{\frac{1}{2}}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$. It is natural to consider the test statistic

$$M_{\boldsymbol{\Omega}^{\frac{1}{2}}} = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \le i \le p} \bar{W}_i^2. \tag{4}$$

As we will show later that the test based on $M_{\boldsymbol{\Omega}}$ uniformly outperforms the test based on $M_{\boldsymbol{\Omega}^{\frac{1}{2}}}$ for testing against sparse alternatives.

Another natural choice is $\boldsymbol{A} = \boldsymbol{I}$. That is, the test is directly based on the difference of the sample means $\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}$. Set $\bar{\boldsymbol{\delta}} = (\bar{\delta}_1, ..., \bar{\delta}_p)' := \bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}$ and define the test statistic

$$M_{\boldsymbol{I}} = \frac{n_1 n_2}{n_1 + n_2} \max_{1 \le i \le p} \frac{\bar{\delta}_i^2}{\sigma_{i,i}} \tag{5}$$

where $\sigma_{i,i}$ are the diagonal elements of $\boldsymbol{\Sigma}$. Here $M_{\boldsymbol{I}}$ is the maximum of the squared two sample t statistics based on the samples $\{\boldsymbol{X}_k\}$ and $\{\boldsymbol{Y}_k\}$ directly. It will be shown that the

test based on $M_{\boldsymbol{I}}$ is uniformly outperformed by the test based on $M_{\boldsymbol{\Omega}}$ for testing against sparse alternatives.

2.2 Data-Driven Procedure

We have so far focused on the oracle case in which the precision matrix $\boldsymbol{\Omega}$ is known. However, in most applications $\boldsymbol{\Omega}$ is unknown and thus needs to be estimated. We consider in this paper two procedures for estimating the precision matrix. When $\boldsymbol{\Omega}$ is known to be sparse, the CLIME estimator proposed in Cai, Liu and Luo (2011) is used to estimate $\boldsymbol{\Omega}$ directly. If such information is not available, we first estimate the covariance matrix $\boldsymbol{\Sigma}$ by the inverse of the adaptive thresholding estimator $\widehat{\boldsymbol{\Sigma}}^{\star}$ introduced in Cai and Liu (2011), and then estimate $\boldsymbol{\Omega}$ by $(\widehat{\boldsymbol{\Sigma}}^{\star})^{-1}$.

We first consider the CLIME estimator. Let $\boldsymbol{\Sigma}_n$ be the pooled sample covariance matrix

$$(\hat{\sigma}_{i,j})_{p\times p} = \boldsymbol{\Sigma}_n = \frac{1}{n_1+n_2}\Big\{\sum_{k=1}^{n_1}(\boldsymbol{X}_k-\bar{\boldsymbol{X}})(\boldsymbol{X}_k-\bar{\boldsymbol{X}})' + \sum_{k=1}^{n_2}(\boldsymbol{Y}_k-\bar{\boldsymbol{Y}})(\boldsymbol{Y}_k-\bar{\boldsymbol{Y}})'\Big\}.$$

Let $\widehat{\boldsymbol{\Omega}}_1 = (\hat{\omega}^1_{i,j})$ be a solution of the following optimization problem:

$$\min \|\boldsymbol{\Omega}\|_1 \quad \text{subject to} \quad |\boldsymbol{\Sigma}_n\boldsymbol{\Omega} - \boldsymbol{I}|_\infty \le \lambda_n,$$

where $\|\cdot\|_1$ is the elementwise ℓ_1 norm, and $\lambda_n = C\sqrt{\log p/n}$ for some sufficiently large constant C. In practice, λ_n can be chosen through cross validation. See Cai, Liu, and Luo (2011) for further details. The estimator of the precision matrix $\boldsymbol{\Omega}$ is defined to be $\widehat{\boldsymbol{\Omega}} = (\hat{\omega}_{i,j})_{p\times p}$, where

$$\hat{\omega}_{i,j} = \hat{\omega}_{j,i} = \hat{\omega}^1_{i,j} I\{|\hat{\omega}^1_{i,j}| \le |\hat{\omega}^1_{j,i}|\} + \hat{\omega}^1_{j,i} I\{|\hat{\omega}^1_{i,j}| > |\hat{\omega}^1_{j,i}|\}.$$

The estimator $\widehat{\boldsymbol{\Omega}}$ is called the CLIME estimator and can be implemented by linear programming. It enjoys desirable theoretical and numerical properties. See Cai, Liu, and Luo (2011) for more details on the properties and implementation of this estimator.

When the precision matrix $\boldsymbol{\Omega}$ is not known to be sparse, we estimate $\boldsymbol{\Omega}$ by $\hat{\boldsymbol{\Omega}} = (\widehat{\boldsymbol{\Sigma}}^{\star})^{-1}$, the inverse of the adaptive thresholding estimator of $\boldsymbol{\Sigma}$. The adaptive thresholding esti-

mator $\widehat{\mathbf{\Sigma}}^\star = (\hat{\sigma}^\star_{i,j})_{p\times p}$ is defined by

$$\hat{\sigma}^\star_{i,j} = \hat{\sigma}_{i,j} I(|\hat{\sigma}_{i,j}| \geq \lambda_{i,j})$$

with $\lambda_{i,j} = \delta\sqrt{\frac{\hat{\theta}_{i,j}\log p}{n}}$, where

$$\begin{aligned}
\hat{\theta}_{i,j} &= \frac{1}{n_1+n_2}\Big\{\sum_{k=1}^{n_1}\Big[(X_{ki}-\bar{X}^i)(X_{kj}-\bar{X}^j)-\hat{\sigma}_{i,j}\Big]^2 \\
&\qquad + \sum_{k=1}^{n_2}\Big[(Y_{ki}-\bar{Y}^i)(Y_{kj}-\bar{Y}^j)-\hat{\sigma}_{i,j}\Big]^2\Big\} \\
\bar{X}^i &= n_1^{-1}\sum_{k=1}^{n_1} X_{ki},\ \bar{Y}^i = n_2^{-1}\sum_{k=1}^{n_2} Y_{ki}
\end{aligned}$$

is an estimate of $\theta_{i,j} = \mathsf{Var}((X_i-\mu_i)(X_j-\mu_j))$. Here δ is a tuning parameter which can be taken as fixed at $\delta = 2$ or can be chosen empirically through cross-validation. This estimator is easy to implement and it enjoys desirable theoretical and numerical properties. See Cai and Liu (2011) for more details on the properties of this estimator.

For testing the hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ in the case of unknown precision matrix $\mathbf{\Omega}$, motivated by the oracle procedure $M_{\mathbf{\Omega}}$ given in Section 2.1, our final test statistic is $M_{\widehat{\mathbf{\Omega}}}$ defined by

$$M_{\widehat{\mathbf{\Omega}}} = \frac{n_1 n_2}{n_1+n_2}\max_{1\leq i\leq p}\frac{\widehat{Z}_i^2}{\hat{\omega}^{(0)}_{i,i}}, \tag{6}$$

where $\widehat{\boldsymbol{Z}} = (\widehat{Z}_1,\ldots,\widehat{Z}_p)^T := \widehat{\mathbf{\Omega}}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})$ and $\hat{\omega}^{(0)}_{i,i} = \frac{n_1}{n_1+n_2}\hat{\omega}^{(1)}_{i,i} + \frac{n_2}{n_1+n_2}\hat{\omega}^{(2)}_{i,i}$ with

$$\begin{aligned}
(\hat{\omega}^{(1)}_{i,j}) &:= \frac{1}{n_1}\sum_{k=1}^{n_1}(\widehat{\mathbf{\Omega}}\boldsymbol{X}_k - \bar{\boldsymbol{X}}_{\widehat{\mathbf{\Omega}}})(\widehat{\mathbf{\Omega}}\boldsymbol{X}_k - \bar{\boldsymbol{X}}_{\widehat{\mathbf{\Omega}}})^T, \\
(\hat{\omega}^{(2)}_{i,j}) &:= \frac{1}{n_2}\sum_{k=1}^{n_2}(\widehat{\mathbf{\Omega}}\boldsymbol{Y}_k - \bar{\boldsymbol{Y}}_{\widehat{\mathbf{\Omega}}})(\widehat{\mathbf{\Omega}}\boldsymbol{Y}_k - \bar{\boldsymbol{Y}}_{\widehat{\mathbf{\Omega}}})^T, \\
\bar{\boldsymbol{X}}_{\widehat{\mathbf{\Omega}}} &= n_1^{-1}\sum_{k=1}^{n_1}\widehat{\mathbf{\Omega}}\boldsymbol{X}_k, \quad \bar{\boldsymbol{Y}}_{\widehat{\mathbf{\Omega}}} = n_2^{-1}\sum_{k=1}^{n_2}\widehat{\mathbf{\Omega}}\boldsymbol{Y}_k.
\end{aligned} \tag{7}$$

It will be shown in Section 3 that $M_{\widehat{\mathbf{\Omega}}}$ and $M_{\mathbf{\Omega}}$ have the same asymptotic null distribution and power under certain regularity conditions. Note that other estimators of the precision matrix $\mathbf{\Omega}$ can also be used to construct a good test. See more discussions in Section 3.2.2.

Remark 1. The CLIME estimator $\widehat{\boldsymbol{\Omega}}$ is positive definite with high probability when $\lambda_{\min}(\boldsymbol{\Omega}) > c > 0$. However, for a given realization, $\hat{\boldsymbol{\Omega}}$ is not guaranteed to be positive definite. The testing procedure still works even when $\hat{\boldsymbol{\Omega}}$ is not positive definite as the procedure uses $\hat{\boldsymbol{\Omega}}$ directly. If a positive semi-definite or positive definite estimator is still desired, the following simple additional step leads to an estimator of $\boldsymbol{\Omega}$ which is positive definite and achieves the same rate of convergence.

Write the eigen-decomposition of $\widehat{\boldsymbol{\Omega}}$ as $\widehat{\boldsymbol{\Omega}} = \sum_{i=1}^{p} \hat{\lambda}_i v_i v_i^T$, where $\hat{\lambda}_i$'s and v_i's are, respectively, the eigenvalues and eigenvectors of $\widehat{\boldsymbol{\Omega}}$. Set $\hat{\lambda}_i^* = \max(\hat{\lambda}_i, 0 \log p/n)$ and define $\widehat{\boldsymbol{\Omega}}^+ = \sum_{i=1}^{p} \hat{\lambda}_i^* v_i v_i^T$. Then $\widehat{\boldsymbol{\Omega}}^+$ is positive definite and attains the same rate of convergence. This method can also be applied to the adaptive thresholding estimator $\hat{\boldsymbol{\Sigma}}^\star$ to ensure the positive definiteness of the estimator. See, for example, Cai and Zhou (2011) and Cai and Yuan (2012). All the results in the present paper hold with the estimator $\widehat{\boldsymbol{\Omega}}$ replaced by $\widehat{\boldsymbol{\Omega}}^+$.

3 Theoretical Analysis

We now turn to the analysis of the properties of $M_{\boldsymbol{\Omega}}$ and $M_{\widehat{\boldsymbol{\Omega}}}$ including the limiting null distribution and the power of the corresponding tests. It is shown that the test based on $M_{\widehat{\boldsymbol{\Omega}}}$ performs as well as that based on $M_{\boldsymbol{\Omega}}$ and enjoys certain optimality under regularity conditions. The asymptotic null distributions of $M_{\boldsymbol{\Omega}^{\frac{1}{2}}}$ and $M_{\boldsymbol{I}}$ are also derived and the power of the corresponding tests is studied.

3.1 Asymptotic Distributions of the Oracle Test Statistics

We first establish the asymptotic null distributions for the oracle test statistics $M_{\boldsymbol{\Omega}}$, $M_{\boldsymbol{\Omega}^{\frac{1}{2}}}$ and $M_{\boldsymbol{I}}$. Let $\boldsymbol{D}_1 = \mathrm{diag}(\sigma_{1,1}, ..., \sigma_{p,p})$ and $\boldsymbol{D}_2 = \mathrm{diag}(\omega_{1,1}, ..., \omega_{p,p})$, where $\sigma_{k,k}$ and $\omega_{k,k}$ are the diagonal entries of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Omega}$ respectively. The correlation matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is then $\boldsymbol{\Gamma} = (\gamma_{i,j}) = \boldsymbol{D}_1^{-1/2} \boldsymbol{\Sigma} \boldsymbol{D}_1^{-1/2}$ and the correlation matrix of $\boldsymbol{\Omega X}$ and $\boldsymbol{\Omega Y}$ is $\boldsymbol{R} = (r_{i,j}) = \boldsymbol{D}_2^{-1/2} \boldsymbol{\Omega} \boldsymbol{D}_2^{-1/2}$. To obtain the limiting null distributions, we assume that the eigenvalues

of the covariance matrix $\mathbf{\Sigma}$ are bounded from above and below, and the correlations in $\mathbf{\Gamma}$ and $\boldsymbol{R}$ are bounded away from -1 and 1. More specifically we assume the following:

(C1): $C_0^{-1} \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq C_0$ for some constant $C_0 > 0$;
(C2): $\max_{1\leq i<j\leq p} |\gamma_{i,j}| \leq r_1 < 1$ for some constant $0 < r_1 < 1$;
(C3): $\max_{1\leq i<j\leq p} |r_{i,j}| \leq r_2 < 1$ for some constant $0 < r_2 < 1$.

Condition (C1) on the eigenvalues is a common assumption in the high-dimensional setting. Conditions (C2) and (C3) are also mild. For example, if $\max_{1\leq i<j\leq p} |r_{i,j}| = 1$, then $\mathbf{\Sigma}$ is singular. The following theorem states the asymptotic null distributions for the three oracle statistics $M_{\mathbf{\Omega}}$, $M_{\mathbf{\Omega}^{\frac{1}{2}}}$ and $M_{\boldsymbol{I}}$.

Theorem 1. *Let the test statistics $M_{\mathbf{\Omega}}$, $M_{\mathbf{\Omega}^{\frac{1}{2}}}$ and $M_{\boldsymbol{I}}$ be defined as in* (2), (4) *and* (5), *respectively.*

(i). Suppose that (C1) and (C3) hold. Then for any $x \in \mathbb{R}$,

$$P_{H_0}\Big(M_{\mathbf{\Omega}} - 2\log p + \log\log p \leq x\Big) \to \exp\Big(-\frac{1}{\sqrt{\pi}}\exp(-\frac{x}{2})\Big), \quad \text{as } p \to \infty.$$

(ii). For any $x \in \mathbb{R}$,

$$P_{H_0}\Big(M_{\mathbf{\Omega}^{\frac{1}{2}}} - 2\log p + \log\log p \leq x\Big) \to \exp\Big(-\frac{1}{\sqrt{\pi}}\exp(-\frac{x}{2})\Big), \quad \text{as } p \to \infty.$$

(iii). Suppose that (C1) and (C2) hold. Then for any $x \in \mathbb{R}$,

$$P_{H_0}\Big(M_{\boldsymbol{I}} - 2\log p + \log\log p \leq x\Big) \to \exp\Big(-\frac{1}{\sqrt{\pi}}\exp(-\frac{x}{2})\Big), \quad \text{as } p \to \infty.$$

Theorem 1 holds for any fixed sample sizes n_1 and n_2 and it shows that $M_{\mathbf{\Omega}}$, $M_{\mathbf{\Omega}^{\frac{1}{2}}}$ and $M_{\boldsymbol{I}}$ have the same asymptotic null distribution. Based on the limiting null distribution, three asymptotically α-level tests can be defined as follows:

$$\begin{aligned}
\Phi_\alpha(\mathbf{\Omega}) &= I\{M_{\mathbf{\Omega}} \geq 2\log p - \log\log p + q_\alpha\},\\
\Phi_\alpha(\mathbf{\Omega}^{\frac{1}{2}}) &= I\{M_{\mathbf{\Omega}^{\frac{1}{2}}} \geq 2\log p - \log\log p + q_\alpha\},\\
\Phi_\alpha(\boldsymbol{I}) &= I\{M_{\boldsymbol{I}} \geq 2\log p - \log\log p + q_\alpha\},
\end{aligned}$$

where q_α is the $1-\alpha$ quantile of the type I extreme value distribution with the cumulative distribution function $\exp\Big(-\frac{1}{\sqrt{\pi}}\exp(-x/2)\Big)$, i.e.,

$$q_\alpha = -\log(\pi) - 2\log\log(1-\alpha)^{-1}.$$

The null hypothesis H_0 is rejected if and only if $\Phi_\alpha(\cdot) = 1$. Although the asymptotic null distribution of the test statistics $M_{\boldsymbol{\Omega}}$, $M_{\boldsymbol{I}}$, and $M_{\boldsymbol{\Omega}^{\frac{1}{2}}}$ are the same, the power of the tests $\Phi_\alpha(\boldsymbol{\Omega})$, $\Phi_\alpha(\boldsymbol{\Omega}^{\frac{1}{2}})$, and $\Phi_\alpha(\boldsymbol{I})$ are quite different. We shall show in Section 1 in the supplementary material Cai, Liu and Xia (2013b) that the power of $\Phi_\alpha(\boldsymbol{\Omega})$ uniformly dominates those of $\Phi_\alpha(\boldsymbol{\Omega}^{\frac{1}{2}})$ and $\Phi_\alpha(\boldsymbol{I})$ when testing against sparse alternatives, and the results are briefly summarized in Section 3.2.3.

3.2 Asymptotic Properties of $\Phi_\alpha(\boldsymbol{\Omega})$ And $\Phi_\alpha(\widehat{\boldsymbol{\Omega}})$

In this section, the asymptotic power of $M_{\boldsymbol{\Omega}}$ is analyzed and the test $\Phi_\alpha(\boldsymbol{\Omega})$ is shown to be minimax rate optimal. In practice, $\boldsymbol{\Omega}$ is unknown and the test statistic $M_{\widehat{\boldsymbol{\Omega}}}$ should be used instead of $M_{\boldsymbol{\Omega}}$. Define the set of k_p-sparse vectors by

$$\mathcal{S}(k_p) = \Big\{\boldsymbol{\delta} : \sum_{j=1}^{p} I\{\delta_j \neq 0\} = k_p\Big\}.$$

Throughout the section, we analyze the power of $M_{\boldsymbol{\Omega}}$ and $M_{\widehat{\boldsymbol{\Omega}}}$ under the alternative

$$\begin{aligned} H_1: \quad & \boldsymbol{\delta} \in \mathcal{S}(k_p) \text{ with } k_p = p^r, 0 \le r < 1, \text{ and the nonzero} \\ & \text{locations are randomly uniformly drawn from } \{1, ..., p\}, \end{aligned}$$

Under H_1, we let $(\boldsymbol{X} - \boldsymbol{\mu}_1, \boldsymbol{Y} - \boldsymbol{\mu}_2)$ be independent with the nonzero locations of $\boldsymbol{\delta}$. As discussed in the introduction, the condition on the nonzero coordinates in H_1 is mild. Similar conditions have been imposed in Hall and Jin (2008), Hall and Jin (2010) and Arias-Castro, Candès and Plan (2011). We show that, under certain sparsity assumptions on $\boldsymbol{\Omega}$, $M_{\widehat{\boldsymbol{\Omega}}}$ performs as well as $M_{\boldsymbol{\Omega}}$ asymptotically. For the following sections, we assume $n_1 \asymp n_2$ and write $n = \frac{n_1 n_2}{n_1+n_2}$.

3.2.1 Asymptotic Power And Optimality of $\Phi_\alpha(\mathbf{\Omega})$

The asymptotic power of $\Phi_\alpha(\mathbf{\Omega})$ is analyzed under certain conditions on the separation between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. Furthermore, a lower bound is derived to show that this condition is minimax rate optimal in order to distinguish H_1 and H_0 with probability tending to 1.

Theorem 2. *Suppose* $(C1)$ *holds. Under the alternative* H_1 *with* $r < 1/4$, *if* $\max_i |\delta_i/\sigma_{i,i}^{\frac{1}{2}}| \geq \sqrt{2\beta \log p/n}$ *with* $\beta \geq 1/(\min_i \sigma_{i,i}\omega_{i,i}) + \varepsilon$ *for some constant* $\varepsilon > 0$, *then as* $p \to \infty$

$$P_{H_1}\Big(\Phi_\alpha(\mathbf{\Omega}) = 1\Big) \to 1.$$

We shall show that the condition $\max_i |\delta_i/\sigma_{i,i}^{\frac{1}{2}}| \geq \sqrt{2\beta \log p/n}$ is minimax rate optimal for testing against sparse alternatives. First we introduce some conditions.

(C4) $k_p = p^r$ for some $r < 1/2$ and $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ is s_p-sparse with $s_p = O((p/k_p^2)^\gamma)$ for some $0 < \gamma < 1$.

(C4′) $k_p = p^r$ for some $r < 1/4$.

(C5) $\|\mathbf{\Omega}\|_{L_1} \leq M$ for some constant $M > 0$.

Define the class of α-level tests by

$$\mathcal{T}_\alpha = \{\Phi_\alpha : \ \mathsf{P}_{H_0}\Big(\Phi_\alpha = 1\Big) \leq \alpha\}.$$

The following theorem shows that the condition $\max_i |\delta_i/\sigma_{i,i}^{\frac{1}{2}}| \geq \sqrt{2\beta \log p/n}$ is minimax rate optimal.

Theorem 3. *Assume that (C4) (or (C4′)) and (C5) hold. Let* $\alpha, \nu > 0$ *and* $\alpha + \nu < 1$. *Then there exists a positive constant* c *such that for all sufficiently large* n *and* p,

$$\inf_{\boldsymbol{\delta} \in \mathcal{S}(k_p) \cap \{|\boldsymbol{\delta}|_\infty \geq c\sqrt{\log p/n}\}} \sup_{\Phi_\alpha \in \mathcal{T}_\alpha} P(\Phi_\alpha = 1) \leq 1 - \nu.$$

Theorem 3 shows that, if c is sufficiently small, then any α level test is unable to reject the null hypothesis correctly uniformly over $\boldsymbol{\delta} \in \mathcal{S}(k_p) \cap \{|\boldsymbol{\delta}|_\infty \geq c\sqrt{\log p/n}\}$ with probability tending to one. So the order of the lower bound $\max_i |\delta_i/\sigma_{i,i}^{\frac{1}{2}}| \geq \sqrt{2\beta \log p/n}$ can not be improved.

3.2.2 Asymptotic Properties And Optimality of $\Phi_\alpha(\widehat{\mathbf{\Omega}})$

We now analyze the properties of $M_{\widehat{\mathbf{\Omega}}}$ and the corresponding test including the limiting null distribution and the asymptotic power. We shall assume the estimator $\widehat{\mathbf{\Omega}} = (\hat{\omega}_{i,j})$ has at least a logarithmic rate of convergence

$$\|\widehat{\mathbf{\Omega}} - \mathbf{\Omega}\|_{L_1} = o_{\mathsf{P}}(\frac{1}{\log p}) \quad \text{and} \quad \max_{1\le i\le p} |\hat{\omega}_{i,i} - \omega_{i,i}| = o_{\mathsf{P}}(\frac{1}{\log p}). \tag{8}$$

This is a rather weak requirement on $\hat{\mathbf{\Omega}}$ and, as will be shown later, can be easily satisfied by the CLIME estimator or the inverse of the adaptive thresholding estimator for a wide range of covariance/precision matrices. We will show that under Condition (8) $M_{\widehat{\mathbf{\Omega}}}$ has the same limiting null distribution as $M_{\mathbf{\Omega}}$. Define the corresponding test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ by

$$\Phi_\alpha(\widehat{\mathbf{\Omega}}) = I\{M_{\widehat{\mathbf{\Omega}}} \ge 2\log p - \log\log p + q_\alpha\}.$$

The following theorem show that $M_{\widehat{\mathbf{\Omega}}}$ and $M_{\mathbf{\Omega}}$ have the same asymptotic distribution and power under Condition (8), and so the test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ is also minimax rate optimal.

Theorem 4. *Suppose that $\widehat{\mathbf{\Omega}}$ satisfies* (8) *and (C1) and (C3) hold.*

(i). *Then under the null hypothesis H_0, for any $x \in \mathbb{R}$,*

$$\mathsf{P}_{H_0}\Big(M_{\widehat{\mathbf{\Omega}}} - 2\log p + \log\log p \le x\Big) \to \exp\Big(-\frac{1}{\sqrt{\pi}}\exp(-\frac{x}{2})\Big), \qquad \text{as } n, p \to \infty.$$

(ii). *Under the alternative hypothesis H_1 with $r < 1/6$, we have, as $n, p \to \infty$,*

$$\frac{\mathsf{P}_{H_1}\Big(\Phi_\alpha(\widehat{\mathbf{\Omega}}) = 1\Big)}{\mathsf{P}_{H_1}\Big(\Phi_\alpha(\mathbf{\Omega}) = 1\Big)} \to 1.$$

Furthermore, if $\max_i |\delta_i/\sigma_{i,i}^{\frac{1}{2}}| \ge \sqrt{2\beta\log p/n}$ with $\beta \ge 1/(\min_i \sigma_{i,i}\omega_{i,i}) + \varepsilon$ for some constant $\varepsilon > 0$, then

$$\mathsf{P}_{H_1}\Big(\Phi_\alpha(\widehat{\mathbf{\Omega}}) = 1\Big) \to 1, \qquad \text{as } n, p \to \infty.$$

As mentioned earlier, Condition (8) is rather weak and is satisfied by the CLIME estimator or the inverse of the adaptive thresholding estimator for a wide range of precision/covariance matrices. It is helpful to give a few examples of collections of precision/covariance matrices for which (8) holds.

We first consider the following class of precision matrices that satisfy an ℓ_q-ball constraint for each row/column. Let $0 \le q < 1$ and define

$$\mathcal{U}_q(s_{p,1}, M_p) = \Big\{ \mathbf{\Omega} \succ 0 : \|\mathbf{\Omega}\|_{L_1} \le M_p, \ \max_{1\le j\le p} \sum_{i=1}^{p} |\omega_{i,j}|^q \le s_{p,1} \Big\}. \tag{9}$$

The class $\mathcal{U}_q(s_{p,1}, M_p)$ covers a range of precision matrices as the parameters q, $s_{p,1}$ and M_p vary. Using the techniques in Cai, Liu and Luo (2011), Proposition 1 below shows that (8) holds for the CLIME estimator if $\mathbf{\Omega} \in \mathcal{U}_q(s_p, M_p)$ with

$$s_{p,1} = o\Big(\frac{n^{(1-q)/2}}{M_p^{1-q} (\log p)^{(3-q)/2}} \Big). \tag{10}$$

We now turn to the covariance matrices. Consider a large class of covariance matrices defined by, for $0 \le q < 1$,

$$\mathcal{U}_q^\star(s_{p,2}, M_p) = \Big\{ \mathbf{\Sigma} : \mathbf{\Sigma} \succ 0, \|\mathbf{\Sigma}^{-1}\|_{L_1} \le M_p, \ \max_i \sum_{j=1}^{p} (\sigma_{i,i}\sigma_{j,j})^{(1-q)/2} |\sigma_{i,j}|^q \le s_{p,2} \Big\}. \tag{11}$$

Matrices in $\mathcal{U}_q^\star(s_{p,2})$ satisfy a weighted ℓ_q-ball constraint for each row/column. Let $\hat{\mathbf{\Omega}} = (\hat{\mathbf{\Sigma}}^\star)^{-1}$, where $\hat{\mathbf{\Sigma}}^\star$ is the adaptive thresholding estimator defined in Section 2.2. Then Proposition 1 below shows that Condition (8) is satisfied by $\hat{\mathbf{\Omega}}$ if $\mathbf{\Sigma} \in \mathcal{U}_q^\star(s_{p,2}, M_p)$ with

$$s_{p,2} = o\Big(\frac{n^{(1-q)/2}}{M_p^{2} (\log p)^{(3-q)/2}} \Big). \tag{12}$$

Besides the class of covariance matrices $\mathcal{U}_q^\star(s_{p,2}, M_p)$ given in (11), Condition (8) is also satisfied by the inverse of the adaptive thresholding estimator $\hat{\mathbf{\Omega}} = (\hat{\mathbf{\Sigma}}^\star)^{-1}$ over the class of bandable covariance matrices defined by

$$\mathcal{F}_\alpha(M_1, M_2) = \left\{ \Sigma : \mathbf{\Sigma} \succ 0, \|\mathbf{\Sigma}^{-1}\|_{L_1} \le M_1, \ \max_j \sum_{i:|i-j|>k} |\sigma_{i,j}| \le M_2 k^{-\alpha}, \text{ for } k \ge 1 \right\}$$

where $\alpha > 0$, $M_1 > 0$ and $M_2 > 0$. This class of covariance matrices arises naturally in time series analysis. See Cai, Zhang and Zhou (2010) and Cai and Zhou (2012).

Proposition 1. *Suppose that* $\log p = o(n^{1/3})$ *and (C1) holds. Then the CLIME estimator for* $\mathbf{\Omega} \in \mathcal{U}_q(s_{p,1}, M_p)$ *with (10) satisfies* (8)*. Similarly, the inverse of the adaptive thresholding estimator of* $\mathbf{\Sigma} \in \mathcal{U}_q^\star(s_{p,2}, M_p)$ *with (12) or* $\mathbf{\Sigma} \in \mathcal{F}_\alpha(M_1, M_2)$ *with* $\log p = o(n^{\alpha/(4+3\alpha)})$ *satisfies* (8)*.*

We should note that the conditions (8) , (10), and (12) are technical conditions and they can be further weakened. For example, the following result holds without imposing a sparsity condition on $\mathbf{\Omega}$.

Theorem 5. *Let* $\hat{\mathbf{\Omega}}$ *be the CLIME estimator. Suppose that (C1) and (C3) hold and* $\min_i \omega_{i,i} \geq c$ *for some* $c > 0$*. If* $\|\mathbf{\Omega}\|_{L_1} \leq M_p$ *and*

$$M_p^2 = o(\sqrt{n}/(\log p)^{3/2}), \tag{13}$$

then

$$P_{H_0}(\Phi_\alpha(\hat{\mathbf{\Omega}}) = 1) \leq \alpha + o(1), \qquad \text{as } n,\, p \to \infty.$$

Furthermore, if $\max_i |\delta_i/\sigma_{i,i}^{\frac{1}{2}}| \geq \sqrt{2\beta \log p/n}$ *with* $\beta \geq 1/(\min_i \sigma_{i,i}\omega_{i,i}) + \varepsilon$ *for some constant* $\varepsilon > 0$*, then*

$$P_{H_1}(\Phi_\alpha(\hat{\mathbf{\Omega}}) = 1) \to 1, \qquad \text{as } n,\, p \to \infty.$$

3.2.3 Power Comparison of the Oracle Tests

The tests $\Phi_\alpha(\mathbf{\Omega})$ and $\Phi_\alpha(\hat{\mathbf{\Omega}})$ are shown in Sections 3.2.1 and 3.2.2 to be minimax rate optimal for testing against sparse alternatives. Under some additional regularity conditions, it can be shown that the test $\Phi_\alpha(\mathbf{\Omega})$ is uniformly at least as powerful as both $\Phi_\alpha(\mathbf{\Omega}^{\frac{1}{2}})$ and $\Phi_\alpha(\boldsymbol{I})$, and the results are stated in Proposition 1 in the supplementary material Cai, Liu and Xia (2013b). Furthermore, we show that, for a class of alternatives, the test $\Phi_\alpha(\mathbf{\Omega})$ is strictly more powerful than both $\Phi_\alpha(\mathbf{\Omega}^{\frac{1}{2}})$ and $\Phi_\alpha(\boldsymbol{I})$. For further details, see Propositions 2 and 3 in the supplementary material Cai, Liu and Xia (2013b). On the other hand, we should also note that the relative performance of the three oracle tests, $\Phi_\alpha(\mathbf{\Omega})$, $\Phi_\alpha(\mathbf{\Omega}^{\frac{1}{2}})$, and $\Phi_\alpha(\boldsymbol{I})$, is not clear in the non-sparse case. It is possible, for example, when $k_p = p^r$ for some $r > 1/2$, $\Phi_\alpha(\mathbf{\Omega})$ might be outperformed by $\Phi_\alpha(\mathbf{\Omega}^{\frac{1}{2}})$ or $\Phi_\alpha(\boldsymbol{I})$.

4 Extension to Non-Gaussian Distributions

We have so far focused on the Gaussian setting and studied the asymptotic null distributions and power of the tests. In this section, the results for the tests $\Phi_\alpha(\mathbf{\Omega})$ and $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ are extended to non-Gaussian distributions.

We require some moment conditions on the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two p-dimensional random vectors satisfying

$$\boldsymbol{X} = \boldsymbol{\mu}_1 + \boldsymbol{U}_1 \text{ and } \boldsymbol{Y} = \boldsymbol{\mu}_2 + \boldsymbol{U}_2,$$

where $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are independent and identical distributed random vectors with mean zero and covariance matrix $\mathbf{\Sigma} = (\sigma_{i,j})_{p\times p}$. Let $\boldsymbol{V}_j = \mathbf{\Omega}\boldsymbol{U}_j =: (V_{1j}, \ldots, V_{pj})^T$ for $j = 1, 2$. The moment conditions are divided into two cases: the sub-Gaussian-type tails and polynomial-type tails.

(C6). (Sub-Gaussian-type tails) Suppose that $\log p = o(n^{1/4})$. There exist some constants $\eta > 0$ and $K > 0$ such that

$$\mathsf{E}\exp(\eta V_{i1}^2/\omega_{i,i}) \le K \text{ and } \mathsf{E}\exp(\eta V_{i2}^2/\omega_{i,i}) \le K \text{ for } 1 \le i \le p.$$

(C7). (Polynomial-type tails) Suppose that for some constants $\gamma_0, c_1 > 0$, $p \le c_1 n^{\gamma_0}$, and for some constants $\epsilon > 0$ and $K > 0$

$$\mathsf{E}|V_{i1}/\omega_{i,i}^{\frac{1}{2}}|^{2\gamma_0+2+\epsilon} \le K \text{ and } \mathsf{E}|V_{i2}/\omega_{i,i}^{\frac{1}{2}}|^{2\gamma_0+2+\epsilon} \le K \text{ for } 1 \le i \le p.$$

Theorem 6. *Suppose that (C1), (C3) and (C6) (or (C7)) hold. Then under the null hypothesis H_0, for any $x \in \mathbb{R}$,*

$$P_{H_0}\Big(M_{\mathbf{\Omega}} - 2\log p + \log\log p \le x\Big) \to \exp\Big(-\frac{1}{\sqrt{\pi}}\exp(-\frac{x}{2})\Big), \quad \text{as } n, p \to \infty.$$

Theorem 6 shows that $\Phi_\alpha(\mathbf{\Omega})$ is still an asymptotically α-level test when the distribution is non-Gaussian. When $\mathbf{\Omega}$ is unknown, as in the Gaussian case, the CLIME estimator $\widehat{\mathbf{\Omega}}$ in Cai, Liu and Luo (2011) or the inverse of adaptive thresholding estimator $(\hat{\mathbf{\Sigma}}^\star)^{-1}$ can be used. The following theorem shows that the test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ shares the same optimality as $\Phi_\alpha(\mathbf{\Omega})$ in the non-Gaussian setting.

Theorem 7. *Suppose that (C1), (C3), (C6) (or (C7)) and (8) hold.*

(i). Under the null hypothesis H_0, for any $x \in \mathbb{R}$,

$$P_{H_0}\Big(M_{\widehat{\mathbf{\Omega}}} - 2\log p + \log\log p \le x\Big) \to \exp\Big(-\frac{1}{\sqrt{\pi}}\exp(-\frac{x}{2})\Big), \quad \text{as } n, p \to \infty.$$

(ii). Under H_1 and the conditions of Theorem 2, we have

$$P_{H_1}\Big(\Phi_\alpha(\widehat{\mathbf{\Omega}}) = 1\Big) \to 1, \quad \text{as } n, p \to \infty.$$

5 Simulation Study

In this section, we consider the numerical performance of the proposed test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ and compare it with a number of other tests, including the tests based on the sum of squares type statistics in Bai and Saranadasa (1996), Srivastava and Du (2009), Chen and Qin (2010), and the commonly used Hotelling's T^2 test. These last four tests are denoted respectively by BS, SD, CQ and T^2 respectively in the rest of this section.

The test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ is easy to implement. A range of covariance structures are considered in the simulation study, including the settings where the covariance matrix $\mathbf{\Sigma}$ is sparse, the precision matrix Ω is sparse, and both $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ are non-sparse. In the case when $\mathbf{\Omega}$ is known to be sparse, the CLIME estimator in Cai, Liu and Luo (2011) is used to estimate it, while the inverse of the adaptive thresholding estimator in Cai and Liu (2011) is used to estimate $\mathbf{\Omega}$ when such information is not available. The simulation results show that the test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ significantly and uniformly outperforms the other four tests when either Σ or Ω is sparse, and the test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ still outperforms the other four tests even when both $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ are non-sparse.

Without loss of generality, we shall always take $\boldsymbol{\mu}_2 = \mathbf{0}$ in the simulations. Under the null hypothesis, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \mathbf{0}$, while under the alternative hypothesis, we take $\boldsymbol{\mu}_1 = (\mu_{11}, \ldots, \mu_{1p})'$ to have m nonzero entries with the support $S = \{l_1, ..., l_m : l_1 < l_2 < \cdots < l_m\}$ uniformly and randomly drawn from $\{1, ..., p\}$. Two values of m are considered: $m = \lfloor 0.05p \rfloor$ and $m = \lfloor \sqrt{p} \rfloor$. Here $\lfloor x \rfloor$ denote the largest integer that is no greater than

x. For each of these two values of m, and for any $l_j \in S$, two settings of the magnitude of μ_{1,l_j} are considered: $\mu_{1,l_j} = \pm\sqrt{\log p/n}$ with equal probability and μ_{1,l_j} has magnitude randomly uniformly drawn from the interval $[-\sqrt{8\log p/n}, \sqrt{8\log p/n}]$. We take $\mu_{1,k} = 0$ for $k \in S^c$.

The specific models for the covariance structure are given as follows. Let $\boldsymbol{D} = (d_{i,j})$ be a diagonal matrix with diagonal elements $d_{i,i} = \text{Unif}(1,3)$ for $i = 1, ..., p$. Denote by $\lambda_{\min}(\boldsymbol{A})$ the minimum eigenvalue of a symmetric matrix $\boldsymbol{A}$. The following three models where the precision matrix $\boldsymbol{\Omega}$ is sparse are considered.

- Model 1 (Block diagonal $\boldsymbol{\Omega}$): $\boldsymbol{\Sigma} = (\sigma_{i,j})$ where $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0.8$ for $2(k-1)+1 \leq i \neq j \leq 2k$, where $k = 1, ..., [p/2]$ and $\sigma_{i,j} = 0$ otherwise.

- Model 2 ("Bandable" Σ): $\boldsymbol{\Sigma} = (\sigma_{i,j})$ where $\sigma_{i,j} = 0.6^{|i-j|}$ for $1 \leq i, j \leq p$.

- Model 3 (Banded Ω): $\boldsymbol{\Omega} = (\omega_{i,j})$ where $\omega_{i,i} = 2$ for $i = 1, ..., p$, $\omega_{i,i+1} = 0.8$ for $i = 1, ..., p-1$, $\omega_{i,i+2} = 0.4$ for $i = 1, ..., p-2$, $\omega_{i,i+3} = 0.4$ for $i = 1, ..., p-3$, $\omega_{i,i+4} = 0.2$ for $i = 1, ..., p-4$, $\omega_{i,j} = \omega_{j,i}$ for $i, j = 1, ..., p$ and $\omega_{i,j} = 0$ otherwise.

We also consider two models where the covariance matrix $\boldsymbol{\Sigma}$ is sparse.

- Model 4 (Sparse Σ): $\boldsymbol{\Omega} = (\omega_{i,j})$ where $\omega_{i,j} = 0.6^{|i-j|}$ for $1 \leq i, j \leq p$. $\boldsymbol{\Sigma} = \boldsymbol{D}^{1/2}\boldsymbol{\Omega}^{-1}\boldsymbol{D}^{1/2}$.

- Model 5 (Sparse Σ): $\boldsymbol{\Omega}^{1/2} = (a_{i,j})$ where $a_{i,i} = 1$, $a_{i,j} = 0.8$ for $2(k-1)+1 \leq i \neq j \leq 2k$, where $k = 1, ..., [p/2]$ and $a_{i,j} = 0$ otherwise. $\boldsymbol{\Omega} = \boldsymbol{D}^{1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{\Omega}^{1/2}\boldsymbol{D}^{1/2}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$.

In addition, three models where neither $\boldsymbol{\Sigma}$ nor $\boldsymbol{\Omega}$ is sparse are considered. In Model 6, $\boldsymbol{\Sigma}$ is a sparse matrix plus a perturbation of a non-sparse matrix $\boldsymbol{E}$. The entries of $\boldsymbol{\Sigma}$ in Model 7 decays as a function of the lag $|i-j|$, which arises naturally in time series analysis. Model 8 considers a non-sparse rank-3 perturbation to a sparse matrix which leads to a non-sparse covariance matrix $\boldsymbol{\Sigma}$. The simulation results show that the proposed test $\Phi_\alpha(\widehat{\boldsymbol{\Omega}})$ still significantly outperforms the other four tests under these non-sparse models.

- Model 6 (Nonsparse case): $\boldsymbol{\Sigma}^* = (\sigma^*_{i,j})$ where $\sigma^*_{i,i} = 1$, $\sigma^*_{i,j} = 0.8$ for $2(k-1)+1 \leq i \neq j \leq 2k$, where $k = 1, ..., [p/2]$ and $\sigma^*_{i,j} = 0$ otherwise. $\boldsymbol{\Sigma} = \boldsymbol{D}^{1/2}\boldsymbol{\Sigma}^*\boldsymbol{D}^{1/2} + \boldsymbol{E} + \delta\boldsymbol{I}$ with $\delta = |\lambda_{\min}(\boldsymbol{D}^{1/2}\boldsymbol{\Sigma}^*\boldsymbol{D}^{1/2} + \boldsymbol{E})| + 0.05$, where $\boldsymbol{E}$ is a symmetric matrix with the support of the off-diagonal entries chosen independently according to the Bernoulli(0.3) distribution with the values of the nonzero entries drawn randomly from Unif$(-0.2, 0.2)$.

- Model 7 (Nonsparse case): $\boldsymbol{\Sigma}^* = (\sigma^*_{i,j})$ where $\sigma^*_{i,i} = 1$ and $\sigma^*_{i,j} = |i-j|^{-5}/2$ for $i \neq j$. $\boldsymbol{\Sigma} = \boldsymbol{D}^{1/2}\boldsymbol{\Sigma}^*\boldsymbol{D}^{1/2}$.

- Model 8 (Nonsparse case): $\boldsymbol{\Sigma} = \boldsymbol{D}^{1/2}(\boldsymbol{F} + \boldsymbol{u}_1\boldsymbol{u}_1' + \boldsymbol{u}_2\boldsymbol{u}_2' + \boldsymbol{u}_3\boldsymbol{u}_3')\boldsymbol{D}^{1/2}$, where $\boldsymbol{F} = (f_{i,j})$ is a $p \times p$ matrix with $f_{i,i} = 1$, $f_{i,i+1} = f_{i+1,i} = 0.5$ and $f_{i,j} = 0$ otherwise, and $\boldsymbol{u}_i$ are orthonormal vectors for $i = 1, 2, 3$.

Under each model, two independent random samples $\{\boldsymbol{X}_k\}$ and $\{\boldsymbol{Y}_l\}$ are generated with the same sample size $n = 100$ from two multivariate normal distributions with the means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ respectively and a common covariance matrix $\boldsymbol{\Sigma}$. The dimension p takes values $p = 50, 100$ and 200. The power and significance level are calculated from 1000 replications.

The numerical results on the proposed test $\Phi_\alpha(\widehat{\boldsymbol{\Omega}})$ and BS, SD, CQ and T^2 under Models 1-5 are summarized in Tables 1 - 2. Table 1 compares the empirical sizes of the tests. It can be seen that the estimated sizes are reasonably close to the nominal level 0.05 for all the tests. Table 2, which compares the powers, shows that the new test $\Phi_\alpha(\widehat{\boldsymbol{\Omega}})$, based on either the CLIME estimator of sparse Ω or the inverse of the adaptive thresholding estimator of sparse Σ, uniformly and significantly outperforms the other four tests over all dimensions ranging from 50 to 200. The powers of these tests are significantly lower than that of $\Phi_\alpha(\widehat{\boldsymbol{\Omega}})$. These numerical results confirm the theoretical analysis given in the last section.

Table 3 summarizes the sizes and powers for the non-sparse cases. We only report here the cases when the magnitudes of the signals vary under the alternative. The performance

p	50	100	200	50	100	200	50	100	200	50	100	200	50	100	200
	Model 1			Model 2			Model 3			Model 4			Model 5		
T^2	0.04	0.06	–	0.04	0.04	–	0.04	0.05	–	0.05	0.06	–	0.05	0.05	–
BS	0.06	0.07	0.06	0.07	0.06	0.06	0.06	0.05	0.04	0.07	0.07	0.06	0.06	0.06	0.05
SD	0.06	0.07	0.06	0.07	0.06	0.06	0.06	0.05	0.02	0.07	0.07	0.06	0.06	0.06	0.05
CQ	0.06	0.07	0.06	0.06	0.06	0.07	0.06	0.05	0.02	0.07	0.07	0.06	0.06	0.06	0.05
$\Phi_\alpha(\widehat{\mathbf{\Omega}})$	0.05	0.05	0.06	0.04	0.05	0.06	0.05	0.06	0.06	0.04	0.05	0.06	0.03	0.04	0.03

Table 1: Empirical sizes based on 1000 replications with $\alpha = 0.05$ and $n = 100$.

of the tests is similar to that in the case of fixed magnitude. It can be seen from Table 3 that the sizes of the sum of square type tests tend to be larger than the nominal level 0.05 while the sizes of the new test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ is smaller than the nominal level. Thus, the new test has smaller type I error probability than those of the sum of square type tests. For the models where both Σ and Ω are non-sparse, the power of the proposed test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ is not as high as in the sparse cases. However, similar phenomena are observed in Table 3 when comparing the powers with the other tests. The tests based on the sum of squares test statistics are not powerful against the sparse alternatives, and they are still significantly outperformed by the new test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$.

More extensive simulations were carried out in the non-sparse settings as well as for non-Gaussian distributions. We also compare the proposed test with the tests based on some other estimators of the precision matrices. In particular, we consider non-sparse covariance structures by adding to the covariance/precision matrices in Models 1-5 a perturbation of a non-sparse matrix $\boldsymbol{E}$, where $\boldsymbol{E}$ is a symmetric matrix with 30% random nonzero entries drawn from Unif$(-0.2, 0.2)$. Furthermore, simulations for five additional non-sparse covariance models are carried out. The comparisons are consistent with the cases reported here. For reasons of space, these simulation results are given in the supplementary material Cai, Liu and Xia (2013b).

In summary, the numerical results show that the proposed test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ is significantly

p	50	100	200	50	100	200	50	100	200	50	100	200	50	100	200
	$m = 0.05p$ with fixed magnitude														
	Model 1			Model 2			Model 3			Model 4			Model 5		
T^2	0.22	0.41	–	0.17	0.27	–	0.14	0.26	–	0.14	0.21	–	0.34	0.39	–
BS	0.11	0.19	0.27	0.11	0.13	0.24	0.12	0.28	0.44	0.10	0.15	0.23	0.07	0.07	0.07
SD	0.11	0.19	0.27	0.11	0.13	0.24	0.12	0.27	0.44	0.10	0.15	0.27	0.07	0.07	0.07
CQ	0.11	0.19	0.27	0.11	0.13	0.24	0.12	0.27	0.44	0.10	0.15	0.23	0.07	0.07	0.07
$\Phi_\alpha(\widehat{\mathbf{\Omega}})$	0.36	0.70	0.89	0.26	0.41	0.66	0.21	0.40	0.62	0.18	0.28	0.61	0.57	0.66	1.00
	$m = \sqrt{p}$ with fixed magnitude														
	Model 1			Model 2			Model 3			Model 4			Model 5		
T^2	0.85	0.81	–	0.64	0.67	–	0.63	0.57	–	0.58	0.49	–	0.81	0.85	–
BS	0.30	0.36	0.40	0.23	0.28	0.35	0.49	0.59	0.63	0.26	0.30	0.36	0.08	0.08	0.08
SD	0.30	0.36	0.40	0.23	0.28	0.35	0.50	0.59	0.63	0.29	0.29	0.35	0.08	0.08	0.09
CQ	0.30	0.35	0.39	0.22	0.28	0.35	0.49	0.59	0.63	0.26	0.30	0.35	0.08	0.08	0.08
$\Phi_\alpha(\widehat{\mathbf{\Omega}})$	0.82	0.89	0.96	0.58	0.79	0.76	0.49	0.57	0.68	0.52	0.69	0.68	0.77	0.93	0.99
	$m = 0.05p$ with varied magnitude														
	Model 1			Model 2			Model 3			Model 4			Model 5		
T^2	0.19	0.44	–	0.08	0.45	–	0.15	0.27	–	0.13	0.14	–	0.19	0.08	–
BS	0.11	0.21	0.32	0.08	0.21	0.31	0.13	0.29	0.51	0.10	0.11	0.23	0.07	0.06	0.06
SD	0.11	0.22	0.32	0.08	0.21	0.31	0.13	0.29	0.51	0.10	0.11	0.27	0.07	0.06	0.06
CQ	0.11	0.21	0.32	0.08	0.21	0.32	0.13	0.29	0.52	0.10	0.11	0.24	0.07	0.06	0.06
$\Phi_\alpha(\widehat{\mathbf{\Omega}})$	0.35	0.76	0.94	0.12	0.85	0.88	0.24	0.51	0.84	0.22	0.30	0.76	0.44	0.11	0.40
	$m = \sqrt{p}$ with varied magnitude														
	Model 1			Model 2			Model 3			Model 4			Model 5		
T^2	0.63	0.70	–	0.72	0.75	–	0.59	0.56	–	0.29	0.74	–	0.28	0.95	–
BS	0.23	0.34	0.44	0.27	0.42	0.38	0.51	0.52	0.72	0.12	0.39	0.38	0.07	0.08	0.08
SD	0.23	0.34	0.44	0.27	0.42	0.38	0.51	0.52	0.72	0.15	0.37	0.41	0.07	0.09	0.08
CQ	0.23	0.33	0.44	0.26	0.41	0.37	0.51	0.51	0.72	0.12	0.39	0.38	0.07	0.08	0.08
$\Phi_\alpha(\widehat{\mathbf{\Omega}})$	0.71	0.90	0.97	0.80	0.91	0.87	0.54	0.70	0.92	0.47	0.96	0.94	0.30	1.00	1.00

Table 2: Powers of the tests based on 1000 replications with $\alpha = 0.05$ and $n = 100$.

and uniformly more powerful than the other four tests in the settings where either Σ or Ω is sparse. When both Σ and Ω are non-sparse, the test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ still outperforms the sum of squares type tests. Based on these numerical results, we recommend using the test $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ with the CLIME estimator of Ω when Ω is known to be sparse and using $\Phi_\alpha(\widehat{\mathbf{\Omega}})$ with the inverse of the adaptive thresholding estimator of Σ when such information is not available.

p	50	100	200	50	100	200	50	100	200
	Model 6			Model 7			Model 8		
	Size								
T^2	0.05	0.05	–	0.05	0.05	–	0.04	0.04	–
BS	0.07	0.07	0.05	0.07	0.06	0.05	0.06	0.06	0.06
SD	0.08	0.05	0.05	0.08	0.06	0.05	0.05	0.05	0.06
CQ	0.07	0.07	0.05	0.07	0.06	0.06	0.06	0.06	0.06
$\Phi_\alpha(\widehat{\mathbf{\Omega}})$	0.05	0.05	0.05	0.02	0.02	0.03	0.03	0.02	0.03
	Power when $m = 0.05p$								
T^2	0.14	0.37	–	0.17	0.43	–	0.31	0.40	–
BS	0.10	0.22	0.24	0.09	0.12	0.20	0.07	0.10	0.16
SD	0.10	0.21	0.26	0.09	0.13	0.20	0.07	0.09	0.16
CQ	0.10	0.22	0.23	0.09	0.12	0.20	0.07	0.11	0.16
$\Phi_\alpha(\widehat{\mathbf{\Omega}})$	0.16	0.40	0.46	0.14	0.41	0.80	0.20	0.50	0.84
	Power when $m = \sqrt{p}$								
T^2	0.47	0.23	–	0.49	0.59	–	0.53	1.00	–
BS	0.17	0.13	0.46	0.16	0.18	0.20	0.11	0.15	0.16
SD	0.14	0.14	0.57	0.16	0.17	0.20	0.12	0.14	0.16
CQ	0.16	0.13	0.46	0.15	0.18	0.20	0.11	0.14	0.16
$\Phi_\alpha(\widehat{\mathbf{\Omega}})$	0.24	0.26	0.77	0.37	0.57	0.53	0.38	0.73	0.85

Table 3: Empirical sizes and powers for Model 6-8 with $\alpha = 0.05$ and $n = 100$. Based on 1000 replications.

6 Discussions

In the present paper it is assumed that the two populations have the same covariance matrix. More generally, suppose we observe $\boldsymbol{X}_k \overset{iid}{\sim} N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $k = 1, ..., n_1$, and $\boldsymbol{Y}_k \overset{iid}{\sim} N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $k = 1, ..., n_2$ and wish to test $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. In order to apply the procedure proposed in this paper, one needs to first test $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ versus $H_1 : \boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$. For this purpose, for example, the test introduced in Cai, Liu and Xia (2013a) can be used. If the null hypothesis $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ is rejected, the test proposed in this paper is not directly applicable. However, a modified version of the procedure can still be used. Note that the covariance matrix of $\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}$ is $\boldsymbol{\Sigma}_1/n_1 + \boldsymbol{\Sigma}_2/n_2$. To apply the test procedure in Section 2, one needs to estimate $(\boldsymbol{\Sigma}_1 + \frac{n_1}{n_2}\boldsymbol{\Sigma}_2)^{-1}$. When both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are sparse, the inverse can be estimated well by $(\hat{\boldsymbol{\Sigma}}_{1,\text{thr}} + \frac{n_1}{n_2}\hat{\boldsymbol{\Sigma}}_{2,\text{thr}})^{-1}$ using the adaptive thresholding estimators $\hat{\boldsymbol{\Sigma}}_{1,\text{thr}}$ and $\hat{\boldsymbol{\Sigma}}_{2,\text{thr}}$ introduced in Cai and Liu (2011). Similarly, when both $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are bandable, $(\boldsymbol{\Sigma}_1 + \frac{n_1}{n_2}\boldsymbol{\Sigma}_2)^{-1}$ can also be estimated well. A more interesting problem is the estimation of $(\boldsymbol{\Sigma}_1 + \frac{n_1}{n_2}\boldsymbol{\Sigma}_2)^{-1}$ when the precision matrices $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2$ are sparse.

Besides testing the means and covariance matrices of two populations, another interesting and related problem is the testing of the equality of two distributions based on the two samples. That is, we wish to test $H_0 : \mathbb{P}_1 = \mathbb{P}_2$ versus $H_1 : \mathbb{P}_1 \neq \mathbb{P}_2$, where $\mathbb{P}_i$ is the distribution of $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. We shall report the details of the results elsewhere in the future as a significant amount of additional work is still needed.

The asymptotic properties in Section 3.2 rely on the assumption that the locations of the nonzero entries of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ are uniformly drawn from $\{1, ..., p\}$. When this assumption does not hold, the asymptotic power results may fail. A simple solution is to first apply a random permutation to the coordinates of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ (and correspondingly the coordinates of $\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}$) so that the nonzero locations are uniformly drawn from $\{1, ..., p\}$, and apply the testing procedures to the permuted data and the results given in Section 3.2 then hold.

It is well known that the convergence rate in distribution of the extreme value type statistics is slow. There are several possible ways to improve the rate of convergence. See,

for example, Hall (1991), Liu, Lin and Shao (2008) and Birnbaum and Nadler (2012). It is interesting to investigate whether these methods can be applied to improve the convergence rate of our test statistic. We leave this to future work.

7 Proof of Main Results

We prove the main results in this section. The proofs of some of the main theorems rely on a few additional technical lemmas. These technical results are collected in Section 7.1 and they are proved in the supplementary material, Cai, Liu and Xia (2013b).

7.1 Technical Lemmas

Lemma 1 (Bonferroni Inequality)**.** *Let $A = \cup_{t=1}^{p} A_t$. For any $k < [p/2]$, we have*

$$\sum_{t=1}^{2k}(-1)^{t-1}E_t \le P(A) \le \sum_{t=1}^{2k-1}(-1)^{t-1}E_t,$$

where $E_t = \sum_{1\le i_1<\cdots<i_t\le p} P(A_{i_1}\cap\cdots\cap A_{i_t})$.

Lemma 2 (Berman, 1962)**.** *If X and Y have a bivariate normal distribution with expectation zero, unit variance and correlation coefficient ρ, then*

$$\lim_{c\to\infty}\frac{P\Big(X>c, Y>c\Big)}{[2\pi(1-\rho)^{\frac{1}{2}}c^2]^{-1}\exp\Big(-\frac{c^2}{1+\rho}\Big)(1+\rho)^{\frac{1}{2}}} = 1,$$

uniformly for all ρ such that $|\rho|\le\delta$, for any δ, $0<\delta<1$.

Lemma 3. *Suppose (C1) holds and $\mathbf{\Sigma}$ has all diagonal elements equal to 1. Then for p^r-sparse $\boldsymbol{\delta}$, with $r<1/4$ and nonzero locations $l_1,...,l_m$, $m=p^r$, randomly and uniformly drawn from $\{1,...,p\}$, we have, for any $2r<a<1-2r$, as $p\to\infty$,*

$$P\Big(\max_{i\in H}\Big|\frac{(\mathbf{\Omega}\boldsymbol{\delta})_i}{\sqrt{\omega_{i,i}}}-\sqrt{\omega_{i,i}}\delta_i\Big| = O(p^{r-a/2})\max_{i\in H}|\delta_i|\Big)\to 1, \tag{14}$$

and

$$P\Big(\max_{i\in H}\Big|(\mathbf{\Omega}^{\frac{1}{2}}\boldsymbol{\delta})_i - a_{i,i}\delta_i\Big| = O(p^{r-a/2})\max_{i\in H}|\delta_i|\Big)\to 1, \tag{15}$$

where $\mathbf{\Omega}^{\frac{1}{2}} =: (a_{i,j})$ *and* H *is the support of* $\boldsymbol{\delta}$.

Lemma 4. *Let* $Y_i \sim N(\mu_i, 1)$ *be independent for* $i = 1, ..., n$. *Let* $a_n = o((\log n)^{-1/2})$. *Then*

$$\sup_{x\in\mathbb{R}} \max_{1\le k\le n} \Big| P\Big(\max_{1\le i\le k} Y_i \ge x + a_n \Big) - P\Big(\max_{1\le i\le k} Y_i \ge x \Big) \Big| = o(1) \tag{16}$$

uniformly in the means μ_i, $1 \le i \le n$. *If* Y_i *is replaced by* $|Y_i|$, *then (16) still holds.*

Lemma 5 (Baraud, 2002)**.** *Let* $\mathcal{F}$ *be some subset of* $l_2(J)$. *Let* μ_ρ *be some probability measure on* $\mathcal{F}_\rho = \{\theta \in \mathcal{F}, \|\theta\| \ge \rho\}$ *and let* $P_{\mu_\rho} = \int P_\theta d\mu_\rho(\theta)$. *Assuming that* P_{μ_ρ} *is absolutely continuous with respect to* P_0, *we define* $L_{\mu_\rho}(y) = \frac{dP_{\mu_\rho}}{dP_0}(y)$. *For all* $\alpha > 0$, $\nu \in [0, 1-\alpha]$, *if* $E_0(L^2_{\mu_{\rho^*}}(Y)) \le 1 + 4(1-\alpha-\nu)^2$, *then*

$$\forall \rho \le \rho^*, \quad \inf_{\Phi_\alpha} \sup_{\theta\in\mathcal{F}_\rho} P_\theta(\Phi_\alpha = 0) \ge \nu.$$

7.2 Proof of Theorem 1

Because we standardize the test statistic first, we shall let $(Z_1, \ldots, Z_p)'$ be a zero mean multivariate normal random vector with covariance matrix $\mathbf{\Omega} = (\omega_{i,j})_{1\le i,j\le p}$ and the diagonal $\omega_{i,i} = 1$ for $1 \le i \le p$. To prove Theorem 1, it suffices to prove the following lemma.

Lemma 6. *Suppose that* $\max_{1\le i\ne j\le p} |\omega_{i,j}| \le r < 1$ *and* $\max_j \sum_{i=1}^p \omega_{i,j}^2 \le C_0$. *Then for any* $x \in \mathbb{R}$ *as* $p \to \infty$

$$P\Big(\max_{1\le i\le p} Z_i^2 - 2\log p + \log\log p \le x \Big) \;\to\; \exp\Big(-\frac{1}{\sqrt{\pi}} \exp(-x/2) \Big), \tag{17}$$

$$P\Big(\max_{1\le i\le p} Z_i \le \sqrt{2\log p - \log\log p + x} \Big) \;\to\; \exp\Big(-\frac{1}{2\sqrt{\pi}} \exp(-x/2) \Big). \tag{18}$$

Proof. We only need to prove (17) because the proof of (18) is similar. Set $x_p = 2\log p - \log\log p + x$. By Lemma 1, we have for any fixed $k \le [p/2]$,

$$\sum_{t=1}^{2k} (-1)^{t-1} E_t \le \mathsf{P}\Big(\max_{1\le i\le p} |Z_i| \ge \sqrt{x_p} \Big) \le \sum_{t=1}^{2k-1} (-1)^{t-1} E_t, \tag{19}$$

where $E_t = \sum_{1\le i_1<\cdots<i_t\le p} \mathsf{P}\Big(|Z_{i_1}| \ge \sqrt{x_p}, \cdots, |Z_{i_t}| \ge \sqrt{x_p} \Big) =: \sum_{1\le i_1<\cdots<i_t\le p} \mathsf{P}_{i_1,\cdots,i_t}$. Define $\mathcal{I} = \Big\{ 1 \le i_1 < \cdots < i_t \le p : \max_{1\le k<l\le t} |\mathsf{Cov}(Z_{i_k}, Z_{i_l})| \ge p^{-\gamma} \Big\}$, where $\gamma > 0$ is a

sufficiently small number to be specified later. For $2 \le d \le t-1$, define

$$\mathcal{I}_d = \Big\{1 \le i_1 < \cdots < i_t \le p : \text{Card}(S) = d, \text{ where } S \text{ is the largest subset of } \{i_1, ..., i_t\} \text{ such that } \forall i_k \ne i_l \in S,\ |\mathsf{Cov}(Z_{i_k}, Z_{i_l})| < p^{-\gamma}\Big\}.$$

For $d = 1$, define $\mathcal{I}_1 = \Big\{1 \le i_1 < \cdots < i_t \le p : |\mathsf{Cov}(Z_{i_k}, Z_{i_l})| \ge p^{-\gamma} \text{ for every } 1 \le k < l \le t\Big\}$. So we have $\mathcal{I} = \cup_{d=1}^{t-1} \mathcal{I}_d$. Let $\text{Card}(\mathcal{I}_d)$ denote the total number of the vectors $(i_1, \ldots, i_t)$ in $\mathcal{I}_d$. We can show that $\text{Card}(\mathcal{I}_d) \le Cp^{d+2\gamma t}$. In fact, the total number of the subsets of $\{i_1, ..., i_t\}$ with cardinality d is C_p^d. For a fixed subset S with cardinality d, the number of i such that $|\mathsf{Cov}(Z_i, Z_j)| \ge p^{-\gamma}$ for some $j \in S$ is no more than $Cdp^{2\gamma}$. This implies that $\text{Card}(\mathcal{I}_d) \le Cp^{d+2\gamma t}$. Define $\mathcal{I}^c = \{1 \le i_1 < \cdots < i_t \le p\} \setminus \mathcal{I}$. Then the number of elements in the sum $\sum_{(i_1,\cdots,i_t)\in\mathcal{I}^c} \mathsf{P}_{i_1,\cdots,i_t}$ is $C_p^t - O(\sum_{d=1}^{t-1} p^{d+2\gamma t}) = C_p^t - O(p^{t-1+2\gamma t}) = (1 + o(1))C_p^t$.

To prove Lemma 6, it suffices to show that

$$\mathsf{P}_{i_1,\cdots,i_t} = (1 + o(1))\pi^{-\frac{t}{2}} p^{-t} \exp(-\frac{tx}{2}) \tag{20}$$

uniformly in $(i_1, \ldots, i_t) \in \mathcal{I}^c$, and for $1 \le d \le t-1$,

$$\sum_{(i_1,\cdots,i_t)\in\mathcal{I}_d} \mathsf{P}_{i_1,\cdots,i_t} \to 0. \tag{21}$$

Putting (19) - (21) together, we obtain that

$$(1 + o(1))S_{2k} \le \mathsf{P}\Big(\max_{1\le i\le p} |Z_i| \ge \sqrt{x_p}\Big) \le (1 + o(1))S_{2k-1}, \tag{22}$$

where $S_k = \sum_{t=1}^{k} (-1)^{t-1} \frac{1}{t!} \pi^{-\frac{t}{2}} \exp(-\frac{tx}{2})$. Note that $\lim_{k\to\infty} S_k = 1 - \exp(-\frac{1}{\sqrt{\pi}} e^{-x/2})$. By letting $p \to \infty$ first and then $k \to \infty$ in (22), we prove Lemma 6.

We now prove (20). Let $\boldsymbol{z} = (z_{i_1}, ..., z_{i_t})'$ and $|\boldsymbol{z}|_{\min} = \min_{1\le j\le t} |z_{i_j}|$. Write

$$\mathsf{P}_{i_1,\cdots,i_t} = \frac{1}{(2\pi)^{t/2} \det(\boldsymbol{\Omega}_t)^{\frac{1}{2}}} \int_{|\boldsymbol{z}|_{\min} \ge \sqrt{x_p}} \exp(-\frac{1}{2} \boldsymbol{z}' \boldsymbol{\Omega}_t^{-1} \boldsymbol{z}) d\boldsymbol{z},$$

where $\boldsymbol{\Omega}_t$ is the covariance matrix of $\boldsymbol{Z} = (Z_{i_1}, ..., Z_{i_t})'$, and $\boldsymbol{\Omega}_t = (a_{kl})_{t\times t}$, where $a_{kl} = \mathsf{Cov}(Z_{i_k}, Z_{i_l})$. Since $i_1, ..., i_t \in \mathcal{I}^c$, $a_{k,k} = 1$ and $|a_{kl}| < p^{-\gamma}$ for $k \ne l$. Write

$$\int_{|\boldsymbol{z}|_{\min} \ge \sqrt{x_p}} \exp(-\frac{1}{2} \boldsymbol{z}' \boldsymbol{\Omega}_t^{-1} \boldsymbol{z}) d\boldsymbol{z} = \int_{|\boldsymbol{z}|_{\min} \ge \sqrt{x_p}, \|\boldsymbol{z}\|^2 > (\log p)^2} \exp(-\frac{1}{2} \boldsymbol{z}' \boldsymbol{\Omega}_t^{-1} \boldsymbol{z}) d\boldsymbol{z}$$

$$+\int_{|\boldsymbol{z}|_{\min}\geq\sqrt{x_p},\|\boldsymbol{z}\|^2\leq(\log p)^2}\exp(-\frac{1}{2}\boldsymbol{z}'\boldsymbol{\Omega}_t^{-1}\boldsymbol{z})d\boldsymbol{z}. \quad (23)$$

Then

$$\int_{|\boldsymbol{z}|_{\min}\geq\sqrt{x_p},\|\boldsymbol{z}\|^2>(\log p)^2}\exp(-\frac{1}{2}\boldsymbol{z}'\boldsymbol{\Omega}_t^{-1}\boldsymbol{z})d\boldsymbol{z}\leq C\exp(-(\log p)^2/2t)\leq Cp^{-2t}, \quad (24)$$

uniformly in $(i_1,\ldots,i_t)\in\mathcal{I}^c$. For the second part of the sum in (23), note that

$$\|\boldsymbol{\Omega}_t^{-1}-\boldsymbol{I}\|_2\leq\|\boldsymbol{\Omega}_t^{-1}\|_2\|\boldsymbol{\Omega}_t-\boldsymbol{I}\|_2\leq Cp^{-\gamma}. \quad (25)$$

Let $A=\{|\boldsymbol{z}|_{\min}\geq\sqrt{x_p},\|\boldsymbol{z}\|^2\leq(\log p)^2\}$. It follows that

$$\begin{aligned}\int_A e^{-\frac{1}{2}\boldsymbol{z}'\boldsymbol{\Omega}_t^{-1}\boldsymbol{z}}d\boldsymbol{z} &= \int_A e^{-\frac{1}{2}\boldsymbol{z}'(\boldsymbol{\Omega}_t^{-1}-\boldsymbol{I})\boldsymbol{z}-\frac{1}{2}\|\boldsymbol{z}\|^2}d\boldsymbol{z}\\ &= (1+O(p^{-\gamma}(\log p)^2))\int_A e^{-\frac{1}{2}\|\boldsymbol{z}\|^2}d\boldsymbol{z}\\ &= (1+O(p^{-\gamma}(\log p)^2))\int_{|\boldsymbol{z}|_{\min}\geq\sqrt{x_p}} e^{-\frac{1}{2}\|\boldsymbol{z}\|^2}d\boldsymbol{z}+Cp^{-2t}, \quad (26)\end{aligned}$$

uniformly in $(i_1,\ldots,i_t)\in\mathcal{I}^c$. This, together with (23) and (24), implies (20).

It remains to prove (21). For $S\subset\mathcal{I}_d$ with $d\geq 1$, without loss of generality, we can assume $S=\{i_{t-d+1},...,i_t\}$. By the definition of S and $\mathcal{I}_d$, for any $k\in\{i_1,\ldots,i_{t-d}\}$, there exists at least one $l\in S$ such that $|\mathsf{Cov}(Z_k,Z_l)|\geq p^{-\gamma}$. We divide $\mathcal{I}_d$ into two parts:

$$\begin{aligned}\mathcal{I}_{d,1} = \Big\{&1\leq i_1<\cdots<i_t\leq p: \text{there exists an } k\in\{i_1,\ldots,i_{t-d}\} \text{ such that}\\ &\text{for some } l_1,l_2\in S \text{ with } l_1\neq l_2,\ |\mathsf{Cov}(Z_k,Z_{l_1})|\geq p^{-\gamma} \text{ and } |\mathsf{Cov}(Z_k,Z_{l_2})|\geq p^{-\gamma}\Big\}\end{aligned}$$

and $\mathcal{I}_{d,2}=\mathcal{I}_d\setminus\mathcal{I}_{d,1}$. Clearly, $\mathcal{I}_{1,1}=\emptyset$ and $\mathcal{I}_{1,2}=\mathcal{I}_1$. Moreover, we can show that $\mathrm{Card}(\mathcal{I}_{d,1})\leq Cp^{d-1+2\gamma t}$. For any $(i_1,\ldots,i_t)\in\mathcal{I}_{d,1}$,

$$\mathsf{P}\Big(|Z_{i_1}|\geq\sqrt{x_p},\ldots,|Z_{i_t}|\geq\sqrt{x_p}\Big)\leq\mathsf{P}\Big(|Z_{i_{t-d+1}}|\geq\sqrt{x_p},\ldots,|Z_{i_t}|\geq\sqrt{x_p}\Big)=O(p^{-d}).$$

Hence by letting γ be sufficiently small,

$$\sum_{\mathcal{I}_{d,1}}\mathsf{P}_{i_1,\cdots,i_t}\leq Cp^{-1+2\gamma t}=o(1). \quad (27)$$

For any $(i_1, \ldots, i_t) \in \mathcal{I}_{d,2}$, without loss of generality, we assume that $|\mathsf{Cov}(Z_{i_1}, Z_{i_{t-d+1}})| \geq p^{-\gamma}$. Note that

$$\mathsf{P}\Big(|Z_{i_1}| \geq \sqrt{x_p}, \ldots, |Z_{i_t}| \geq \sqrt{x_p}\Big) \leq \mathsf{P}\Big(|Z_{i_1}| \geq \sqrt{x_p}, |Z_{i_{t-d+1}}| \geq \sqrt{x_p}, \ldots, |Z_{i_t}| \geq \sqrt{x_p}\Big).$$

Let $\boldsymbol{U}_l$ be the covariance matrix of $(Z_{i_1}, Z_{i_{t-d+1}}, \ldots, Z_{i_t})$. We can show that $\|\boldsymbol{U}_l - \bar{\boldsymbol{U}}_l\|_2 = O(p^{-\gamma})$, where $\bar{\boldsymbol{U}}_l = diag(\boldsymbol{D}, I_{d-1})$ and $\boldsymbol{D}$ is the covariance matrix of Z_{i_1} and $Z_{i_{t-d+1}}$. Using the similar arguments as in (23)-(26), we can get

$$\begin{aligned}
&\mathsf{P}\Big(|Z_{i_1}| \geq \sqrt{x_p}, |Z_{i_{t-d+1}}| \geq \sqrt{x_p}, \ldots, |Z_{i_t}| \geq \sqrt{x_p}\Big) \\
&\quad \leq (1 + o(1))\mathsf{P}(|Z_{i_1}| \geq \sqrt{x_p}, |Z_{i_{t-d+1}}| \geq \sqrt{x_p}) \times O(p^{-d+1}) \leq Cp^{-\frac{2}{1+r}} \times O(p^{-d+1}),
\end{aligned}$$

where the last inequality follows from Lemma 2 and the assumption $\max_{1 \leq i \neq j \leq p} |\omega_{i,j}| \leq r < 1$. Thus by letting γ be sufficiently small,

$$\sum_{\mathcal{I}_{d,2}} \mathsf{P}_{i_1, \cdots, i_t} \leq Cp^{d+2\gamma t - d + 1 - \frac{2}{1+r}} = o(1). \tag{28}$$

Combining (27) and (28), we prove (21). The proof of Lemma 6 is then complete. □

7.3 Proof of Theorem 2

It suffices to prove $\mathsf{P}\Big(\max_{1 \leq i \leq p} \Big|(\boldsymbol{\Omega}\boldsymbol{\delta})_i / \sqrt{\omega_{i,i}}\Big| \geq \sqrt{(2 + \varepsilon/2) \log p/n}\Big) \to 1$. By Lemma 3 and the condition $\max_i |\delta_i / \sigma_{i,i}^{\frac{1}{2}}| \geq \sqrt{2\beta \log p/n}$ with $\beta \geq 1/(\min_i \sigma_{i,i}\omega_{i,i}) + \varepsilon$ for some constant $\varepsilon > 0$, we can get $\max_{1 \leq i \leq p} |(\boldsymbol{\Omega}\boldsymbol{\delta})_i / \sqrt{\omega_{i,i}}| \geq \sqrt{(2 + \varepsilon/2) \log p/n}$ with probability tending to one. So Theorem 2 follows. □

7.4 Proof of Theorem 3

First we assume $k_p = o(p^r)$ for some $r < 1/4$, and we can get similar argument if $k_p = O(p^r)$ for some $r < 1/2$ and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is s_p sparse with $s_p = O((p/k_p^2)^\gamma)$ for some $0 < \gamma < 1$. Let $\mathcal{M}_{s,p}$ denote the set of all subsets of $\{1, ..., p\}$ with cardinality k_p. Let $\hat{m}$ be a random set of $\{1, ..., p\}$, which is uniformly distributed on $\mathcal{M}$. Let ω_j, $1 \leq j \leq p$ be i.i.d. variables with $\mathsf{P}(\omega_j = 1) = \mathsf{P}(\omega_j = -1) = 1/2$. We construct a class of $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ by letting

$\boldsymbol{\mu}_1 = 0$ and $\boldsymbol{\delta} = -\boldsymbol{\mu}_2$ satisfy $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^{'}$ with $\delta_j = \frac{\rho}{\sqrt{k_p}}\omega_j \mathbf{1}_{j\in\hat{m}}$, where $\rho = c\sqrt{\frac{k_p \log p}{n}}$ and $c > 0$ is sufficiently small that will be specified later. Clearly, $|\boldsymbol{\delta}|_2 = \rho$. Let μ_ρ be the distribution of $\boldsymbol{\delta}$. Note that μ_ρ is a probability measure on $\{\boldsymbol{\delta} \in \mathcal{S}_{k_p} : |\boldsymbol{\delta}| = \rho\}$. We now calculate the likelihood ratio $L_{\mu_\rho} = \frac{d\mathsf{P}_{\mu_\rho}}{d\mathsf{P}_0}(\{\boldsymbol{X}_n, \boldsymbol{Y}_n\})$. It is easy to see that $L_{\mu_\rho} = \mathsf{E}_{\hat{m},\omega}\Big(\exp(-\sqrt{n}\boldsymbol{Z}^{'}\boldsymbol{\delta} - \frac{n}{2}\boldsymbol{\delta}^{'}\boldsymbol{\Omega}\boldsymbol{\delta})\Big)$, where $\boldsymbol{Z}$ is a multivariate normal vector with mean 0 and $\mathsf{Cov}(\boldsymbol{Z}) = \boldsymbol{\Omega}$, and is independent with $\hat{m}$ and $\boldsymbol{\Omega}$. For any fixed $\hat{m} = m$, let $\boldsymbol{\delta}_m^i$, $1 \le i \le 2^{k_p}$ be all the possible values of $\boldsymbol{\delta}$. That is, $\mathsf{P}\Big(\boldsymbol{\delta} = \boldsymbol{\delta}_m^i | \hat{m} = m\Big) = 2^{-k_p}$. Thus

$$\mathsf{E}_{\hat{m},\omega}\Big(\exp(-\sqrt{n}\boldsymbol{Z}^{'}\boldsymbol{\delta} - \frac{n}{2}\boldsymbol{\delta}'\boldsymbol{\Omega}\boldsymbol{\delta})\Big) = \frac{1}{\binom{p}{k_p}}\frac{1}{2^{k_p}}\sum_{m\in\mathcal{M}}\sum_{i=1}^{2^{k_p}}\exp\Big(-\sqrt{n}\boldsymbol{Z}'\boldsymbol{\delta}_{(m)}^{(i)} - \frac{n}{2}\boldsymbol{\delta}_{(m)}^{(i)'}\boldsymbol{\Omega}\boldsymbol{\delta}_{(m)}^{(i)}\Big).$$

It follows that

$$\begin{aligned}
\mathsf{E}L_{\mu_\rho}^2 &= \mathsf{E}\Big\{\frac{1}{\binom{p}{k_p}}\frac{1}{2^{k_p}}\sum_{m\in\mathcal{M}}\sum_{i=1}^{2^{k_p}}\exp\Big(-\sqrt{n}\boldsymbol{Z}'\boldsymbol{\delta}_{(m)}^{(i)} - \frac{n}{2}\boldsymbol{\delta}_{(m)}^{(i)'}\boldsymbol{\Omega}\boldsymbol{\delta}_{(m)}^{(i)}\Big)\Big\}^2 \\
&= \frac{1}{\binom{p}{k_p}^2}\frac{1}{2^{2k_p}}\mathsf{E}\Big\{\sum_{m,m'\in\mathcal{M}}\sum_{i,j=1}^{2^{k_p}}\exp\Big(-\sqrt{n}\boldsymbol{Z}'(\boldsymbol{\delta}_{(m)}^{(i)} + \boldsymbol{\delta}_{(m')}^{(j)}) - \frac{n}{2}(\boldsymbol{\delta}_{(m)}^{(i)'}\boldsymbol{\Omega}\boldsymbol{\delta}_{(m)}^{(i)} + \boldsymbol{\delta}_{(m')}^{(j)'}\boldsymbol{\Omega}\boldsymbol{\delta}_{(m')}^{(j)})\Big)\Big\} \\
&= \frac{1}{\binom{p}{k_p}^2}\frac{1}{2^{2k_p}}\sum_{m,m'\in\mathcal{M}}\sum_{i,j=1}^{2^{k_p}}\Big[\exp\Big(-\frac{n}{2}(\boldsymbol{\delta}_{(m)}^{(i)'}\boldsymbol{\Omega}\boldsymbol{\delta}_{(m)}^{(i)} + \boldsymbol{\delta}_{(m')}^{(j)'}\boldsymbol{\Omega}\boldsymbol{\delta}_{(m')}^{(j)})\Big) \\
&\qquad\qquad \times \exp\Big(\frac{n}{2}(\boldsymbol{\delta}_{(m)}^{(i)'} + \boldsymbol{\delta}_{(m')}^{(j)'})\boldsymbol{\Omega}(\boldsymbol{\delta}_{(m)}^{(i)} + \boldsymbol{\delta}_{(m')}^{(j)})\Big)\Big] \\
&= \frac{1}{\binom{p}{k_p}^2}\frac{1}{2^{2k_p}}\sum_{m,m'\in\mathcal{M}}\sum_{i,j=1}^{2^{k_p}}\exp\Big(n\boldsymbol{\delta}_{(m)}^{(i)'}\boldsymbol{\Omega}\boldsymbol{\delta}_{(m')}^{(j)}\Big) = \frac{1}{\binom{p}{k_p}^2}\frac{1}{2^{2k_p}}\sum_{m,m'\in\mathcal{M}}\sum_{i,j=1}^{2^{k_p}}\exp\Big(\frac{n\rho^2}{k_p}\sum_{k\in m,l\in m'}a_{k,l}\omega_k^{(i)}\omega_l^{(j)}\Big),
\end{aligned}$$

where $\frac{\rho}{\sqrt{k_p}}(\omega_k^{(i)} I_{k\in m}) := \boldsymbol{\delta}_{(m)}^{(i)}$ and $\boldsymbol{\Omega} = (a_{kl})_{p\times p}$. Thus

$$\begin{aligned}
\mathsf{E}L_{\mu_\rho}^2 &= \frac{1}{\binom{p}{k_p}^2}\frac{1}{2^{2k_p}}\sum_{m,m'\in\mathcal{M}}2^{k_p}\prod_{k,l=1}^{k_p}\Big(\exp(\frac{n\rho^2}{k_p}a_{kl}) + \exp(-\frac{n\rho^2}{k_p}a_{kl})\Big) \\
&= \frac{1}{\binom{p}{k_p}^2}\sum_{m,m'\in\mathcal{M}}\prod_{k\in m,l\in m'}\cosh(\frac{n\rho^2}{k_p}a_{kl}) \le \frac{1}{\binom{p}{k_p}^2}\sum_{m,m'\in\mathcal{M}}\prod_{k\in m,l\in m'}\exp\Big(\frac{n\rho^2}{k_p}|a_{kl}|\Big).
\end{aligned}$$

For every m, let $B := B_m = \{l : \ |a_{k,l}| \ge \frac{M}{d}, k \in m\}$, where $d = \Big(\frac{p}{k_p^2}\Big)^{1-\gamma}$ and γ is sufficiently small. For every k, the number of l such that $|a_{kl}| \ge \frac{M}{d}$ is at most d. Hence

$$\mathsf{E}L_{\mu_\rho}^2 \ \le \ \frac{1}{\binom{p}{k_p}^2}\sum_{m\in\mathcal{M}}\sum_{j=0}^{k_p} I\{|m^{'}\cap B| = j\}\exp\Big(\sum_{k\in m,l\in m'}\frac{n\rho^2}{k_p}|a_{kl}|\Big)$$

$$
\begin{aligned}
&= \frac{1}{\binom{p}{k_p}^2}\sum_{m\in\mathcal{M}}\sum_{j=0}^{k_p} I\{|m^{'}\cap B|=j\}\exp\Big(\sum_{k\in m,l\in m^{'}\cap B}\frac{n\rho^2}{k_p}|a_{kl}|+\sum_{k\in m,l\in m^{'}\cap B^c}\frac{n\rho^2}{k_p}|a_{kl}|\Big)\\
&\le \frac{1}{\binom{p}{k_p}^2}\sum_{m\in\mathcal{M}}\sum_{j=0}^{k_p}\binom{k_p d}{j}\binom{p-k_p}{k_p-j}\exp\Big(\frac{Mn\rho^2}{k_p}j+\frac{Mk_p^2\log p}{d}\Big)\\
&\le \frac{1}{\binom{p}{k_p}}\sum_{j=0}^{k_p}\binom{k_p d}{j}\binom{p-k_p}{k_p-j}\exp\Big(\frac{Mn\rho^2}{k_p}j+\frac{Mk_p^2\log p}{d}\Big)\\
&\le (1+o(1))\sum_{j=0}^{k_p}\binom{k_p}{j}\frac{(dk_p)^j}{p^j}\exp\Big(\frac{Mn\rho^2}{k_p}j+\frac{Mk_p^2\log p}{d}\Big)\\
&= (1+o(1))\Big(1+\frac{dk_p t}{p}\Big)^{k_p}\exp\Big(\frac{Mk_p^2\log p}{d}\Big),
\end{aligned}
$$

where $t=\exp\left(\frac{Mn\rho^2}{k_p}\right)=p^{Mc^2}$. It follows that

$$
\begin{aligned}
\mathsf{E}L_{\mu_\rho}^2 &\le (1+o(1))\exp\Big(k_p\log\Big(1+\frac{dk_p t}{p}\Big)+\frac{Mk_p^2\log p}{d}\Big)\\
&\le (1+o(1))\exp\Big(k_p\frac{dk_p t}{p}+\frac{Mk_p^2\log p}{d}\Big)\le 1+4(1-\alpha-\nu)^2
\end{aligned}
$$

by letting c be sufficiently small. If $k_p=O(p^r)$ for some $r<1/2$ and $\mathbf{\Omega}=\mathbf{\Sigma}^{-1}$ is s_p sparse with $s_p=O((p/k_p^2)^\gamma)$ for some $0<\gamma<1$, we let $B:=B_m=\{l:\ a_{k,l}\neq 0, k\in m\}$. Then we can similarly get

$$
\begin{aligned}
\mathsf{E}L_{\mu_\rho}^2 &\le \frac{1}{\binom{p}{k_p}^2}\sum_{m\in\mathcal{M}}\sum_{j=0}^{k_p} I\{|m^{'}\cap B|=j\}\exp\Big(\sum_{k\in m,l\in m^{'}}\frac{n\rho^2}{k_p}|a_{kl}|\Big)\\
&\le \frac{1}{\binom{p}{k_p}^2}\sum_{m\in\mathcal{M}}\sum_{j=0}^{k_p} I\{|m^{'}\cap B|=j\}\exp\Big(\frac{Mn\rho^2}{k_p}j\Big)\le (1+o(1))\Big(1+\frac{s_p k_p t}{p}\Big)^{k_p},
\end{aligned}
$$

So we can still get $\mathsf{E}L_{\mu_\rho}^2\le 1+4(1-\alpha-\nu)^2$ by letting c be sufficiently small. Theorem 3 now follows from Lemma 5. □

7.5 Proof of Theorem 4

We only prove part (ii) of Theorem 4 in this section, part (i) follows from the proof of part (ii) directly. Without loss of generality, we assume that $\sigma_{i,i}=1$ for $1\le i\le p$. Define the

event $\boldsymbol{A} = \{\max_{1\le i\le p}|\delta_i| \le 8\sqrt{\log p/n}\}$. By conditions in Theorem 4, we have

$$\max_{1\le i\le p}|\hat{\omega}_{i,i}^{(0)} - \omega_{i,i}| = o_{\mathsf{P}}(1/\log p).$$

Hence, as in the proof of Proposition 1 (i) in the supplement material, it is easy to show that $\mathsf{P}\Big(M_{\widehat{\boldsymbol{\Omega}}} \in R_\alpha, \boldsymbol{A}^c\Big) = \mathsf{P}(\boldsymbol{A}^c) + o(1)$ and $\mathsf{P}\Big(M_{\boldsymbol{\Omega}} \in R_\alpha, \boldsymbol{A}^c\Big) = \mathsf{P}(\boldsymbol{A}^c) + o(1)$. Note that $\widehat{\boldsymbol{\Omega}}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}) = (\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}} - \boldsymbol{\delta}) + (\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\boldsymbol{\delta} + \boldsymbol{\Omega}(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}})$. On $\boldsymbol{A}$, we have

$$\Big|(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})(\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}} - \boldsymbol{\delta}) + (\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})\boldsymbol{\delta}\Big|_\infty = o_{\mathsf{P}}\Big(\frac{1}{\sqrt{n\log p}}\Big).$$

To prove Theorem 4, it suffices to show that

$$\mathsf{P}\Big(\max_{1\le i\le p}|Z_i^o| \ge \sqrt{x_p} + a_n, \boldsymbol{A}\Big) = \mathsf{P}\Big(\max_{1\le i\le p}|Z_i^o| \ge \sqrt{x_p}, \boldsymbol{A}\Big) + o(1), \tag{29}$$

for any $a_n = o((\log p)^{-1/2})$, where $Z_i^o = (\boldsymbol{\Omega}\boldsymbol{Z})_i/\sqrt{\omega_{i,i}}$ defined in the proof of Proposition 1 (i) in the supplement material. From the proof of Proposition 1 (i), let $H = \mathrm{supp}(\boldsymbol{\delta}) = \{l_1, ..., l_{p^r}\}$, then we can get

$$\begin{aligned}
\mathsf{P}\Big(\max_{1\le i\le p}|Z_i^o| \ge \sqrt{x_p} + a_n, \boldsymbol{A}\Big) &= \alpha\mathsf{P}(\boldsymbol{A}) + (1-\alpha)\mathsf{P}(\max_{i\in H}|Y_i| \ge \sqrt{x_p} + a_n, \boldsymbol{A}) + o(1),\\
\mathsf{P}\Big(\max_{1\le i\le p}|Z_i^o| \ge \sqrt{x_p}, \boldsymbol{A}\Big) &= \alpha\mathsf{P}(\boldsymbol{A}) + (1-\alpha)\mathsf{P}(\max_{i\in H}|Y_i| \ge \sqrt{x_p}, \boldsymbol{A}) + o(1),
\end{aligned}$$

where given $\boldsymbol{\delta}$, Y_i, $i \in H$ are independent normal random variables with unit variance. This, together with Lemma 4, implies (29). □

7.6 Proof of Theorem 6

Let $(V_1, ..., V_p)'$ be a zero mean random vector with covariance matrix $\boldsymbol{\Omega} = (\omega_{i,j})$ and the diagonal $\omega_{i,i} = 1$ for $1 \le i \le p$ satisfying moment conditions (C6) or (C7). Let $\hat{V}_{li} = V_{li}I\{|V_{li}| \le \tau_n\}$ for $l = 1, ..., n$, where $\tau_n = \eta^{-1/2}2\sqrt{\log(p+n)}$ if $(C6)$ holds and $\tau_n = \sqrt{n}/(\log p)^8$ if (C7) holds. Let $W_i = \sum_{l=1}^n V_{li}/\sqrt{n}$ and $\hat{W}_i = \sum_{l=1}^n \hat{V}_{li}/\sqrt{n}$. Then

$$\mathsf{P}(\max_{1\le i\le p}|W_i - \hat{W}_i| \ge \frac{1}{\log p}) \le \mathsf{P}(\max_{1\le i\le p}\max_{1\le l\le n}|V_{li}| \ge \tau_n) \le np\max_{1\le i\le p}\mathsf{P}(|V_{1i}| \ge \tau_n) = O(p^{-1} + n^{-\epsilon/8}). \tag{30}$$

Note that

$$\Big|\max_{1\le i\le p}W_i^2 - \max_{1\le i\le p}\hat{W}_i^2\Big| \le 2\max_{1\le i\le p}|W_i|\max_{1\le i\le p}|W_i - \hat{W}_i| + \max_{1\le i\le p}|W_i - \hat{W}_i|^2. \tag{31}$$

By (30) and (31), it is enough to prove that for any $x \in R$, as $p \to \infty$

$$\mathsf{P}(\max_{1\le i\le p} \hat{W}_i^2 - 2\log p + \log\log p \le x) \to \exp(-\frac{1}{\sqrt{\pi}}\exp(-\frac{x}{2})).$$

It follows from Lemma 1 that for any fixed $k \le [p/2]$,

$$\sum_{t=1}^{2k}(-1)^{t-1}\sum_{1\le i_1<\ldots<i_t\le p}\mathsf{P}(|\hat{W}_{i_1}| \ge x_p, ..., |W_{i_t}| \ge x_p) \le \mathsf{P}(\max_{1\le i\le p}|\hat{W}_i| \ge x_p)$$
$$\le \sum_{t=1}^{2k-1}(-1)^{t-1}\sum_{1\le i_1<\ldots<i_t\le p}\mathsf{P}(|\hat{W}_{i_1}| \ge x_p, ..., |W_{i_t}| \ge x_p). \quad (32)$$

Define $|\hat{\boldsymbol{W}}|_{\min} = \min_{1\le l\le t}|\hat{W}_{i_l}|$. Then by Theorem 1 in Zaitsev (1987), we have

$$\mathsf{P}(|\hat{\boldsymbol{W}}|_{\min} \ge x_p) \le \mathsf{P}(|\boldsymbol{Z}|_{\min} \ge x_p - \epsilon_n(\log p)^{-1/2}) + c_1 d^{5/2}\exp(-\frac{n^{\frac{1}{2}}\epsilon_n}{c_2 d^3\tau_n(\log p)^{\frac{1}{2}}}), \quad (33)$$

where $c_1 > 0$ and $c_2 > 0$ are absolute constants, $\epsilon_n \to 0$ which will be specified later and $\boldsymbol{Z} = (Z_{i_1}, ..., Z_{i_t})'$ is a t dimensional normal vector as defined in Theorem 1. Because $\log p = o(n^{1/4})$, we can let $\epsilon \to 0$ sufficiently slow such that

$$c_1 d^{5/2}\exp(-\frac{n^{\frac{1}{2}}\epsilon_n}{c_2 d^3\tau_n(\log p)^{\frac{1}{2}}}) = O(p^{-M}) \quad (34)$$

for any large $M > 0$. It follows from (32), (33) and (34) that

$$\mathsf{P}(\max_{1\le i\le p}|\hat{W}_i| \ge x_p) \le \sum_{t=1}^{2k-1}(-1)^{t-1}\sum_{1\le i_1<\ldots<i_t\le p}\mathsf{P}(|Z|_{\min} \ge x_p - \epsilon_n(\log p)^{-1/2}) + o(1). \quad (35)$$

Similarly, using Theorem 1 in Zaïtsev (1987) again, we can get

$$\mathsf{P}(\max_{1\le i\le p}|\hat{W}_i| \ge x_p) \ge \sum_{t=1}^{2k}(-1)^{t-1}\sum_{1\le i_1<\ldots<i_t\le p}\mathsf{P}(|Z|_{\min} \ge x_p - \epsilon_n(\log p)^{-1/2}) - o(1). \quad (36)$$

So by (35), (36) and the proof of Theorem 1, the theorem is proved. □

7.7 Proof of Theorem 7

(i). (i) follows from the proof of Theorem 4.

(ii). Note that $\widehat{\boldsymbol{\Omega}}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}) = (\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega})(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}-\boldsymbol{\delta})+(\widehat{\boldsymbol{\Omega}}-\boldsymbol{\Omega})\boldsymbol{\delta}+\boldsymbol{\Omega}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}})$. It suffices to prove $\mathsf{P}\Big(\max_{1\le i\le p}\Big|\sqrt{n}(\widehat{\boldsymbol{\Omega}}\boldsymbol{\delta})_i/\sqrt{\omega_{i,i}}+\sqrt{n}\boldsymbol{\Omega}(\bar{\boldsymbol{X}}-\bar{\boldsymbol{Y}}-\boldsymbol{\delta})_i/\sqrt{\omega_{i,i}}\Big|\ge\sqrt{\rho\log p}\Big)\to 1$ for some $\rho>2$. To this end, we only need to show $\mathsf{P}\Big(\max_{1\le i\le p}\Big|(\widehat{\boldsymbol{\Omega}}\boldsymbol{\delta})_i/\sqrt{\omega_{i,i}}\Big|\ge\sqrt{(2+\varepsilon/4)\log p/n}\Big)\to 1$. Note that

$$\max_{1\le i\le p}|(\widehat{\boldsymbol{\Omega}}\boldsymbol{\delta})_i/\sqrt{\omega_{i,i}}|\ge\max_{1\le i\le p}|(\boldsymbol{\Omega}\boldsymbol{\delta})_i/\sqrt{\omega_{i,i}}|+o_{\mathsf{P}}(1)\max_{1\le i\le p}|\boldsymbol{\delta}_i|\ge(1+o_{\mathsf{P}}(1))\max_{1\le i\le p}|(\boldsymbol{\Omega}\boldsymbol{\delta})_i/\sqrt{\omega_{i,i}}|.$$

By the condition $\max_i|\delta_i/\sigma_{i,i}^{\frac{1}{2}}|\ge\sqrt{2\beta\log p/n}$ with $\beta\ge 1/(\min_i\sigma_{i,i}\omega_{i,i})+\varepsilon$ for some constant $\varepsilon>0$, we can get $\max_{1\le i\le p}|(\boldsymbol{\Omega}\boldsymbol{\delta})_i/\sqrt{\omega_{i,i}}|\ge\sqrt{(2+\varepsilon/2)\log p/n}$ with probability tending to one. This proves (ii). □

Acknowledgements

The authors would like to thank the Associate Editor and three referees for their helpful constructive comments which have helped to improve quality and presentation of the paper.

This research was supported in part by NSF FRG Grant DMS-0854973. Weidong Liu's research was also supported by NSFC, Grant No.11201298, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, Foundation for the Author of National Excellent Doctoral Dissertation of PR China and the startup fund from SJTU.

References

[1] Anderson, T.W. (2003). *An introduction to multivariate statistical analysis.* Third edition. Wiley-Interscience.

[2] Arias-Castro, E., Candès, E. and Plan, Y. (2011). Global Testing under Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism. *Ann. Statist.* 39:2533-2556.

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

[4] Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* 8:577-606.

[5] Berman, S.M. (1962). A law of large numbers for the maximum of a stationary Gaussian sequence. *Ann. Math. Statist.* 33:93-97.

[6] Berman, S.M. (1964). Limit theorems for the maximum term in stationary sequences. *Ann. Math. Statist.* 35:502-516.

[7] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* 36:2577-2604.

[8] Birnbaum, A. and Nadler, B. (2012). High dimensional sparse covariance estimation: accurate thresholds for the maximal diagonal entry and for the largest correlation coefficient. Technical Report.

[9] Cai, T. and Liu, W.D. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* 106:672-684.

[10] Cai, T., Liu, W.D. and Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* 106:594-607.

[11] Cai, T., Liu, W.D. and Xia, Y. (2013a). Two-Sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.*, 108: 265-277.

[12] Cai, T., Liu, W.D. and Xia, Y. (2013b). Supplement to "Two-Sample test of high dimensional means under dependency". Technical report.

[13] Cai, T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.* 40: 2014-2042.

[14] Cai, T. and Zhou, H. (2012). Minimax estimation of large covariance matrices under ℓ_1 norm (with discussion). *Statist. Sinica* 22:1319-1378.

[15] Cai, T., Zhang, C.-H. and Zhou, H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* 38: 2118-2144.

[16] Cao, J. and Worsley, K.J. (1999). The detection of local shape changes via the geometry of Hotelling's T^2 fields. *Ann. Statist.* 27:925-942.

[17] Castagna J. P., Sun S., and Siegfried R. W. (2003). Instantaneous spectral analysis: Detection of low-frequency shadows associated with hydrocarbons. *The Leading Edge* 22:120-127.

[18] Chen, S. and Qin, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* 38:808-835.

[19] Hall, P. (1991). On convergence rates of suprema. *Probab. Theory Related Fields* 89:447-455.

[20] Hall, P. and Jin, J. (2008). Properties of higher criticism under strong dependence. *Ann. Statist.* 36:381-402.

[21] Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* 38:1686-1732.

[22] James D., Clymer B. D., and Schmalbrock P. (2001). Texture detection of simulated microcalcification susceptibility effects in magnetic resonance imaging of breasts. *J. Magn. Reson. Imaging* 13:876-881.

[23] Liu, W., Lin, Z.Y. and Shao, Q.M. (2008), The asymptotic distribution and Berry-Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *Ann. Appl. Probab.* 18:2337-2366.

[24] Ravikumar, P., Raskutti, G., Wainwright, M.J. and Yu, B. (2008). Model selection in Gaussian Graphical Models: high-dimensional consistency of l_1-regularized MLE. In *Advances in Neural Information Processing Systems (NIPS)* 21.

[25] Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* 2:494-515.

[26] Srivastava, M. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* 99:386-402.

[27] Srivastava, M. (2009). A test for the mean vector with fewer observations than the dimension under non-normality. *J. Multivariate Anal.* 100:518-532.

[28] Taylor, J.E. and Worsley, K.J. (2008). Random fields of multivariate test statistics, with applications to shape analysis. *Ann. Statist.* 36:1-27.

[29] Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* 11:2261-2286.

[30] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* 94:19-35.

[31] Zaïtsev, A. Yu. (1987), On the Gaussian approximation of convolutions under multidimensional analogues of S.N. Bernstein's inequality conditions. *Probab. Theory Related Fields* 74:535-566.

[32] Zhang G., Zhang S., and Wang Y. (2000). Application of adaptive time-frequency decomposition in ultrasonic NDE of highly-scattering materials. *Ultrasonics* 38:961-964.